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

ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC

MÔ HÌNH KẾT HỢP HÀNH VI ĐÁNH GIÁ VÀ BÌNH LUẬN CHO TƯ VẤN KHÁCH SẠN

VŨ QUANG SƠN

KỸ SƯ NGÀNH HỆ THỐNG THÔNG TIN

HÀ NỘI, NĂM 2021

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

ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC

MÔ HÌNH KẾT HỢP HÀNH VI ĐÁNH GIÁ VÀ BÌNH LUẬN CHO TƯ VẤN KHÁCH SẠN

VŨ QUANG SƠN - B17DCCN545

GIẢNG VIÊN HƯỚNG DẪN
TRẦN ĐÌNH QUẾ

HÀ NỘI, NĂM 2021

Lời cảm ơn

Lời đầu tiên, em xin gửi lời cảm ơn chân thành tới tất cả thầy cô đang giảng dạy trong mái 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 khoảng thời gian hơn 4 năm là sinh viên của học viện. Em xin gửi lời biết ơn sâu sắc đến thầy PGS.TS Trần Đình Quế, người đã tận tình hướng dẫn, chỉ bảo, định hướng và nhắc nhở em trong suốt quá trình học tập cũng như hoàn thành đồ án này.

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

Cuối cùng, cho tôi gửi lời cảm ơn đến những người bạn, người anh, người chị của tôi, những người luôn chia sẻ, động viên, giúp đỡ và ở bên tôi mỗi khi tôi gặp khó khăn nhất!

Em xin chân thành cảm ơn!

Hà Nội, ngày 10 tháng 12 năm 2021

Sinh viên thực hiện

Vũ Quang Sơn

Mục lục

Danh sách hình vẽ

Danh sách bảng

Danh sách đoạn mã

Danh mục từ viết tắt và tạm dịch

Từ viết tắt	**Tiếng Anh**	**Tạm dịch**
CF	Collaborative Filtering	Lọc cộng tác
SVD	Singular Value Decomposition	Phân tích giá trị đơn vị đặc biệt
PCA	Principal Component Analysis	Phân tích thành phần chính
MBCF	Memory-based Collaborative Filtering	Lọc cộng tác dựa trên bộ nhớ
UBCF	User-based Collaborative Filtering	Lọc cộng tác dựa trên người dùng
IBCF	Item-based Collaborative Filtering	Lọc cộng tác dựa trên sản phẩm
KNN	K-Nearest Neighbors	K láng giềng gần nhất

Danh mục từ tạm dịch

Machine Learning	Học máy
Deep Learning	Học sâu
Reinforcement learning	Học tăng cường
Federated Learning	Học liên kết

Mở đầu

Cuộc sống của con người ngày càng phát triển, các nhu cầu cá nhân như: giao lưu, kết bạn, tiêu dùng, du lịch, ... ngày tăng. Nhu cầu tiêu dùng ngày càng tăng, cùng với sự phát triển của công nghệ thống tin, các hệ thống thương mại điện tử ra đời và ngày càng lớn mạnh, tiêu biểu như: Facebook, Youtube, Tripadvisor, ... Những trang thương mại điện tử này hỗ trợ doanh nghiệp quảng bá sản phẩm tới tay người tiêu dùng nhanh hơn so với bán hàng truyền thống. Tuy nhiên, khi người dùng được tiếp cận sản phẩm, dịch vụ một cách nhanh chóng thì họ cũng phải đối mặt với vấn đề có quá nhiều sản phẩm và dịch vụ và đâu thực sự là sản phẩm họ cần. Đây là tình trạng quá tải thông tin, khi người dùng có quá nhiều lựa chọn. Tuy nhiên, đôi khi họ cũng phải đối mặt với tình huống nghịch lý rằng có rất nhiều thông tin, nhưng thường rất khó để có thông tin phù hợp [3]. Với hiện trạng nêu trên, nhu cầu cấp thiết đặt ra cần có các hệ thống tự động hóa, hỗ trợ người dùng lọc thông tin cũng như cá nhân hóa đối với từng người dùng.

Hệ tư vấn ra đời nhằm giải quyết vấn đề quá tải thông tin từ người dùng, giúp họ khám phá những sản phẩm khác nhau nằm trong sở thích của mình. Có rất nhiều trang thương mại điện tử lớn sử dụng hệ tư vấn nhằm cải thiện doanh thu và tăng sự thân thiện với người dùng, một trong số đó là Youtube. Youtube, ra đời vào tháng 2, 2005 với sự phát triển nhanh chóng đã trở thành nền tảng chia sẻ video trực tuyến lớn nhất hiện nay với hơn 1 tỷ lượt xem mỗi ngày từ hàng triệu người dùng và mỗi phút có hơn 24 giờ thời lượng video được tải lên nền tảng này. Hệ tư vấn là một phần trong sự thành công của Youtube khi đóng góp 60% lượt bấm xem

video từ trang chủ và các video được gợi ý từ hệ thống có tỷ lệ bấm xem gấp 2 lần những video được nhiều người xem nhất và được đánh giá cao nhất [2].

