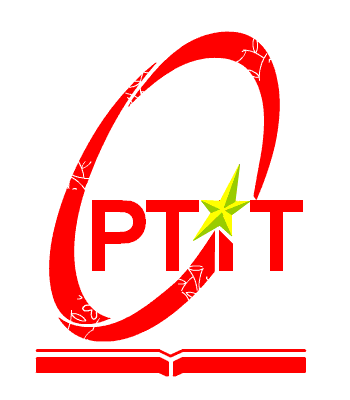


**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

**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 11/2020**

TÓM TẲT

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 2020*

*Sinh viên thực hiện*

**Nguyễn Thị Lan Anh**

# **NHẬN XÉT**

(Của giảng viên phản biện)

# **NHẬN XÉT**

(Của giảng viên hướng dẫn)

MỤC LỤC

[TÓM TẲT 2](#_Toc59865816)

[LỜI CẢM ƠN 3](#_Toc59865817)

[**NHẬN XÉT** 4](#_Toc59865818)

[**NHẬN XÉT** 5](#_Toc59865819)

[MỤC LỤC 6](#_Toc59865820)

[DANH MỤC HÌNH VẼ 8](#_Toc59865821)

[DANH MỤC BẢNG BIỂU 10](#_Toc59865822)

[BẢNG THUẬT NGỮ TIẾNG ANH 11](#_Toc59865823)

[LỜI NÓI ĐẦU 12](#_Toc59865824)

[CHƯƠNG 1: GIỚI THIỆU CHUNG VỀ HỆ KHUYẾN NGHỊ VÀ HỆ KHUYẾN NGHỊ SÁCH TRONG BÁN HÀNG TRỰC TUYẾN 13](#_Toc59865825)

[1.1. Giới thiệu hệ khuyến nghị – Recommender System (nêu được lý do chọn đề tài) 13](#_Toc59865826)

[1.2. Lí do chọn đề tài 14](#_Toc59865827)

[1.3. Giới thiệu một số hệ khuyến nghị sách 15](#_Toc59865828)

[1.4. Phát biểu bài toán khuyến nghị 20](#_Toc59865829)

[1.5. Kết luận chương 21](#_Toc59865830)

[**CHƯƠNG 2: CÁC KỸ THUẬT KHUYẾN NGHỊ** 22](#_Toc59865831)

[2.1. Kỹ thuật khuyến nghị lọc dựa trên nội dung 22](#_Toc59865832)

[**2.1.1.** **Khái niệm** 22](#_Toc59865833)

[**2.1.2.** **Phát biểu bài toán tư vấn lọc theo nội dung** 22](#_Toc59865834)

[**2.1.3.** **Xây dựng hồ sơ sản phẩm – Item Profiles** 22](#_Toc59865835)

[**2.1.4.** **Các phương pháp lọc theo nội dung** 23](#_Toc59865836)

[**2.1.5.** **Ưu điểm và nhược điểm của phương pháp lọc nội dung** 27](#_Toc59865837)

[2.2. Kỹ thuật khuyến nghị sách dựa trên lọc cộng tác 27](#_Toc59865838)

[**2.2.1.** **Khái niệm** 28](#_Toc59865839)

[**2.2.2.** **Phát biểu bài toán lọc cộng tác** 29](#_Toc59865840)

[**2.2.3.** **Phương pháp tư vấn lọc cộng tác dựa trên bộ nhớ** 30](#_Toc59865841)

[**2.2.4.** **Phương pháp tư vấn dựa lọc cộng tác dựa trên mô hình Matrix Factorization** 36](#_Toc59865842)

[**2.2.5.** **Ưu điểm và nhược điểm của phương pháp lọc cộng tác** 38](#_Toc59865843)

[2.3. Tổng kết chương 39](#_Toc59865844)

[CHƯƠNG 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ VỚI DỮ LIỆU SÁCH 40](#_Toc59865845)

[3.1. Thu thập dữ liệu 40](#_Toc59865846)

[3.2. Xử lý dữ liệu 45](#_Toc59865847)

[3.2.1. Tiền xử lý dữ liệu 45](#_Toc59865848)

[3.2.2. Chia dữ liệu thực nghiệm 46](#_Toc59865849)

[3.3. Thực nghiệm mô hình tư vấn 47](#_Toc59865850)

[3.3.1. Thực nghiệm mô hình tư vấn sách theo kỹ thuật lọc dựa trên nội dung 47](#_Toc59865851)

[3.3.2. Thực nghiệm mô hình tư vấn sách theo kỹ thuật lọc cộng tác 47](#_Toc59865852)

[3.4. Đánh giá thuật toán 48](#_Toc59865853)

[**3.4.1.** **Các thông số đánh giá giải thuật** 48](#_Toc59865854)

[3.4.2. Kết quả thực nghiệm 49](#_Toc59865855)

[3.4.3. Nhận xét kết quả và phân tích lỗi trong quá trình thực nghiệm 57](#_Toc59865856)

[3.5. Kết luận chương 57](#_Toc59865857)

[CHƯƠNG 4: ỨNG DỤNG MÔ HÌNH KHUYẾN NGHỊ SÁCH VÀO HỆ THỐNG BÁN HÀNG TRỰC TUYẾN 58](#_Toc59865858)

[4.1. Mô tả hệ thống 58](#_Toc59865859)

[4.2. Phân tích thiết kế hệ thống 58](#_Toc59865860)

[4.3. Thiết kế hệ thống 59](#_Toc59865861)

[**4.3.1.** **Các công nghệ sử dụng** 59](#_Toc59865862)

[**4.3.2.** **Mô hình tổng quát của hệ thống** 59](#_Toc59865863)

[**4.3.3.** **Triển khai ứng dụng thực tế** 60](#_Toc59865864)

[**KẾT LUẬN** 61](#_Toc59865865)

[**DANH MỤC TÀI LIỆU THAM KHẢO** 62](#_Toc59865866)

DANH MỤC HÌNH VẼ

[Hình 1: Các công ty tích hợp hệ gợi ý trong sản phẩm 14](#_Toc59844440)

[Hình 2: Minh họa về hệ gợi ý của Amazon 15](#_Toc59844441)

[Hình 3: Minh họa về hệ gợi ý của Goodread 16](#_Toc59844442)

[Hình 4: Minh họa sản phẩm của Tiki 17](#_Toc59844443)

[Hình 5: Minh họa sản phẩm gợi ý của Tiki - 1 17](#_Toc59844444)

[Hình 6: Minh họa sản phẩm gợi ý của Tiki - 2 17](#_Toc59844445)

[Hình 7: Minh họa sản phẩm của Fahasa 18](#_Toc59844446)

[Hình 8: Minh họa sản phẩm gợi ý của Fahasa - 1 18](#_Toc59844447)

[Hình 9: Minh họa sản phẩm gợi ý của Fahasa - 2 19](#_Toc59844448)

[Hình 10: Minh họa sản phẩm của Nhã Nam 19](#_Toc59844449)

[Hình 11: Minh họa sản phẩm gợi ý của Nhã Nam 20](#_Toc59844450)

[Hình 12: Minh họa phương pháp lọc nội dung dựa vào bộ nhớ 23](#_Toc59844451)

[Hình 13: Minh họa khoảng cách cosine giữa hai vector 24](#_Toc59844452)

[Hình 14: Mô hình lọc cộng tác 30](#_Toc59844453)

[Hình 15: Minh họa phương pháp lọc cộng tác dựa trên người dùng 31](#_Toc59844454)

[Hình 16: Chuẩn hóa ma trận - 1 32](#_Toc59844455)

[Hình 17: Chuẩn hóa ma trận - 2 32](#_Toc59844456)

[Hình 18: Chuẩn hóa ma trận - 3 33](#_Toc59844457)

[Hình 19: Chuẩn hóa ma trận - 4 34](#_Toc59844458)

[Hình 20: Minh họa chuẩn hóa ma trận 34](#_Toc59844459)

[Hình 21: Minh họa chuẩn hóa utility matrix 36](#_Toc59844460)

[Hình 22: Kỹ thuật phân rã ma trận 36](#_Toc59844461)

[Hình 23: Bài toán tối ưu ma trận 37](#_Toc59844462)

[Hình 24: Các bước xây dựng mô hình khuyến nghị 40](#_Toc59844463)

[Hình 25: Biểu đồ phân phối đánh giá (dữ liệu thô) 41](#_Toc59844464)

[Hình 26: Biểu đồ phân phối số sao đánh giá theo sản phẩm (dữ liệu thô) 41](#_Toc59844465)

[Hình 27: Biểu đồ phân phối số sao đánh giá theo người dùng (dữ liệu thô) 42](#_Toc59844466)

[Hình 28: Một vài mẫu dữ liệu sách 43](#_Toc59844467)

[Hình 29: Biểu đồ phân phối đánh giá (dữ liệu sau khi xử lý) 44](#_Toc59844468)

[Hình 30: Biểu đồ phân phối số sao đánh giá theo sản phẩm (dữ liệu sau khi xử lý) 44](#_Toc59844469)

[Hình 31: Biểu đồ phân phối số sao đánh giá theo người dùng (dữ liệu sau khi xử lý) 45](#_Toc59844470)

[Hình 32: Dữ liệu đánh giá của ngưởi dùng 45](#_Toc59844471)

[Hình 33: Tiền xử lý dữ liệu văn bản 46](#_Toc59844472)

[Hình 34: Minh họa cách chia dữ liệu 47](#_Toc59844473)

[Hình 35: Sơ đồ usecase toàn hệ thống 57](#_Toc59844474)

[Hình 36: Biểu đồ lớp toàn hệ thống 58](#_Toc59844475)

[Hình 37: Lược đồ cơ sở dữ liệu 58](#_Toc59844476)

DANH MỤC BẢNG BIỂU

[Bảng 1: Ví dụ Utility matrix 23](#_Toc60057545)

[Bảng 2: Utility matrix với vector đặc trưng của sản phẩm 24](#_Toc60057546)

[Bảng 3: Ma trận đánh giá của người dùng 26](#_Toc60057547)

[Bảng 4: Ma trận đánh giá R 27](#_Toc60057548)

[Bảng 5: Kết quả đo precision và recall theo các ngưỡng bằng phương pháp lọc theo bộ nhớ 47](#_Toc60057549)

[Bảng 6: Kết quả đo precision và recall theo k bằng phương pháp lọc theo bộ nhớ 47](#_Toc60057550)

[Bảng 7: Kết quả đo precision và recall theo các ngưỡng bằng phương pháp lọc phân rã ma trận 49](#_Toc60057551)

[Bảng 8: Kết quả đo precision và recall theo k bằng phương pháp lọc phân rã ma trận 50](#_Toc60057552)

BẢNG THUẬT NGỮ TIẾNG ANH

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 (hay hệ tư vấn) ra đời giúp chúng ta dễ dàng tìm thấy thông tin thích hợp nhất trên mạng internet.