Một trong các thuật toán tư vấn điển hình và phổ biến là lọc cộng tác và hoạt động rất hiệu quả. Các hệ tư vấn truyền thống thường sử dụng dữ liệu điểm đánh giá để làm cơ sở tư vấn. Tuy nhiên, theo [6], thuật toán này vẫn còn những vấn đề còn tồn tại như:

- Vấn đề người dùng mới, sản phẩm mới (Cold Start)
- Vấn đề thưa thớt dữ liệu

Do thói quen lười đánh giá từ người dùng, gây ra những vấn đề trên ảnh hưởng tới độ chính xác của hệ tư vấn lọc cộng tác.

Với sự bùng nổ của các trang thương mại điện tử, các hành vi bày tỏ quan điểm ngày càng đa dạng và phong phú. Do đó, các phương pháp phân loại văn bản ngày càng được cải thiện và trở nên chính xác hơn. Những dữ liệu văn bản này cũng mang ý nghĩa bày tỏ quan điểm đối với sản phẩm.

Để hệ tư vấn có những đề xuất chính xác hơn cũng như tận dụng dữ liệu văn bản cùng các kỹ thuật phân loại được phát triển, đồ án lựa chọn đề tài **"Mô hình kết hợp hành vi đánh giá và bình luận cho tư vấn khách sạn"** với mục tiêu nghiên cứu lý thuyết về hệ tư vấn, các kỹ thuật tư vấn, tiền xử lý văn bản và phân loại văn bản về lĩnh vực cụ thể là gợi ý các khách sạn trên các bộ dữ liệu thu thập được.

Đồ án được chia thành 4 chương với nội dung như sau:

Chương 1: Tổng quan về hệ tư vấn - Nội dung trong Chương 1 giới thiệu tổng quan về hệ tư vấn và các kỹ thuật lọc cộng tác. Ngoài ra, Chương 1 còn trình bày ngắn gọn các vấn đề còn tồn tại của hệ tư vấn lọc cộng tác.

Chương 2: Tư vấn dựa trên mô hình kết hợp - Trong chương này, đồ án trình bày về mô hình kết hợp giữa hành vi đánh giá và hành vi bình luận và cách ứng dụng mô hình kết hợp vào hệ tư vấn lọc cộng tác.

Ngoài ra, nội dung Chương 2 còn trình bày về các kỹ thuật tiền xử lý dữ liệu văn bản cùng với 3 kỹ thuật phân loại văn bản: Naïve Bayes, Logistic Regression, SVM.

Chương 3: Thử nghiệm và đánh giá - Chương 3 tập trung trình bày về bộ dữ liệu được thử nghiệm, phương pháp thực nghiệm, bộ dữ liệu được sử dụng và kết quả thực nghiệm và đánh giá.

Chương 1

Đặt vấn đề

Sự phát triển mạnh mẽ của lĩnh vực công nghệ thông tin đã góp phần giúp cuộc sống của con người ngày trở nên dễ dàng và tiện lợi. Tận dụng các thành tựu của khoa học công nghệ, nhiều trang thương mại điện tử ra đời và ngày càng lớn mạnh với sự tham gia của đông đảo người dùng tiêu biểu như: Facebook, Youtube, Netflix, Amazon, Twitter, v.v.. Thông qua các trang thương mại điện tử này, quá trình tiếp thị của những nhà cung cấp dịch vụ và sản phẩm trở nên đơn giản và dễ dàng thông qua các hình thức quảng cáo. Tuy nhiên, số lượng sản phẩm và dịch vụ ngày càng nhiều, người dùng cần phải tốn nhiều thời gian hơn trong quá lựa chọn. Đây là tình trạng quá tải thông tin, gây ra sự bất tiện và khó khăn trong quá trình trích lọc thông tin của người dùng. Ngoài ra, người dùng cũng phải đối mặt với nghịch lý rằng có rất nhiều sản phẩm để lựa chọn nhưng lại không chọn ra được một sản phẩm thích hợp.

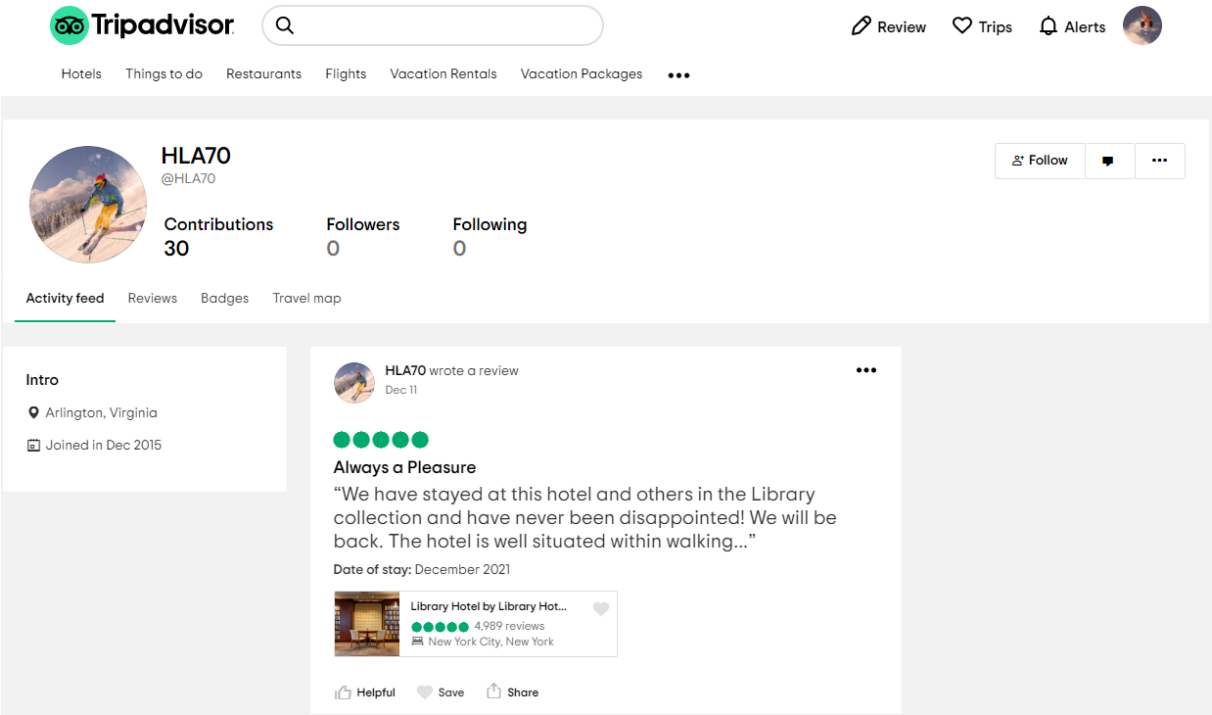
Với hiện trạng nêu trên, hệ tư vấn ngày càng đóng vai trò quan trọng trong sự phát triển của thường mại điện tử. Theo Wikipedia, hệ tư vấn là các kỹ thuật được sử dụng nhằm mục đích dự đoán điểm đánh giá mà người dùng có thể dành cho một sản phẩm. Các điểm đánh giá dự đoán này là cơ sở để thực hiện tư vấn sản phẩm phù hợp cho người dùng. Hiện nay, các hệ thống lớn cung cấp sản phẩm, dịch vụ đều phát triển hệ tư vấn của riêng mình, tiêu biểu như: hệ tư vấn phim của Netflix, hệ tư vấn âm nhạc của Pandora, hệ tư vấn sách của Amazon [4]. Theo [4], khi sử dụng

hệ tư vấn, nhà cung cấp sản phẩm và dịch vụ có thể nhận lại rất nhiều lợi ích trong đó có: tăng doanh thu bán hàng và sự hài lòng của khách hàng. Tuy nhiên, để có thể tư vấn chính xác, hệ tư vấn cần được cung cấp các dữ liệu liên quan tới sở thích và nhu cầu của người dùng. Sở thích và nhu cầu của người dùng thể hiện qua: lịch sử tìm kiếm, lịch sử mua hàng, đánh giá sản phẩm, v.v.. Những dữ liệu này đóng vai trò quyết định tới kết quả tư vấn của hệ thống.