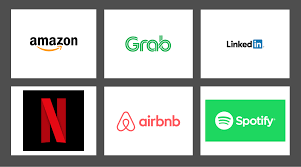
Hệ tư vấn ra đời hướng đến việc giảm tải thông tin cho mỗi người dùng bằng cách đưa ra những gợi ý thông tin phù hợp và gỡ bỏ những thông tin không phù hợp cho mỗi người dùng. Đối với người dùng, hệ tư vấn tự động trợ giúp lựa chọn thông tin phù hợp trong vô số thông tin không phù hợp. Đối với các nhà cung cấp thông tin, hệ tư vấn 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 tư vấn 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 tư vấn sẽ sử dụng các phương pháp khuyến nghị 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 các kỹ thuật khuyến nghị. Chương 3 mô tả thực nghiệm 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.

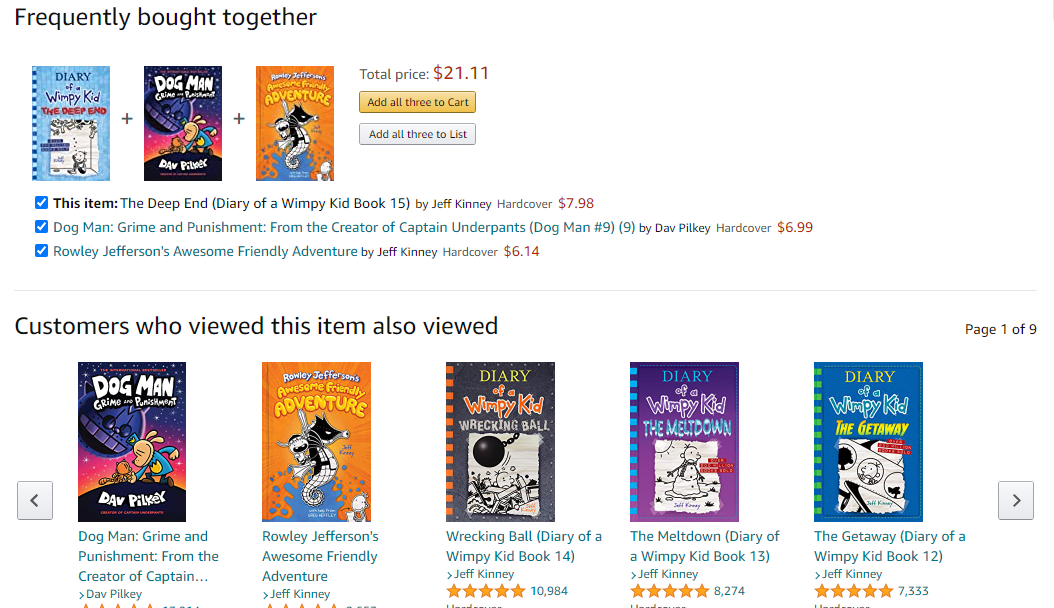
CHƯƠNG 1: GIỚI THIỆU CHUNG VỀ HỆ KHUYẾN NGHỊ VÀ HỆ KHUYẾN NGHỊ SÁCH TRONG BÁN HÀNG TRỰC TUYẾN

* 1. Giới thiệu hệ khuyến nghị – Recommender System
* Wikipedia định nghĩa, hệ thống khuyến nghị (Recommender System), 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…
* 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 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 gợi ý 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.



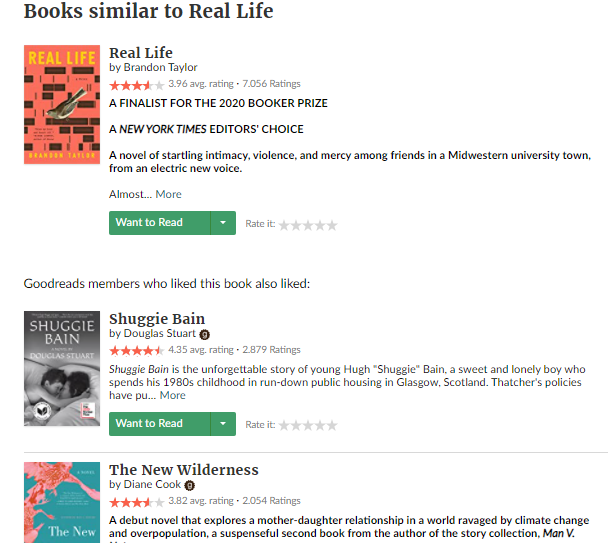
Hình 1: Các công ty tích hợp hệ gợi ý trong sản phẩm

* Hệ thống khuyến nghị sẽ đưa ra các gợi ý dựa trên quá trình thu thập, xử lý và phân tích dữ liệu của người dùng. Dữ liệu 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…).
  1. 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. Theo ghi nhận của tiki, đơn vị lớn nhất trong lĩnh vực bán sách online hiện nay cho thấy, trong 2 tháng đầu năm 2020, tỷ lệ tăng trưởng mặt hàng sách tăng 150% so với cùng kỳ năm ngoái. 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%.
* 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. 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 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.