Trong mạng xã hội, mỗi người dùng có một không gian riêng và có thể kết nối với nhau thông qua danh sách bạn bè. Trong không gian này, người dùng có quyền làm những gì họ muốn trong phạm vi hỗ trợ của nền tảng mạng xã hội, chẳng hạn như: chia sẻ một bộ phim, bình luận về một bài viết, kết bạn và theo dõi. Những hành động trên được gọi chung là hành vi của người dùng trên mạng xã hội. Các hành vi của người dùng trên mạng xã hội phản ánh một phần sở thích, tính cách và quan điểm của họ đối với những sự kiện xảy ra trên mạng xã hội. Điều này có ảnh hưởng không nhỏ tới những người trong danh sách bạn bè của họ. Hình 1.1 mô tả trang cá nhân của một người dùng trên mạng xã hội Tripadvisor. Tripadvisor là một trang chuyên cung cấp thông tin về những địa điểm du lịch: nhà hàng, khách sạn, danh lam thắng cảnh. Những người dùng trên nền tảng này để lại đánh giá cho những địa điểm mà họ đã trải nghiệm. Những đánh giá này ảnh hưởng tới quyết định trải nghiệm du lịch của những người dùng khác. Càng có nhiều người theo dõi thì mức độ ảnh hưởng của người dùng càng lớn, thể hiện thông qua: số lượng người theo dõi (Followers), tương tác của bài đánh giá (Helpful, Save, Share). Đăng bài đánh giá, theo dõi, tương tác là những hành vi chính trên mạng xã hội này. Với một bài đánh giá, phần đánh giá điểm và bình luận là 2 phần thể hiện rõ nhất quan điểm của người dùng. Vì vậy, trong phần tiếp theo, đồ án sẽ tập trung trình bày về hành vi đánh giá và hành vi bình luận của người dùng trên mạng xã hội.

Cần viết lại cho mượt

Tuy nhiên, hệ khuyến nghị cũng có những điểm yếu cần khắc phục.



Hình 1.1: Trang cá nhân của một người dùng trên mạng xã hội Tripadvisor

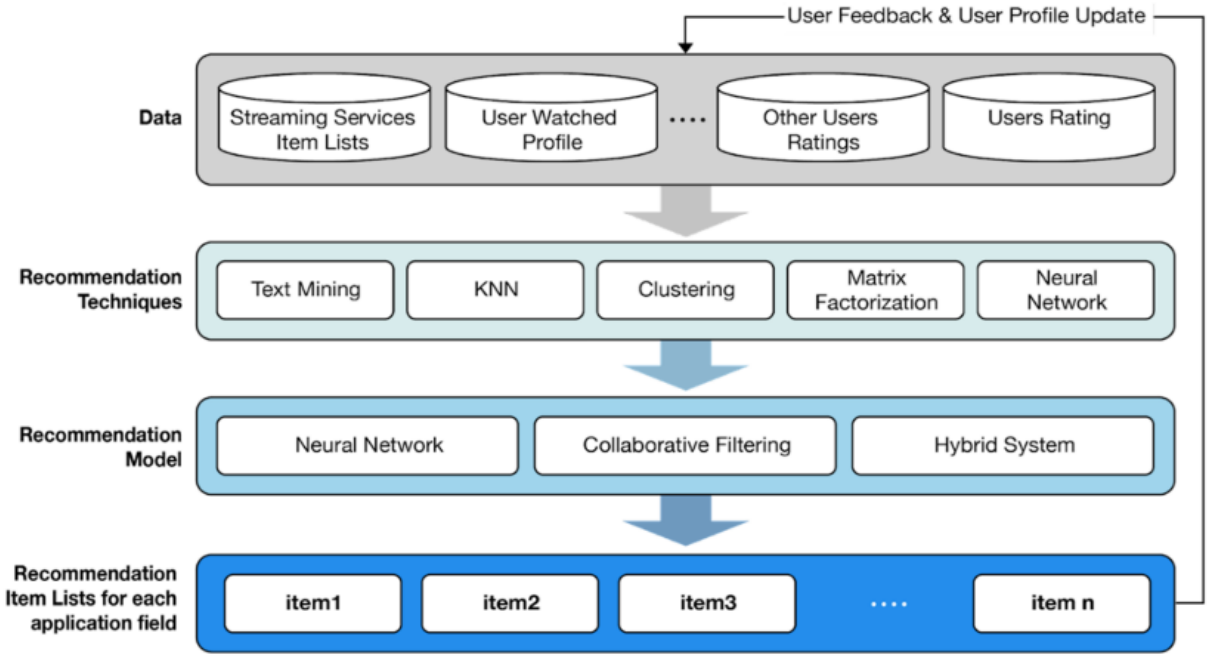
Theo [3], hệ tư vấn lọc cộng tác là kỹ thuật được sử dụng phổ biến hiện nay nhưng vẫn còn phải đối mặt với những vấn đề điển hình như: khởi đầu lạnh, thưa thớt dữ liệu và khả năng mở rộng. Đầu tiên, vấn đề thưa thớt dữ liệu, một trong những vấn đề chính của hệ tư vấn và ảnh hưởng rất nhiều đến chất lượng của hệ thống. Thông thường, dữ liệu để thực hiện tư vấn của hệ thống được biểu diễn dưới dạng ma trận người dùng-sản phẩm, giá trị của các ô trong ma trận là điểm đánh giá người dùng dành cho sản phẩm đó. Tuy nhiên, do thói quen lười đánh giá của người dùng khiến mật độ các giá trị được điền của ma trận trở nên thưa thớt. Sự thưa thớt này càng ngày càng tăng lên khi hệ thống phát triển, số lượng người dùng và sản phẩm tăng lên. Đây vẫn là một vấn đề cần phải được nghiên cứu thêm. Tiếp theo, vấn đề khởi đầu lạnh xảy ra khi gặp 1 trong 3 tình huống: người dùng mới, sản phẩm mới và hệ thống mới. Trong những tình huống này, người dùng, sản phẩm hay hệ thống chưa có dữ liệu để thực hiện khai thác, dự đoán thói quen, nhu cầu của người dùng. Vì vậy hệ thống rất khó