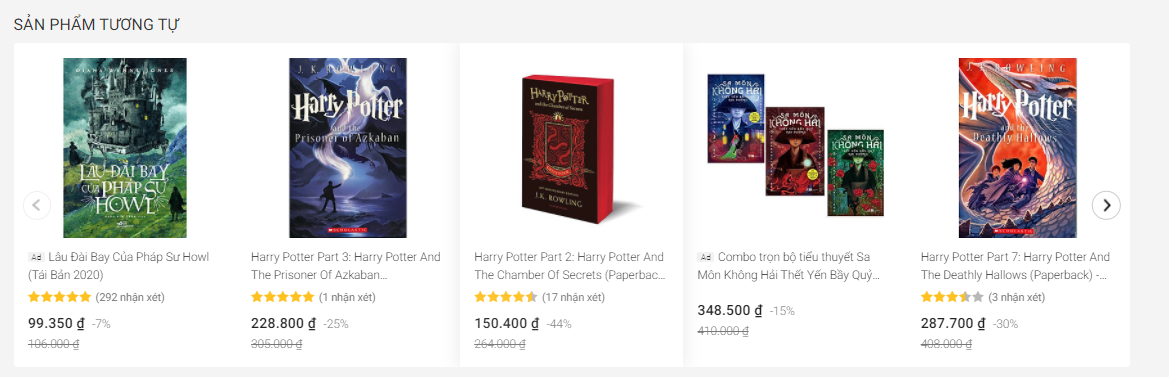


Hình 3: Minh họa về hệ gợi ý của Goodread

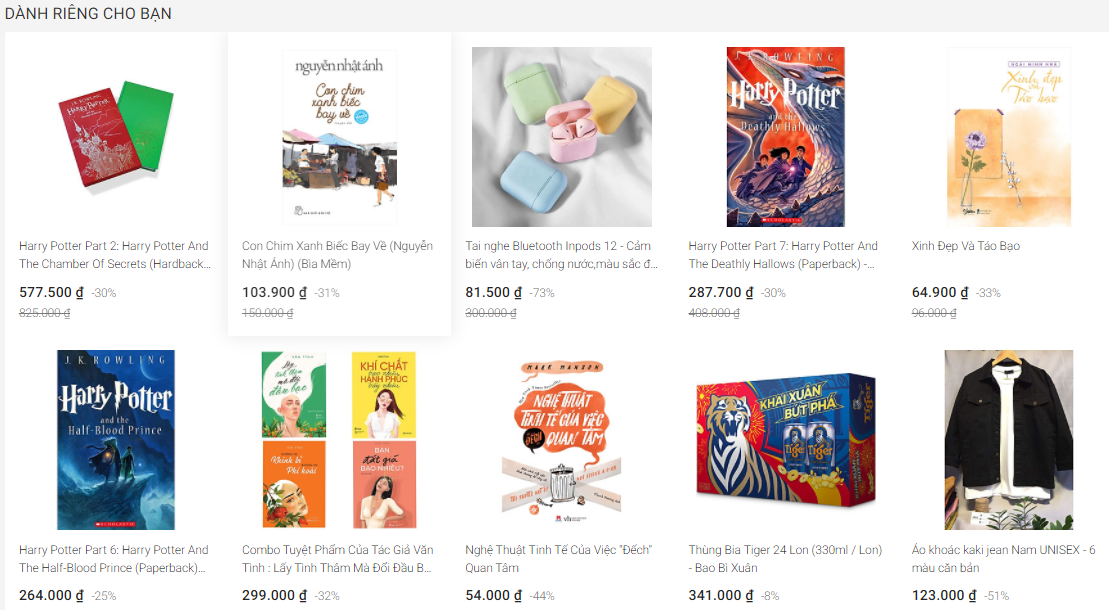
* Tại Việt Nam:
* Tiki.vn



Hình 4: Minh họa sản phẩm của Tiki

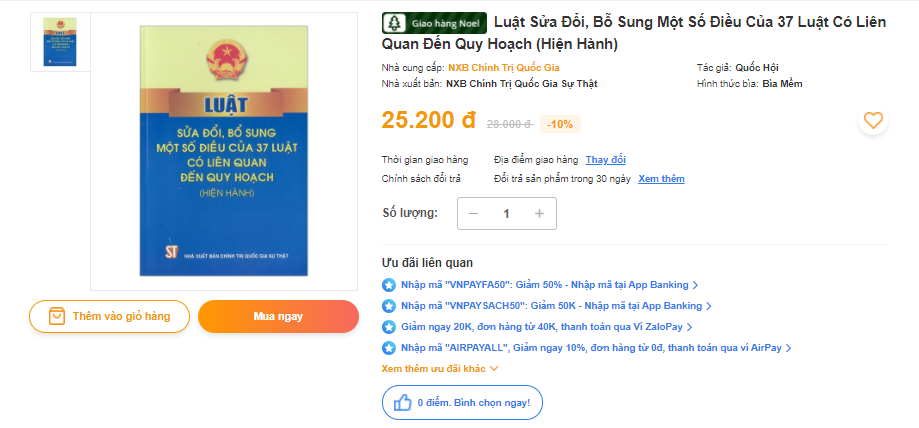


Hình 5: Minh họa sản phẩm gợi ý của Tiki - 1

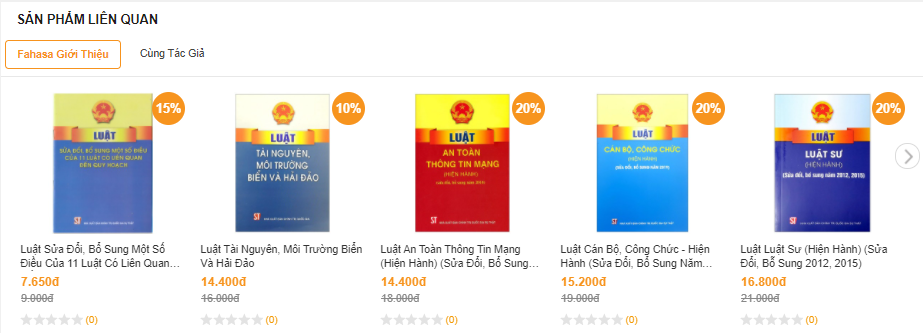


Hình 6: Minh họa sản phẩm gợi ý của Tiki - 2

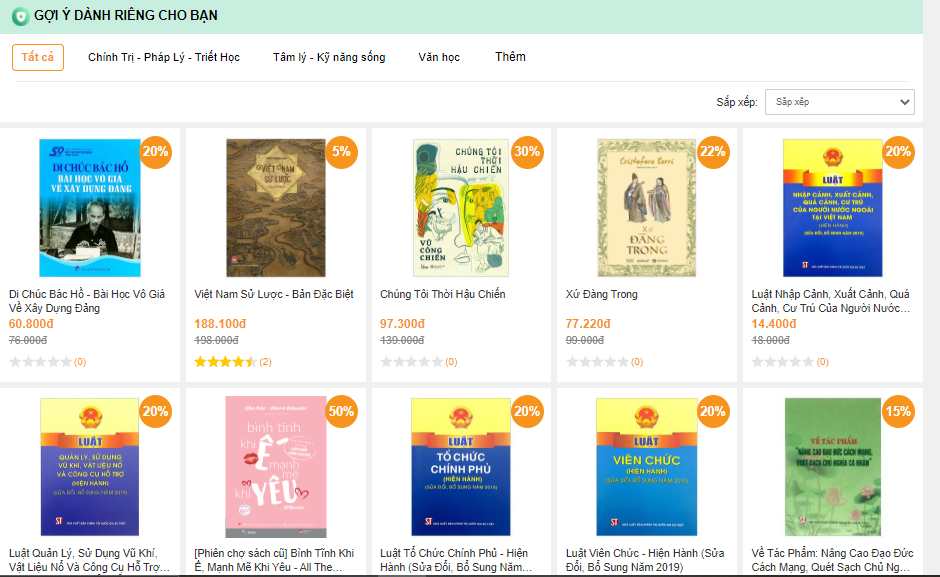
* Fahasa.com



Hình 7: Minh họa sản phẩm của Fahasa



Hình 8: Minh họa sản phẩm gợi ý của Fahasa - 1

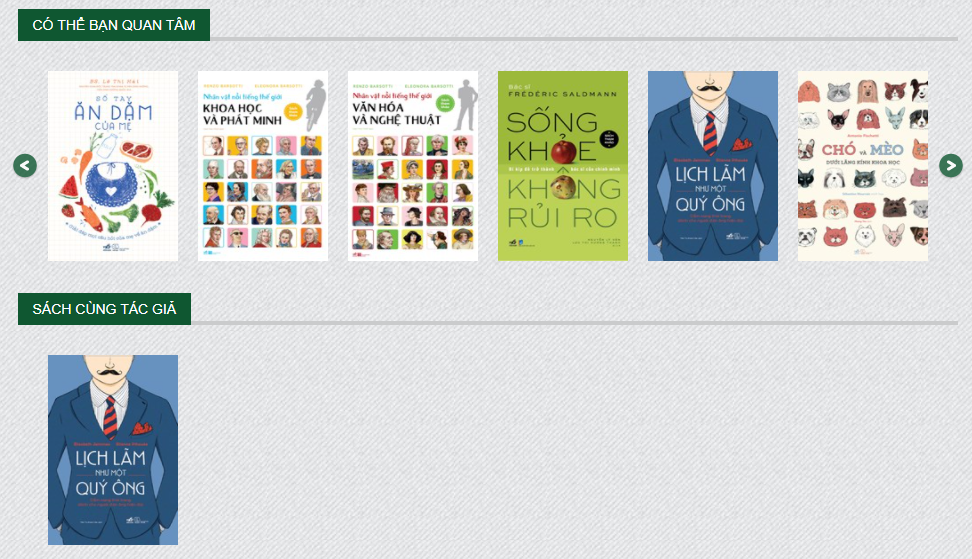


Hình 9: Minh họa sản phẩm gợi ý của Fahasa - 2

* Nhamam.com.vn



Hình 10: Minh họa sản phẩm của Nhã Nam



Hình 11: Minh họa sản phẩm gợi ý của Nhã Nam

* Các website 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.
  1. Phát biểu bài toán khuyến nghị
* Một cách hình thức, bài toán khuyến nghị hay bài toán tư vấn được phát biểu như sau:
* Gọi U là tập tất cả người dùng, I là tập tất cả các sản phẩm có thể tư vấn. Tập I có thể rất lớn, từ hàng trăm ngàn đến hàng triệu bản ghi.
* Hàm r(u, i) đo độ phù hợp hay hạng của sản phẩm i với người dùng u. Với mỗi người dùng u thuộc U, cần tìm sản phẩm i’ sao cho hàm r(u, i’) đạt giá trị lớn nhất:
* Trong hệ thống gợi ý, độ phù hợp của một sản phẩm thường được đo bằng điểm, ví dụ người dùng A đánh giá cuốn sách B được 4/5 sao. Tuy nhiên, độ phù hợp có thể là một hàm bất kỳ phụ thuộc vào ứng dụng cụ thể. Giá trị của hàm r có thể được xác định bởi người dùng hoặc tính toán bằng một công thức nào đó.
* Mỗi người dùng trong không gian U được xác định bởi 1 hồ sơ (profile). Hồ sơ này bao gồm thông tin người dùng (id, tên). Tương tự mỗi sản phẩm s trong không gian I cũng được xác định bởi 1 tập đặc trưng. Trong hệ thống gợi ý sách, đặc trưng của tập I có thể là: id, tên sách, mô tả nội dung, giá bán….
* Quá trình xây dựng hệ gợi ý bao gồm các bước sau:
* Tập hợp thông tin (Information Recollection)
* Lựa chọn, chuyển đổi, cấu trúc hóa thông tin
* Biểu diễn thông tin
* Hệ gợi ý được xây dựng dựa trên hai kỹ thuật lọc thông tin chính: Lọc theo nội dung (Content-Based Filtering) và lọc cộng tác (Collaborative Filtering). Lọc theo nội dung khai thác những khía cạnh liên quan đến nội dung thông tin sản phẩm người dùng đã từng truy cập trong quá khứ để tạo nên tư vấn. Trái lại, lọc cộng tác khai thác những khía cạnh liên quan đến thói quen sử dụng sản phẩm của cộng đồng người dùng có cùng sở thích để tạo nên tư vấn. Các kỹ thuật gợi ý sẽ được trình bày cụ thể trong các chương tiếp theo.
  1. Kết luận chương

Trong chương này, tôi đã trình bày một cách tổng quan về những lý thuyết cơ bản nhất của hệ khuyến nghị: khái niệm, vai trò, tầm quan trọng của hệ khuyến nghị trong đời sống cũng như một vài ví dụ thực tế về hệ khuyến nghị. Tôi cũng đưa ra mô tả về bài toán gợi ý và các bước xây dựng hệ gợi ý.

Trong chương tiếp theo, tôi sẽ trình bày chi tiết hơn 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: CÁC KỸ THUẬT KHUYẾN NGHỊ**

## Kỹ thuật khuyến nghị lọc dựa trên nội dung

### **Khái niệm**

* Lọc dựa trên 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.

### **Phát biểu bài toán tư vấn lọc theo nội dung**

* Cho là tập gồm n sản phẩm. Nội dung sản phẩm được ký hiệu là được biểu diễn thông qua tập K đặc trưng nội dung của I. Tập các đặc trưng sản phẩm I hay 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
* Cho là tập gồm m người dùng. Với mỗi người dùng , gọi 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 và hồ sơ người dùng

### **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) 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
* 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.

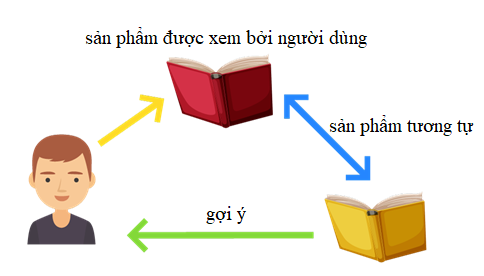
Khi tính tần suất xuất hiện TF, các từ đều được coi là quan trọng như nhau. Tuy nhiên, trong văn bản thường xuất hiện nhiều từ không quan trọng trong việc thể hiện ý nghĩa của văn bản như: từ nối (và, nhưng...), giới từ (trên, trong…), từ chỉ định (ấy, đó…). Do đó, ta cần giảm đi mức độ quan trọng của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

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

### **Các phương pháp lọc theo nội dung**

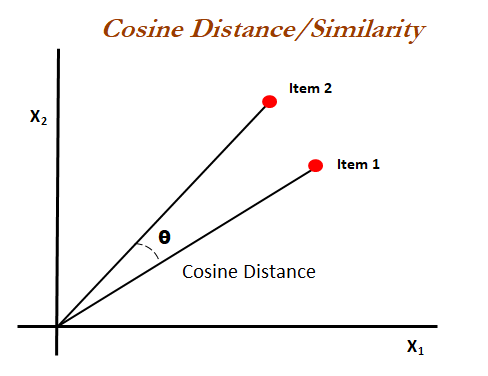
Lọc theo nội dung có thể tiếp cận theo 2 xu hướng: lọc dựa trên bộ nhớ và lọc dựa trên mô hình. Nội dung cụ thể các phương pháp được trình bày dưới đây:

* + - 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 12: 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:
* Xây dựng hồ sơ sản phẩm – item profiles
* Đo độ tương tự giữa các sản phẩm
* Bước 1 xây dựng hồ sơ sản phẩm đã được trình bày ở mục 2.1.3
* Ở 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.
* Công thức cosine tính góc giữa 2 vector



Hình 13: Minh họa khoảng cách cosine giữa hai vector

* 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.
  + - 1. *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 dưng 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

* + - * 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 (user) và sản phẩm (item). 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 dưới đây:

Bảng : 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 như đã trình bày ở mục 2.1.3

Bảng : 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 | x1 = [0.99, 0.52,..., 0.25, 0.05] |
| Mắt biếc | ? | 2 | ? | ? | 4 | ? | x2 = [0.12, 0.43,..., 0.15, 0.95] |
| Kính vạn hoa | 4 | ? | 3 | 2 | 4 | 4 | x3 = [0.64, 0.59,..., 0.31, 0.09] |
| Nghĩ giàu, làm giàu | ? | ? | ? | ? | ? | 5 | x4 = [0.82, 0.12,..., 0.43, 0.75] |
| Thám tử lừng danh Conan tập 1 | 5 | 1 | 3 | 3 | ? | 1 | x5 = [0.03, 0.02,..., 0.85, 0.55] |
| Mô hình người dùng |  |  |  |  |  |  |  |

* Bài toán đi tìm mô hình 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 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 lên chúng.
  + - * 1. 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 nếu sản phẩm thứ i đã được đánh giá bởi người dùng thứ j và ngược lại
* Mô hình tuyến tính: giả sử rằng ta có thể tìm ra được một mô hình cho mỗi người dùng, minh họa bởi vector cột hệ số và bias sao cho rating của một người dùng cho một sản phẩm có thể tính được bằng hàm tuyến tính:

Trong đó là một vector hàng, 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 , ta có thể xây dựng hàm mất mát theo Ridge Regression như sau:

Trong đó thành phần thứ hai là regularization term và là một số dương. Regularization thường không được áp dụng lên . Trong thực hành, trung bình cộng của lỗi thường được dùng, và mất mát được viết lại thành:

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

* Vì hàm mục tiêu chỉ phụ thuộc vào các sản phẩm đã được đánh giá, ta có thể rút gọn nó bằng cách đặt 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 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:

Trong đó là vector cột chứa phần tử 1.

* Mục tiêu của chúng ta là tìm cặp nghiệm , để tối ưu hàm mất mát.

### **Ưu điểm và nhược điểm của phương pháp lọc nội dung**

* Ưu điểm:
* Mô hình không cần bất cứ dữ liệu gì về người dùng khác, gợi ý riêng theo từng đối tượng người dùng. Do vậy, mô hình 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
* Nhược điểm:
* 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.
* 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 mô hình dự đoán, mô hình gặp vấn đề trong việc đưa ra gợi ý.

## 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 tư vấn lọc cộng tác và các kỹ thuật gợi ý của phương pháp này.

### **Khái niệm**

* Không giống với phương pháp lọc theo nội dung, phương pháp lọc cộng tác gợi ý dựa trên sự tương quan (similarity) giữa các người dùng và/hoặc sản phẩm. Hệ thống sẽ so sánh, tính toán độ tương tự nhay 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. Các hệ gợi ý dựa trên lọc cộng tác không yêu cầu quá nặng vào việc tính toán, do đó nó có thể đưa ra những gợi ý có độ chính xác cao và nhanh chóng cho một số lượng lớn người dùng. Hơn nữa, hệ gợi ý này không yêu cầu mô tả nội dung tường minh mà chỉ sự dụng đánh giá của người dùng để ước lượng, do đó những hệ này có khả năng gợi ý phong phú và thường tạo ra những gợi ý bất ngờ cho người dùng. Với phương pháp này, sở thích của người dùng là đầu vào duy nhất để quyết định kết quả gợi ý.
* 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.
* Tùy theo hệ thống mà đánh giá của người dùng được quy ước những giá trị nào. Trong ví dụ này, các đánh giá có giá trị từ 1->5.

Bảng : 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
* 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.
* Trong ví dụ trên, 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.

### **Phát biểu bài toán lọc cộng tác**

* Ký hiệu là tập gồm m người dùng, là tập gồm n sản phẩm mà người dùng có thể lựa chọn.
* Ký hiệu R = . Trong đó mỗi người dùng đưa ra đánh giá của mình cho một số sản phẩm bằng một số . Giá trị phản ánh mức độ ưa thích của người dùng đối với sản phẩm , giá trị 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 trong trường hợp người dùng chưa đánh giá hoặc chưa bao giờ biết đến sản phẩm
* Với một người dùng cần được gợi ý (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 đối với mặt hàng mà chưa đánh giá (), trên cơ sở đó gợi ý cho những sản phẩm được đánh giá cao.
* Bảng dưới thể hiện một ví dụ với ma trận đánh giá R trong hệ gồm 5 người dùng và 4 sản phẩm . Mỗi người dùng đều đưa ra ác đánh giá của mình về các sản phẩm theo thang bậc {0, 1, 2, 3, 4, 5}. Giá trị được hiểu là người dùng chưa đánh giá hoặc chưa bao giờ biết đến sản phẩm . Giá trị là sản phẩm hệ thống cần dự đoán cho người dùng

Bảng : Ma trận đánh giá R

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  | 2 | 1 | 3 | 5 |
|  | 4 | 2 | 1 | 0 |
|  | 3 | 0 | 2 | 4 |
|  | 4 | 4 | 0 | 0 |
|  | 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 (active user) 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 mỗi liên hệ trên ma trận đánh giá người dùng.

### **Phương pháp tư vấn 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 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 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 trưng 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)
  + - 1. Phương pháp lọc cộng tác dựa trên người dùng



Hình 15: Minh họa phương pháp lọc cộng tác dựa trên người dùng

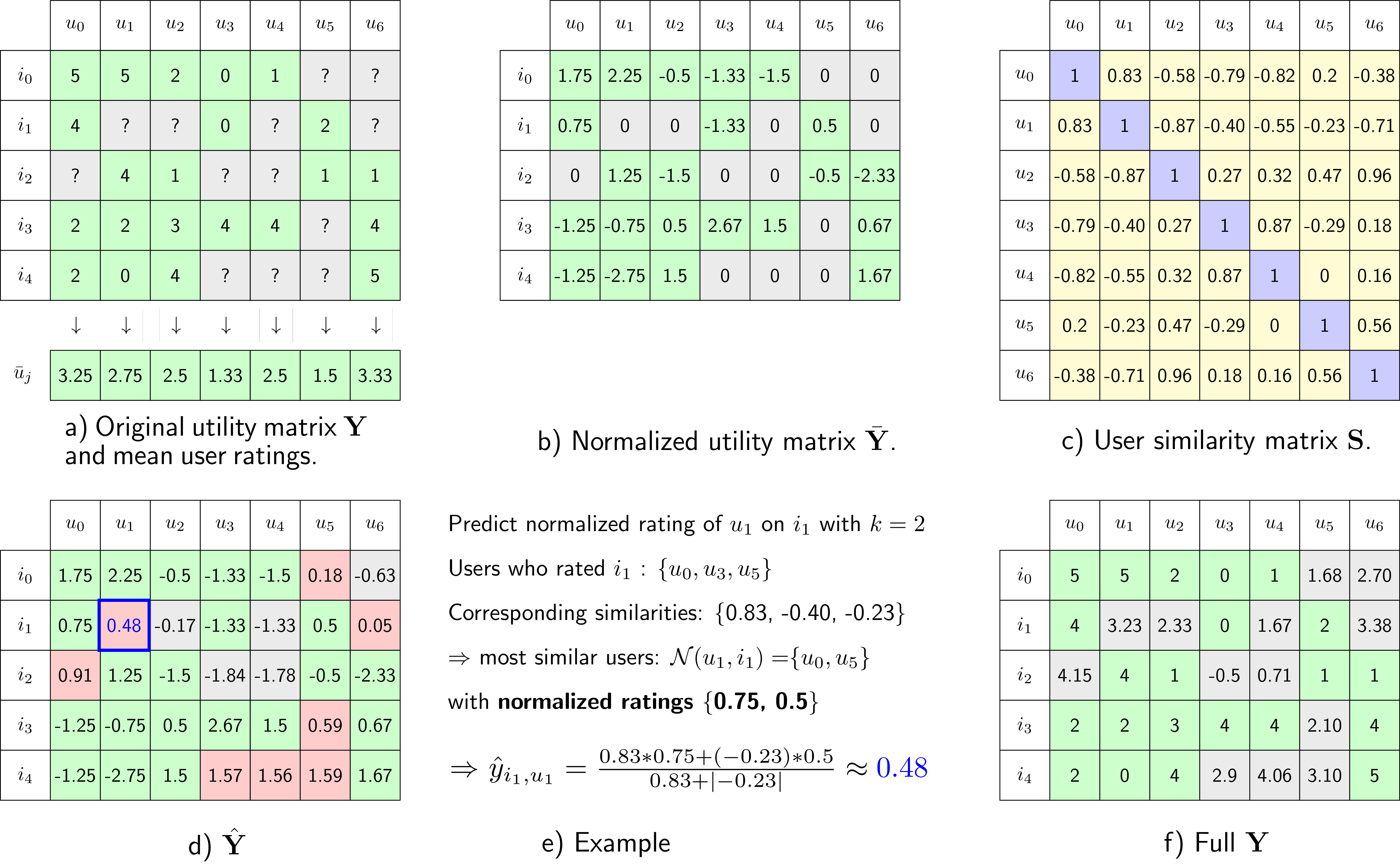
* Chuẩn hóa ma trận Utility matrix:
* Để có thể sử dụng ma trận này vào việc tính toán, chúng ta cần thay những dấu ‘?’ bởi một giá trị. Đơn giản nhất thì chúng ta có thể thay vào đó giá trị 0 hoặc giá trị trung bình ratings là 2.5. Tuy nhiên, những giá trị này sẽ hạn chế với cách đánh giá của mỗi người 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 20-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 sparce 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 20-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:

Trong đó là tập tất cả các sản phẩm người dùng x và y cùng đánh giá , 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 = là số lượng các sản phẩm cả hai người dùng cùng đánh giá. *(Hình 20-c)*
* 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:

Trong đó là tập k người dùng có độ tương đồng cao nhất với người dùng u và đã từng đánh giá sản phẩm i *(Hình 20-d, 20-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 20-f)

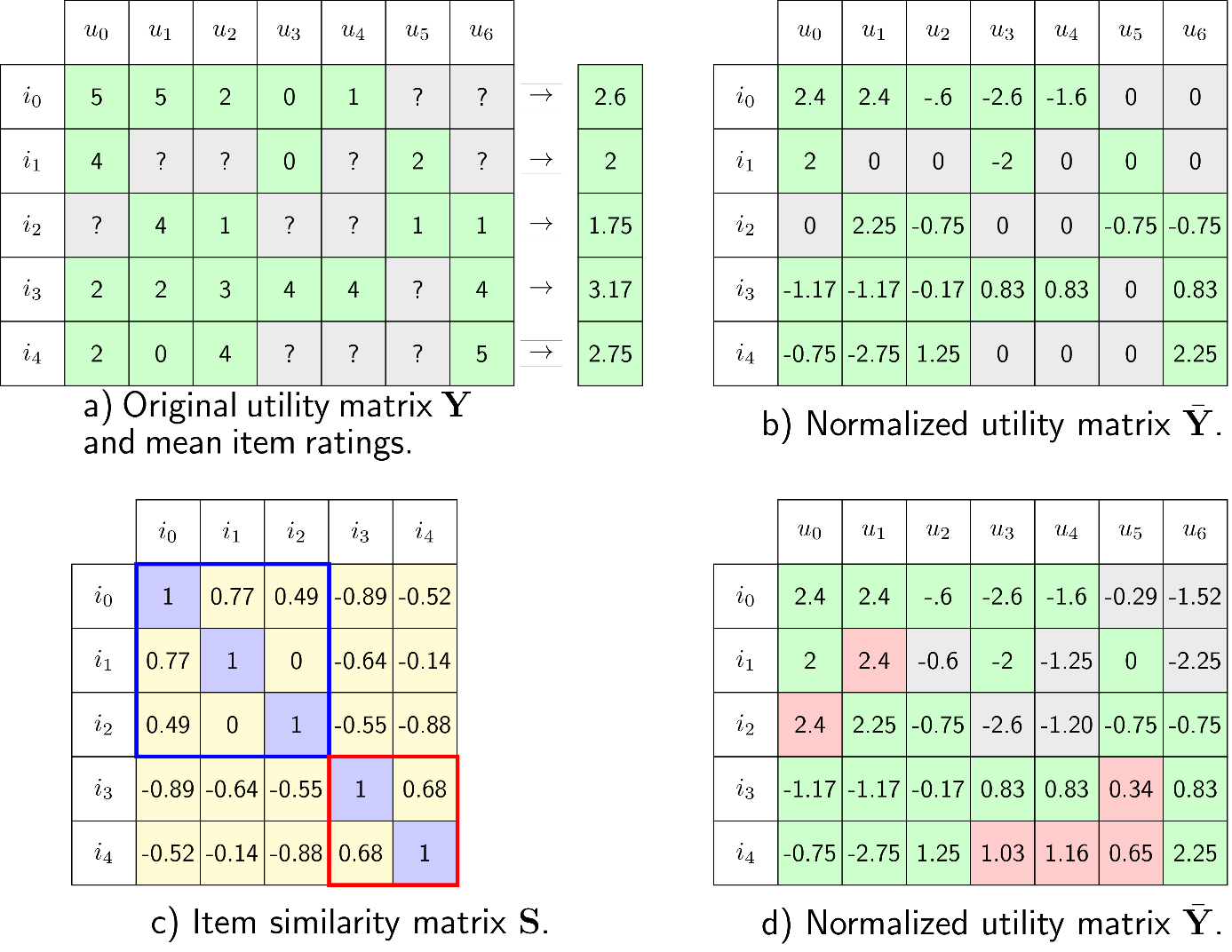


Hình : Minh họa chuẩn hóa ma trận

* + - 1. 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:

Trong đó là tập tất cả người dùng cùng đánh giá sản phẩm x và y, , là đánh giá trung bình cho sản phẩm x và sản phẩm y

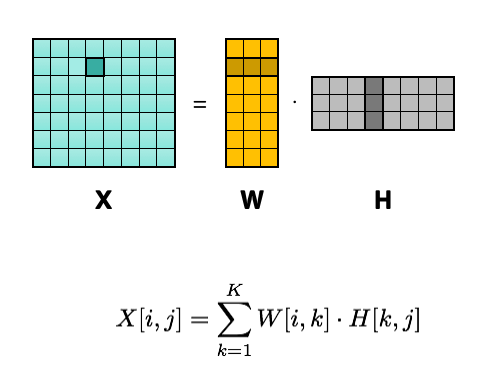
* Độ tương tự vector giữa hai sản phẩm x, y là cosine của hai vector x và y theo công thức dưới đây. Trong đó, hai sản phẩm x và y được xem xét như hai vector cột n chiều, n = là số lượng các người dùng đánh giá cả 2 sản phẩm x và y.
* Dự đoán đánh giá của người dùng u cho sản phẩm i, ta thực hiện:
* Tìm tập các sản phẩm mà người dùng u đã đánh giá
* Tính độ tương tự của sản phẩm i với các sản phẩm trong tập . Chọn ra k sản phẩm có độ tương đồng cao nhất với i
* Tính ratings theo công thức:
* 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.



Hình : Minh họa chuẩn hóa utility matrix

### **Phương pháp tư vấn dựa 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, sao cho ta có thể xây dựng lại X từ hai ma trận nhỏ hơn này càng chính xác càng tốt, nghĩa là



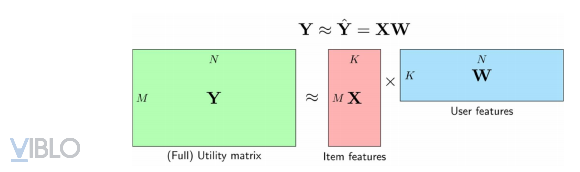
Hình 22: Kỹ thuật phân rã ma trận

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

=

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

* 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 bằng tích của hai ma trận con là và



Hình 23: Bài toán tối ưu ma trận

* Xây dựng hàm mất mát:

Trong đó nếu sản phẩm thứ m đã được đánh giá bởi người dùng thứ n, là Frobineous 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 lorm 2 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:
* Khi cố định X, việc tối ưu W chính là bài toán tối ưu trong lọc cộng tác dựa trên mô hình
* Khi cố định W, việc tối ưu X được đưa về tối ưu hàm
* Hai bài toán trên được tối ưu bằng phương pháp Gradient Descent.
* 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
* Vì biểu thức trong dấu  chỉ phụ thuộc vào các sản phẩm đã được đánh giá bởi người dùng đang xét, ta có thể đơn giản nó bằng cách đặt  là ma trận được tạo bởi các hàng tương ứng với các sản phẩm đã được đánh giá đó, và   là các ratings tương ứng. Khi đó:

Và đạo hàm của nó

* Vậy công thức cập nhật cho mỗi cột của W là:
* Tương tự như thế, mỗi cột của X, tức vector cho mỗi sản phẩm, sẽ được tính bằng cách tối ưu:
* Đặt là ma trận được tạo bằng các cột của W ứng với các người dùng đã đánh giá sản phẩm đó và là vector ratings tương ứng
* Tương tự như trên, công thức cập nhật cho mỗi hàng của X có dạng:
* Sau mỗi lần lặp, điểm hiện tại sẽ giảm một lượng tùy thuộc vào tốc độ và lặp đến khi có được điểm cực tiểu chính là điểm cần tìm để hàm mất mát có giá trị nhỏ nhất.
  + 1. **Ưu điểm và nhược điểm của phương pháp lọc cộng tác**
* *Ưu điểm*
* 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, chính vì thế các tính chất của đối tượng được gợi ý không có ảnh hưởng đến quá trình gợi ý
* 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*
* 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 tư vấn 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 tư vấn 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.
  1. Tổng 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.
* Chương 3 sẽ mô tả thực nghiệm 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

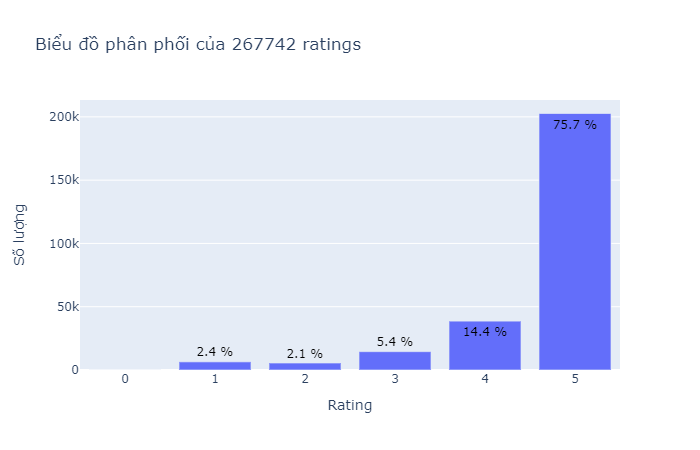
Phương pháp thực nghiệm được mô tả theo các pha dưới đây:

* 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

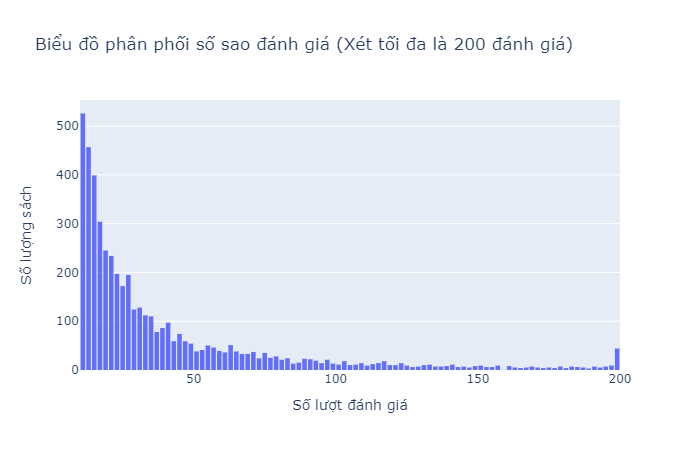


Hình 24: Các bước xây dựng mô hình khuyến nghị

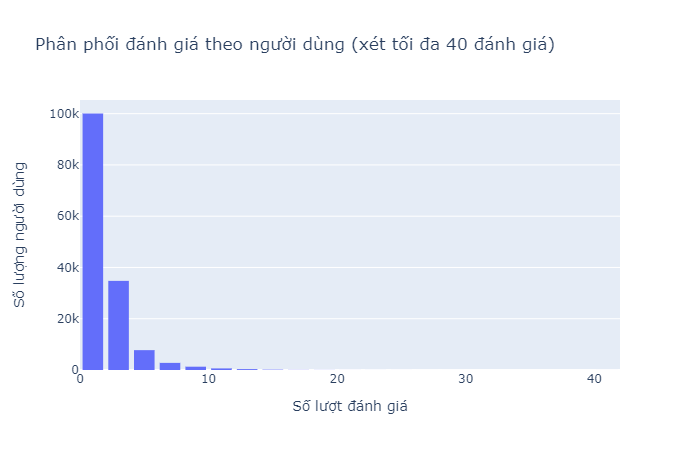
* 1. Thu thập dữ liệu
* Dữ liệu là một thành phần vô cùng quan trọng trong các hệ gợi ý. Dữ liệu và nguồn tri thức có sẵn cho các hệ gợi ý có thể rất đa dạng. Tuy nhiên, trong bất kỳ trường hợp nào, dữ liệu được sử dụng bởi hệ gợi ý thuộc ba loại: sản phẩm (Item), người dùng (User) và tương tác giữa 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 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 : Biểu đồ phân phối đánh giá (dữ liệu thô)

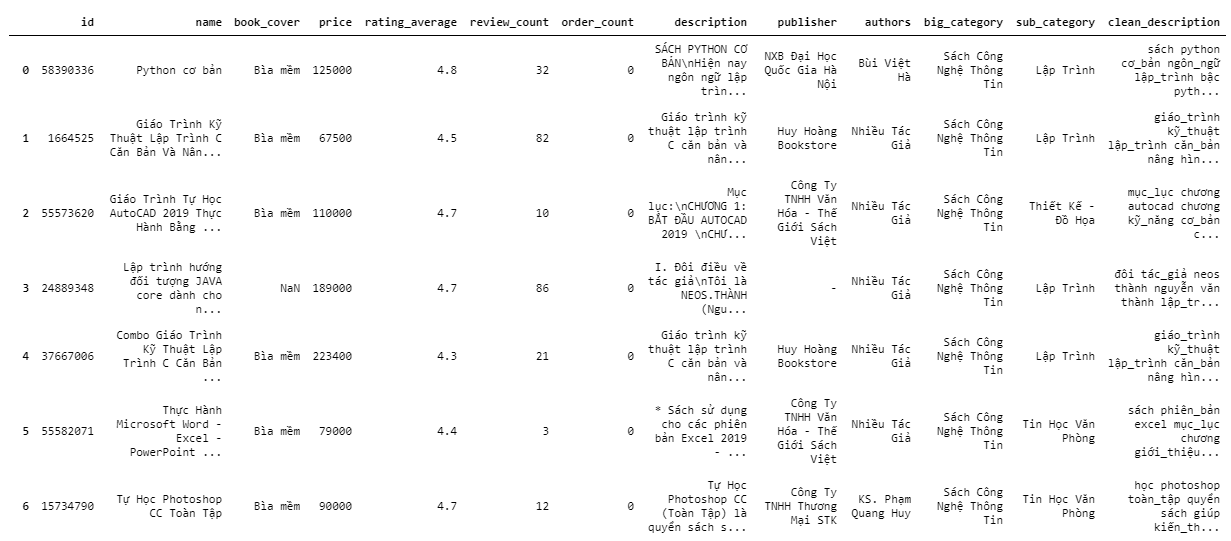


Hình : Biểu đồ phân phối số sao đánh giá theo sản phẩm (dữ liệu thô)



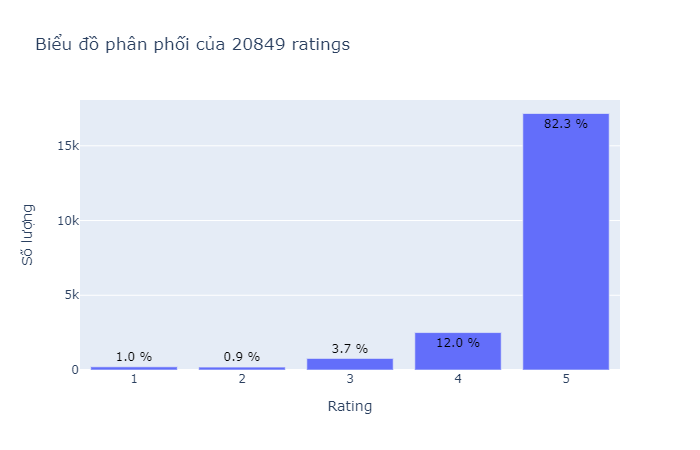
Hình : Biểu đồ phân phối số sao đánh giá theo người dùng (dữ liệu thô)

* 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

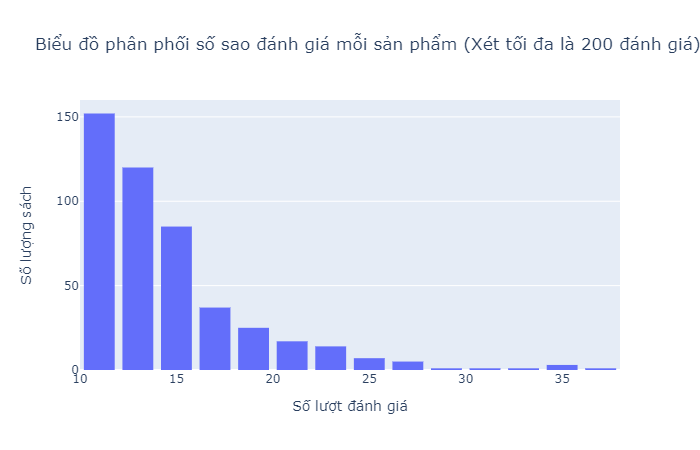


Hình 28: Một vài mẫu dữ liệu sách

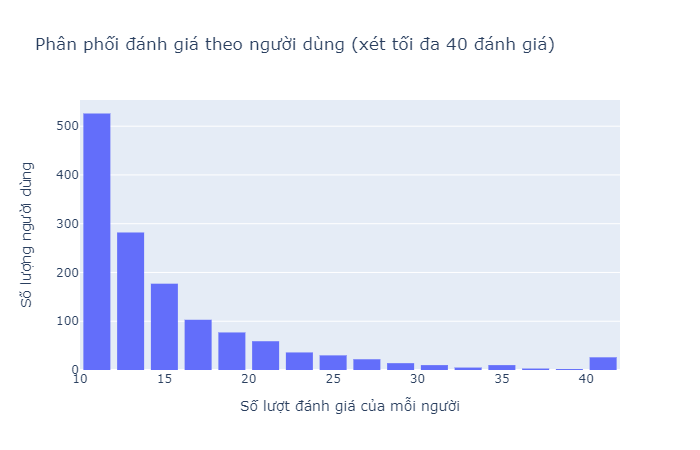
* Qua dữ liệu đánh giá, đa phần người dùng chỉ đánh giá dưới 10 quyển sách. Để tránh vấn đề cold start của dữ liệu người dùng, cũng như đảm bảo tính gợi ý của mô hình, tôi chỉ giữ lại những người dùng có đánh giá tối thiểu là 10 quyển sách.