để thực hiện tư vấn. Cuối cùng, khả năng mở rộng là thuộc tính của hệ thống cho thấy khả năng xử lý lượng thông tin ngày càng tăng một cách hiệu quả. Với sự bùng nổ dữ liệu, đây là một thách thức lớn đối với các hệ thống khi nhu cầu xử lý thông tin liên tục tăng. Trong lọc cộng tác, các phép tính phát triển theo cấp số nhân và tốn kém tài nguyên, đôi khi dẫn đến kết quả không chính xác.

Chương 2

Cơ sở lý thuyết

Hệ thống khuyến nghị là 1 công nghệ hỗ trợ đắc lực cho con người, giúp phân tích lượng dữ liệu khổng lồ được cung cấp bởi người dùng. Hệ thống dự đoán điểm của các sản phẩm, tạo 1 danh sách sắp xếp thứ các sản phẩm này cho mỗi người dùng, và giới thiệu tới người dùng những sản phẩm mà họ có thể thích. Nội dung trong phần này trình bày tổng quan về hệ khuyến nghị, các mô hình thuật toán cùng với các kỹ thuật thuật khai phá dữ liệu trong hệ khuyến nghị hiện có. Hình 2.1 là tổng quan luồng



Hình 2.1: Tổng quan hệ thống khuyến nghị

hoạt động của 1 hệ khuyến nghị, gồm có các bước xử lý: (1) thu thập dữ liệu, (2) khai phá dữ liệu, (3) mô hình hóa dữ liệu, (4) và đưa ra gợi ý. Dữ liệu sử dụng trong hệ khuyến nghị có thể là các đánh giá, bình luận về sản phẩm, danh sách sản phẩm mà người dùng theo dõi, v.v. Các kỹ thuật khai phá dữ liệu truyền thống, có thể kể đến như: phân cụm, khai phá dữ liệu văn bản, KNN hay là học sâu, sử dụng mạng nơ-ron. Tiếp đó, các mô hình khuyến nghị sử dụng các đặc trưng đã được trích chọn để có thể mô hình hóa dữ liệu, từ đó đưa ra các khuyến nghị phù hợp tới người dùng.

Nội dung trong chương này tập trung giới thiệu và phân loại một cách tổng quát về các mô hình khuyến nghị hiện nay, có thể áp dụng vào bất kỳ 1 hệ thống khuyến nghị nào.

2.1 Mô hình khuyến nghị

Các mô hình khuyến nghị có thể được chia thành 3 nhóm chính [4]:

- **Lọc dựa trên nội dung**: Trong cách tiếp cận này, hệ thống sẽ thu thập các dữ liệu rõ ràng (điểm đánh giá sản phẩm) hoặc dữ liệu ngầm (bấm vào một đường dẫn) và tạo ra hồ sơ người dùng. Hệ thống sẽ thực hiện tư vấn những sản phẩm dựa trên những sản phẩm và hành vi liên quan tới hồ sơ người dùng. Do sở thích của người dùng thường được chia thành vài nhóm cơ bản, việc chỉ sử dụng hồ sơ của 1 người dùng khiến hệ thống không tận dụng được thông tin từ những người dùng khác, từ đó hạn chế sự linh hoạt của hệ tư vấn.

- **Lọc cộng tác**: Không giống với lọc dựa trên nội dung, lọc cộng tác tìm kiếm những người dùng có sở thích tương tự nhau. Từ giả định những người dùng A có sở thích giống với người dùng B, hệ thống sẽ tiến hành tư vấn cho người dùng B những sản phẩm phù hợp người dùng A. Lọc cộng tác có 2 hướng tiếp cận: dựa trên bộ nhớ và dựa trên mô hình. Hướng tiếp cận dựa trên bộ nhớ tính toán độ tương tự giữa các người dùng từ đó thực hiện tư vấn. Nhược điểm của hướng

tiếp cận này là sự tốn kém tài nguyên khi số lượng người dùng và sản phẩm tăng lên. Hướng tiếp cận dựa trên mô hình sử dụng các mô hình đã được huấn luyện thông qua các thuật toán học máy hoặc khai phá dữ liệu để thực hiện tư vấn.