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

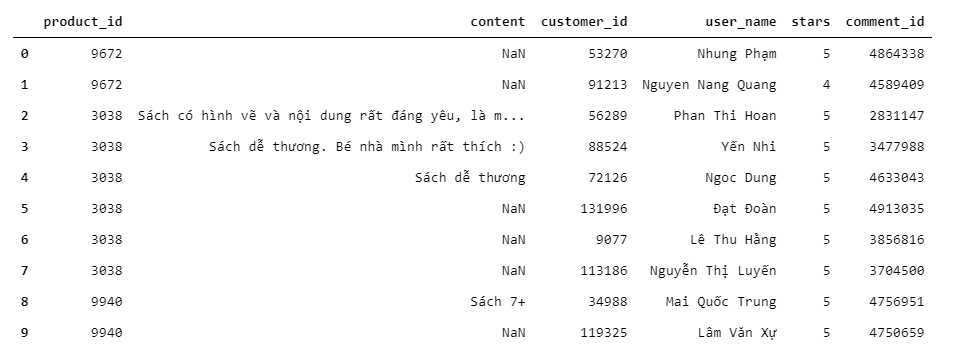


Hình : Biểu đồ phân phối số sao đánh giá theo sản phẩm (dữ liệu sau khi xử lý)



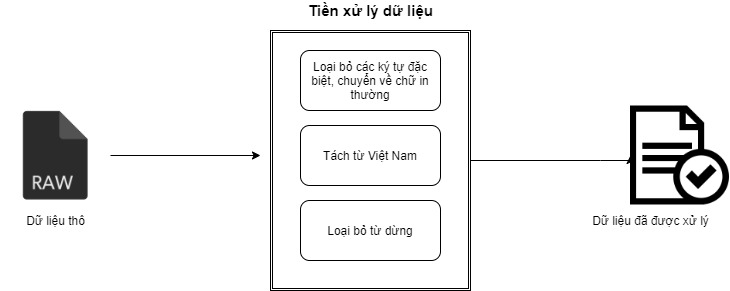
Hình : Biểu đồ phân phối số sao đánh giá theo người dùng (dữ liệu sau khi xử lý)

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



Hình 32: Dữ liệu đánh giá của ngưởi dùng

* 1. Xử lý dữ liệu
     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: phần mô tả sách, còn tồn tại các ký tự đặc biệt, thẻ html, từ dừng
* Quá trình tiền xử lý dữ liệu bao gồm:
* Làm sạch dữ liệu bằng cách loại bỏ một số dấu câu, kí 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

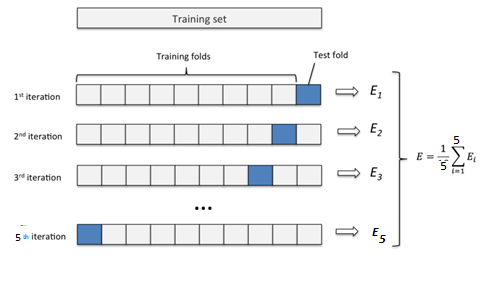


Hình 33: Tiền xử lý dữ liệu văn bản

* 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 |

* + 1. 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 lí 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 mô tả như sau:



Hình : Minh họa cách chia dữ liệu

* 1. Thực nghiệm mô hình tư vấn
     1. Thực nghiệm mô hình tư vấn sách theo kỹ thuật lọc dựa trên nội dung

Công cụ thực nghiệm: thư viện sklearn (hay scikit-learn) – 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

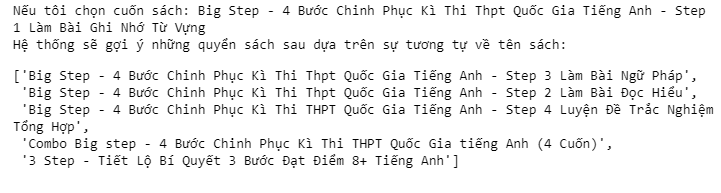
* Xây dựng hồ sơ sản phẩm: Chuyển đổi từ sang vector đặc trưng TF-IDF
* Sử dụng gói *sklearn.feature\_extraction.text.TfidfVectorizer*
* Tham số sử dụng mặc định.
* Sau khi vector hóa, thu được ma trận hồ sơ sản phẩm có kích thước 10320x7193. Trong đó 10320 tương ứng với 10320 sản phẩm, 7193 là kích thước từ điển trích từ bộ dữ liệu.
  + - 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,* 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 để gợi ý.
  + - 1. 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.
  + 1. Thực nghiệm mô hình tư vấn sách theo kỹ thuật lọc cộng tác
       1. Công cụ thực nghiệm
* Giới thiệu thư viện sciki-surprise (<http://surpriselib.com/>): 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
  + - 1. Lọc cộng tác dựa trên bộ nhớ
* Mô hình sử dụng: KNNBasic
  + - 1. Lọc cộng tác bằng phương pháp matrix factorization
* Qua quá trính 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 bộ tham số

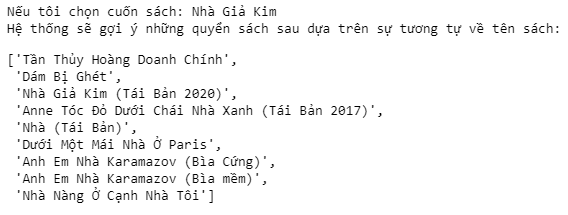
|  |  |  |
| --- | --- | --- |
| Tham số | Ý nghĩa | Giá trị tối ưu |
| n\_factors | Số lượng các nhân tử | 25 |
| n\_epochs | Số lần lặp đào tạo | 20 |
| lr\_all | Tốc độ học cho các tham sô | 0.008 |
| reg\_all | Regularization cho các tham số | 0.08 |

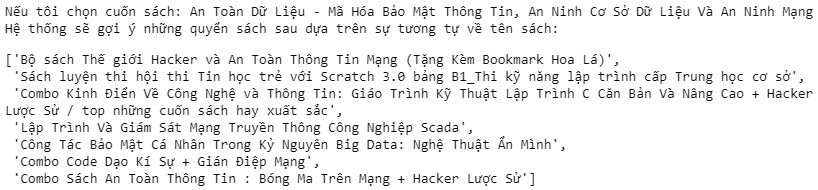
* 1. Đánh giá thuật toán
     1. Các thông số đánh giá giải thuật

Trong phạm vi đồ án sẽ sử dụng các tiêu chí định lượng nhằm đánh giá số lượng các gợi ý liên quan.

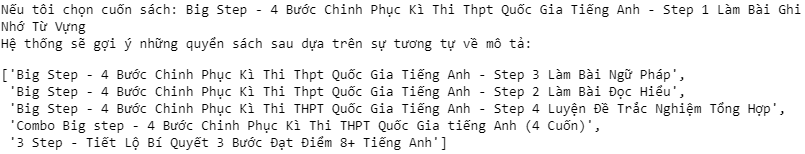
* + - 1. Các tiêu chí đánh giá độ chính xác của các dự đoán
* Việc đánh giá tính chính xác các dự đoán có thể sử dụng sai số bình phương trung bình (Mean Square Error – MSE), căn của sai số bình phương trung bình (Root Mean Square Error – RMSE), sai số tuyệt đối trung bình (Mean Absolute Error – MAE). Tính chính xác của dự đoán được đo trên n quan sát, trong đó là giá trị dự đoán đánh giá của sản phẩm i và là giá trị đánh giá thực tế của sản phẩm i
* 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.
  + - 1. 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.
* 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.
* 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
* Giá trị của các độ đo này càng cao thì độ chính xác của phương pháp càng cao.
  + 1. **Kết quả thực nghiệm**
       1. Mô hình gợi ý theo lọc nội dung
* *Lọc nội dung dựa trên bộ nhớ*
* Gợi ý dựa trên tương đồng về tên sách

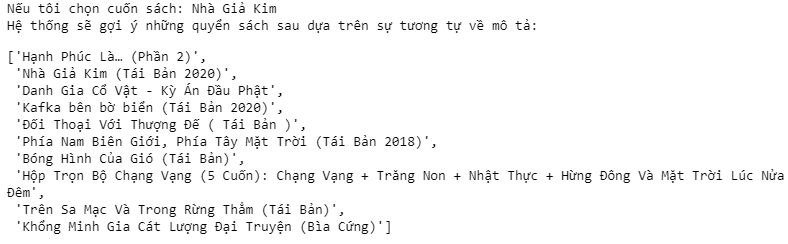


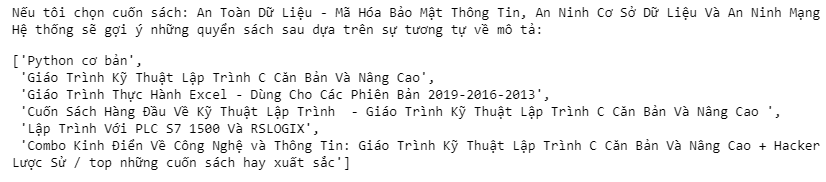




* Gợi ý dựa trên tương đồng về nội dung mô tả sách





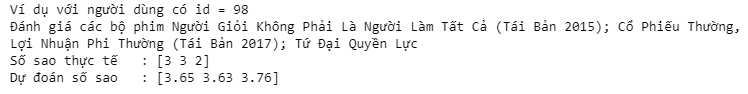


* *Lọc nội dung dựa trên mô hình*
* Kết quả thử nghiệm độ đo RMSE trên bộ dữ liệu huấn luyện và bộ dữ liệu kiểm thử

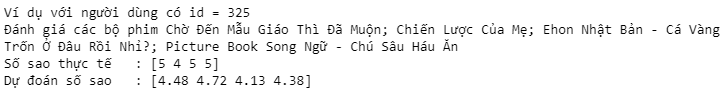
|  |  |
| --- | --- |
|  | RMSE |
| Tập huấn luyện | 0.02533 |
| Tập kiểm thử | 0.53477 |

Minh họa một số gợi ý:

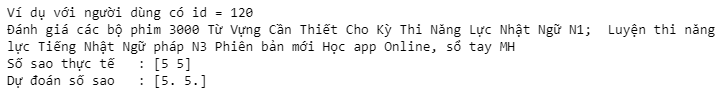
* Ví dụ userid = 98



* Ví dụ userid = 325



* Ví dụ userid = 120



* + - 1. Mô hình gợi ý theo lọc cộng tác
* Lọc cộng tác theo bộ nhớ
* Kêt quả lọc cộng tác dựa trên sản phẩm trên tập huấn luyện (Item Based)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 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 |

* Kết quả lọc cộng tác dựa trên người dùng trên tập huấn luyện (Item Based)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 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. Dưới đây là kết quả đo được tương ứng với các giá trị ngưỡng.

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

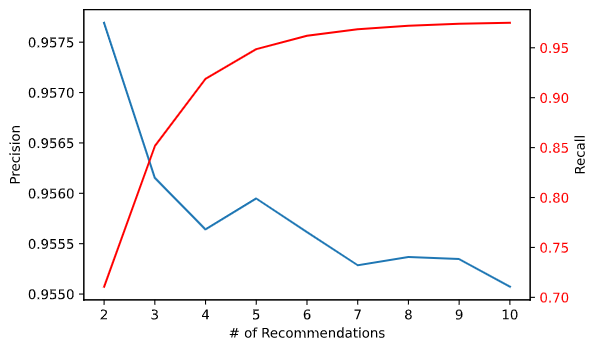
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| threshold | TP | FP | TN | FN | Precision | Recall | F1 |
| 0.0 | 4170 | 0 | 0 | 0 | 1.000000 | 1.000000 | 1.000000 |
| 0.5 | 4170 | 0 | 0 | 0 | 1.000000 | 1.000000 | 1.000000 |
| 1.0 | 4170 | 0 | 0 | 0 | 1.000000 | 1.000000 | 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.

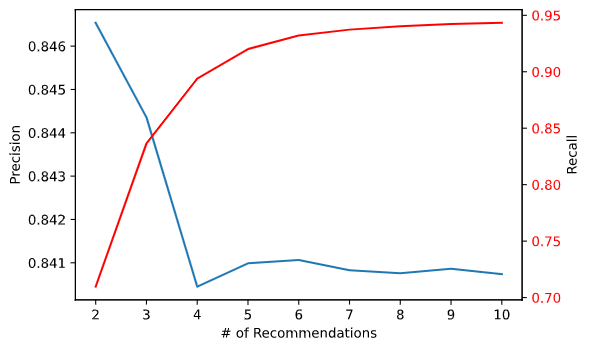
Bảng : Kết quả đo precision và recall theo k bằng phương pháp lọc theo bộ nhớ

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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



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



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

* Lọc cộng tác sử dụng matrix factorization

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | RMSE | MAE | Fit time | Test time |
| SVD | 0.5795 | 0.3383 | 1.528 | 0.112 |

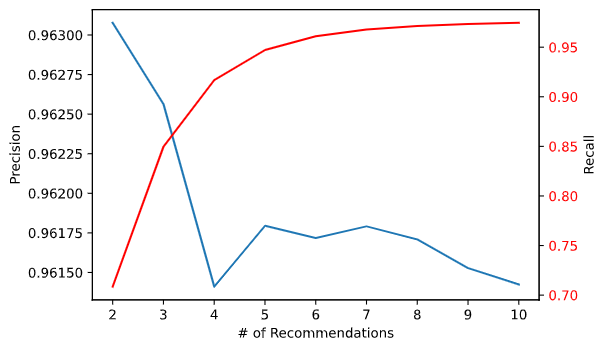
Bảng : Kết quả đo precision và recall theo các ngưỡng bằng phương pháp lọc phân rã ma trận

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 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 |

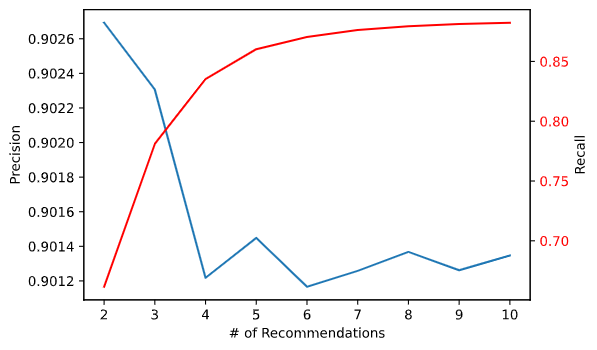
Bảng : Kết quả đo precision và recall theo k bằng phương pháp lọc phân rã ma trận

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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



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



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

* + 1. **Nhận xét kết quả và phân tích lỗi trong quá trình thực nghiệm**
* Qua thử nghiệm, ta thu được các kết quả tốt nhất sau

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

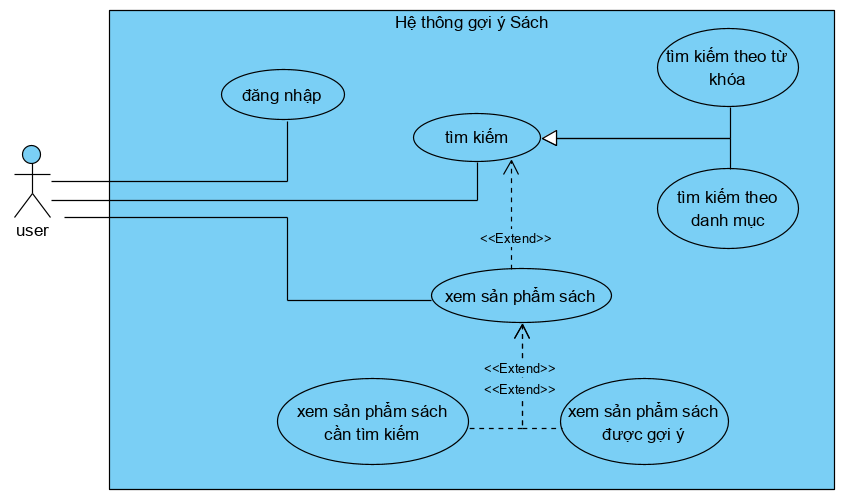
* Phương pháp gợi ý theo lọc nội dung cho kết quả thực nghiệm tốt nhất với bộ dữ liệu sách và dữ liệu đánh giá thu được
* Các phương pháp lọc cộng tác thường cho kết quả với phương pháp phân rã ma trận tốt hơn.
* Tuy nhiên, do bộ dữ liệu chủ yếu với đánh giá = 5, dữ liệu còn thưa thớt nên các mô hình gợi ý chưa thể cho kết quả tối ưu nhất.
  1. Kết luận 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.

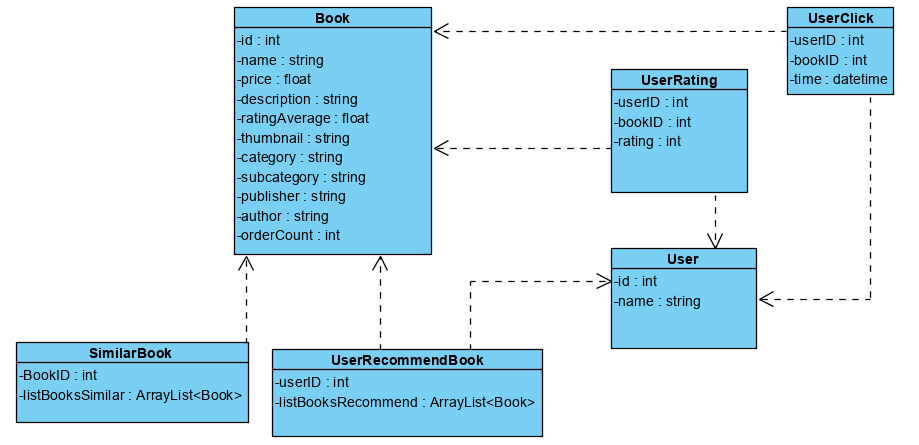
Ở chương cuối, tôi sẽ trình bày cách xây dựng ứng dụng web gợi ý sách ứng dụng những mô hình gợi ý ở chương 3, nhằm mục đích mô phỏng những kết quả và ưu điểm của hệ gợi ý.

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

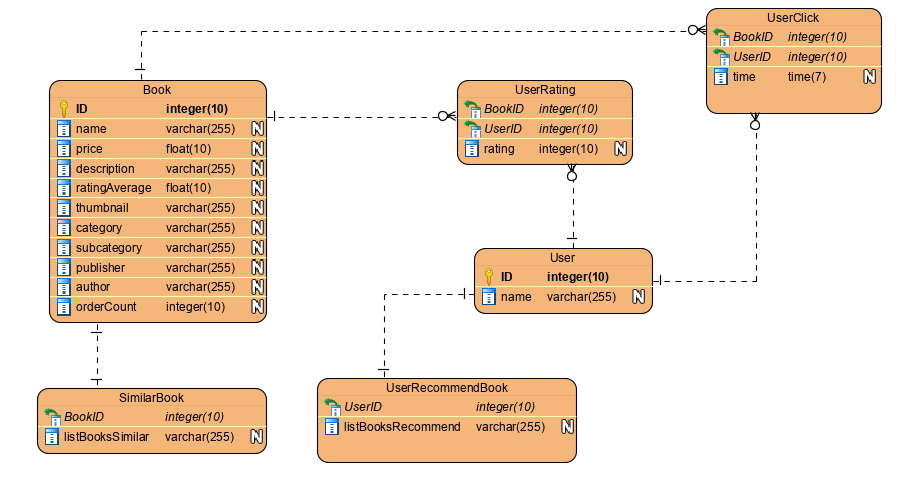
* 1. Mô tả hệ thống
* Trang web được xây dựng với mục đích tìm hiểu, nghiên cứu hoạt động của hệ khuyến nghị
* Hệ thống giới thiệu và gợi ý sách cho phép khách hàng bất kỳ có thể tìm kiếm hoặc xem các sản phẩm theo thể loại, danh sách sách được ưa thích nhất …
* Hệ thống cho phép người dùng đăng nhập qua Facebook
* 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.
  1. Phân tích thiết kế hệ thống



Hình : Sơ đồ usecase toàn hệ thống

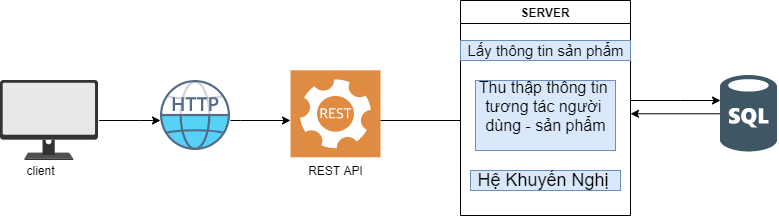


Hình : Biểu đồ lớp toàn hệ thống



Hình : Lược đồ cơ sở dữ liệu

* 1. Thiết kế hệ thống
     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
  + 1. **Mô hình tổng quát của hệ thống**



* Kiến trúc tổng quát của hệ thống 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.
* 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
  + 1. **Triển khai ứng dụng thực tế**
* Một số giao diện minh họa của hệ thống
* Trang chủ
* Trang hiển thị sách theo phân loại.
* Trang chi tiết sản phẩm:

# **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 tư vấn.
* **Xây dựng ứng dụng hệ tư vấn 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 tư vấn 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:**

* Thử nghiệm bộ dữ liệu với các mô hình mạng nơ ron
* Phát triển web gợi ý sách thành website bán hàng
* Quan tâm đến các tính khác: tính đa dạng, novelty…

# **DANH MỤC TÀI LIỆU THAM KHẢO**