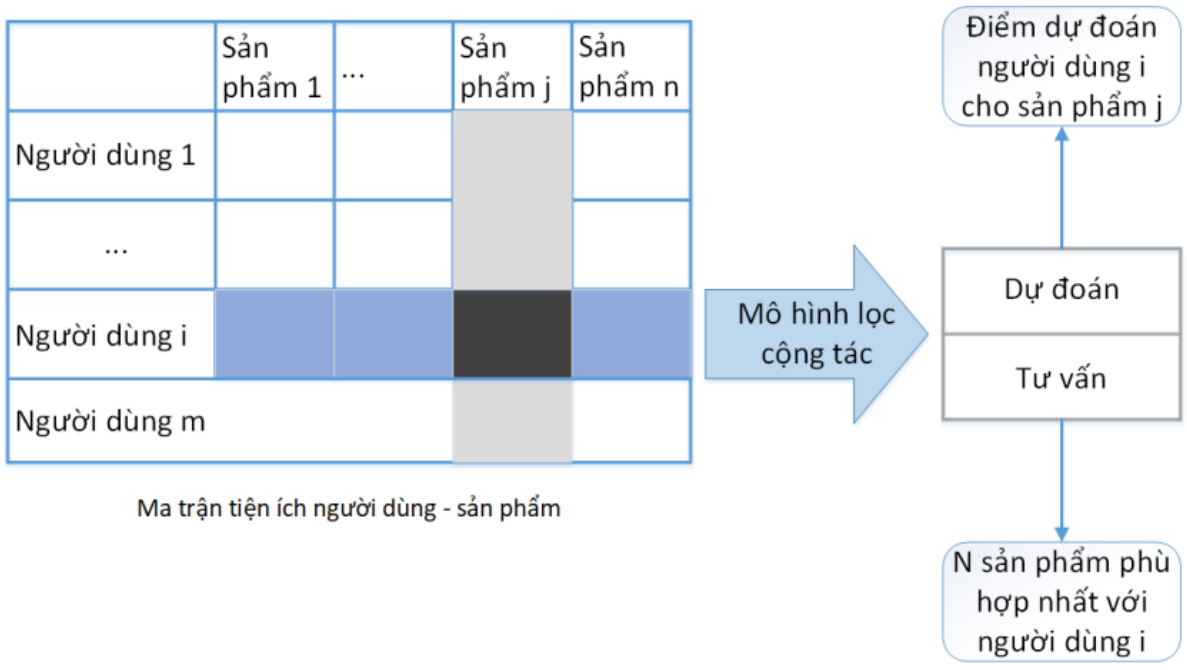
- **Hệ tư vấn lai** Lọc dựa trên nội dung và lọc cộng tác đều có ưu điểm và nhược điểm riêng. Để giải quyết vấn đề này, hệ tư vấn lai được sinh ra, là sự kết hợp của 2 kỹ thuật trên.

Lọc cộng tác là một mô hình lọc thông tin, xây dựng 1 cơ sở dữ liệu sở thích người dùng thông qua dữ liệu tưởng tác giữa họ với sản phẩm để dự đoán các sản phẩm phù hợp với sở thích của họ, từ đó đưa ra các khuyến nghị về sản phẩm. Ý tưởng của mô hình lọc cộng tác là từ dữ liệu hành vi tương tác giữa người dùng và sản phẩm, hệ thống sẽ tính toán mức độ tương đồng giữa các người dùng hoặc giữa các sản phẩm, tạo cơ sở thực hiện khuyến nghị. Những người dùng có mức độ tương đồng cao sẽ có xu hướng mua những sản phẩm giống nhau. Với mỗi cách tính độ tương đồng sẽ cho một mô hình lọc cộng tác khác nhau.

Các mô hình lọc cộng tác có thể được chia ra thông qua 2 hướng tiếp cận: lọc cộng tác dựa trên bộ nhớ và lọc cộng tác dựa trên mô hình. Hướng tiếp cận dựa trên bộ nhớ tính toán độ tương tự giữa các người dùng từ đó thực hiện tư vấn. Nhược điểm của hướng tiếp cận này là sự tốn kém tài nguyên khi số lượng người dùng và sản phẩm tăng lên. Ngoài ra, hệ thống cần tính toán tại thời điểm khuyến nghị, điều này sẽ ảnh hưởng tới thời gian đưa ra dự đoán. Hướng tiếp cận dựa trên mô hình sử dụng các mô hình đã được huấn luyện thông qua các thuật toán học máy hoặc khai phá dữ liệu để thực hiện tư vấn. Với hướng tiếp cận này, mô hình sẽ cần phải thực hiện huấn luyện trước, nhưng khi thực hiện khuyến nghị sẽ rất nhanh. Trong lọc cộng tác dựa trên bộ nhớ, ta có thể phân loại thành: lọc cộng tác dựa trên người dùng và lọc cộng tác dựa trên sản phẩm. Lọc cộng tác dựa trên người dùng là 1 mô hình so sánh sự tương đồng giữa các người dùng thông qua dữ liệu tương tác của họ lên các sản phẩm, từ

đó khuyến nghị các sản phẩm phù hợp. Lọc cộng tác dựa trên sản phẩm dự đoán bằng cách sử dụng độ tương đồng giữa sản phẩm và sản phẩm được chọn bởi người dùng thông qua 1 ma trận tương tác của người dùng và sản phẩm. Nói cách khác, lọc cộng tác dựa trên bộ nhớ sử dụng các kỹ thuật như: độ tương quan Pearson, độ tương quan cô-sin, KNN để tạo các nhóm có đặc tính giống nhau, từ đó khuyến nghị các sản phẩm tới người dùng trong nhóm. Do cách hoạt động dựa trên dữ liệu đánh giá, nên mô hình khó có thể hoạt động tốt khi không có đủ dữ liệu cần thiết. Để khắc phục vấn đề này, lọc cộng tác dựa trên mô hình đưa ra khuyến nghị nhờ sử dụng các thuật toán như: phân cụm, SVD hay PCA.

2.1.1 Lọc cộng tác dựa trên bộ nhớ



Hình 2.2: Mô hình thuật toán lọc cộng tác

Thông thường, hồ sơ người dùng – sản phẩm thường được xây dựng từ điểm đánh giá người dùng chấm cho sản phẩm, được gọi là ma trận tương tác. Ma trận tương tác sẽ có dạng như trong Hình 2.2, với các hàng/cột là danh sách người dùng, cột/hàng là danh sách sản phẩm, các giá trị trong

mỗi ô tương ứng với điểm đánh giá người dùng danh cho sản phẩm. Trong thực tế, người dùng thường ít đánh giá sản phẩm nên ma trận tiện ích trở nên thưa thớt, nghĩa là có nhiều giá trị chưa được điền. Hình 2.2 là mô hình xử lý, mô tả cho thuật toán lọc cộng, tác được chia thành 3 bước thực hiện:

1. Chuẩn hóa dữ liệu
2. Tính toán độ tương đồng
3. Dự đoán mức độ quan tâm của người dùng lên sản phẩm

Lọc cộng tác dựa trên người dùng

Bước 1: Chuẩn hóa dữ liệu

Trong thực tế, người dùng "lười" đánh giá sản phẩm khiến ma trận tiện ích trở nên thưa thớt. Do đó cần chuẩn hóa dữ liệu để loại bỏ những giá trị chưa biết trong ma trận. Xét ví dụ trong Bảng 1.1 là ma trận tiện ích được xây dựng từ tập người dùng $W = w_1, ..., w_5$ và tập sản phẩm $X = x_1, ..., x_5$. Mỗi sản phẩm được người dùng đánh giá trên thang điểm từ 0 đến 5. Các giá trị "?" nghĩa là người dùng chưa đánh giá những sản phẩm tương ứng.

	x_1	x_2	x_3	x_4	x_5
w_1	5	5	2	0	?
w_2	2	4	0	?	?
w_3	0	1	3	4	5
w_4	5	?	?	?	1
w_5	?	?	3	2	4

Bảng 2.1: Ví dụ ma trận tương tác

Cách dễ nhất để điền các giá trị còn thiếu vào trong ma trận này là chọn điểm cao nhất hoặc điểm thấp nhất (5 hoặc 0). Tuy nhiên, khi chọn giá trị này sẽ gây mất cân bằng và giảm độ chính xác của hệ thống. Một

giá trị an toàn có thể điển là điểm trung bình của thang đo (2,5). Tuy nhiên, giá trị này sẽ không đúng với những người dùng khó tính hoặc dễ tính. Vì người dùng khó tính sẽ chỉ cho 4 với những sản phẩm họ thích, ngược lại người dùng dễ tính sẽ cho 1, 2 với những sản phẩm họ không thích. Do đó cần có một cách chuẩn hóa khác để khắc phục vấn đề này. Các bước chuẩn hóa sẽ được trình bày ngay sau đây.

1. Tính trung bình các điểm đánh giá mà mỗi người dùng đã đưa ra. Ví dụ, người dùng w_1 đã chấm 4 sản phẩm với số điểm lần lượt là: 5, 5, 2, 0. Như vậy, điểm trung bình người dùng w_1 đưa ra là: $\frac{5+5+2+0}{4} = 3$.

	x_1	x_2	x_3	x_4	x_5	Điểm TB
w_1	5	5	2	0	?	3
w_2	2	4	0	?	?	2
w_3	0	1	3	4	5	2.6
w_4	5	?	?	?	1	3
w_5	?	?	3	2	4	3

Bảng 2.2: Ví dụ ma trận tương tác

2. Thực hiện trừ điểm đánh giá của người dùng với điểm đánh giá trung bình của họ

	x_1	x_2	x_3	x_4	x_5	Điểm TB
w_1	2	2	-1	-3	?	3
w_2	0	2	-2	?	?	2
w_3	-2.6	-1.6	0.4	1.4	2.4	2.6
w_4	2	?	?	?	-2	3
w_5	?	?	0	-1	1	3

Bảng 2.3: Ví dụ ma trận tương tác

	x_1	x_2	x_3	x_4	x_5	Điểm TB
w_1	2	2	-1	-3	0	3
w_2	0	2	-2	?	0	2
w_3	-2.6	-1.6	0.4	1.4	2.4	2.6
w_4	2	0	0	0	-2	3
w_5	0	0	0	-1	1	3

Bảng 2.4: Ví dụ ma trận tương tác

3. Các ô chưa biết thì điền 0.

Cách chuẩn hóa trên có 2 ưu điểm: (1) Việc trừ đi điểm đánh giá trung bình của người dùng khiến ma trận có giá trị âm, dương. Những giá trị dương ứng với những sản phẩm được người dùng quan tâm hơn. Những ô có giá trị 0 biểu diễn cho người dùng chưa đánh giá sản phẩm này. Đây là những giá trị cần dự đoán. (2) Số chiều của ma trận tiện ích là rất lớn khi người dùng và sản phẩm tăng lên. Vì vậy, để tiết kiệm bộ nhớ, ma trận tiện ích sẽ được lưu dưới dạng ma trận thưa do những dấu "?" đã được thay bằng giá trị 0.

Bước 2: Tính toán độ tương đồng và dự đoán

Với mỗi cách tính độ tương đồng sẽ cho ra một thuật toán lọc cộng tác khác nhau. Nếu tính độ tương đồng giữa các cặp người dùng ta có thuật toán lọc cộng tác người dùng. Nếu tính độ tương đồng giữa các cặp sản phẩm, ta có thuật toán lọc cộng tác sản phẩm. Để tính độ tương đồng giữa người dùng w_i và w_j , ta sử dụng công thức cô-sin:

$$cosin\_similarity(w_i, w_j) = cos(w_i, w_j) = \frac{w_i^T w_j}{||w_i||_2 ||w_j||_2} \qquad (2.1)$$

Trong đó, w_i và w_j là các véc-tơ tương ứng với hàng/cột w_i và w_j trong ma trận tương tác. Sau khi tính toán được độ tương đồng giữa các cặp người dùng, thuật toán sẽ dự đoán mức độ quan tâm của người dùng u lên sản phẩm i dựa trên thông tin từ K người dùng giống u nhất, được định nghĩa

theo công thức:

$$\hat{y}_{i,u} = \frac{\sum_{u,j \in \mathrm{N}(u,i)} \bar{y}_{i,u_j} sim(u,u_j)}{\sum_{u,j \in \mathrm{N}(u,i)} |sim(u,u_j)|} \tag{2.2}$$

Với $N(u,i)$ là tập hợp K người dùng gần giống u nhất và đã đánh giá sản phẩm i.

Lọc cộng tác người dùng thường hoạt động không hiệu quả trên các hệ thống lớn do số lượng người dùng khổng lồ. Khi đó, việc tính toán độ tương đồng giữa các cặp người dùng trở nên tốn kém tài nguyên là thời gian.

Lọc cộng tác dựa trên sản phẩm

Lọc cộng tác sản phẩm là hướng tiếp cận có thể khắc phục nhược điểm của lọc cộng tác người dùng do số lượng sản phẩm trên hệ thống thường không biến động mạnh. Thay vì tính toán độ tương động giữa các cặp người dùng, lọc cộng tác sản phẩm tính toán độ tương đồng giữa các sản phẩm.

Chuẩn hóa dữ liệu

- Tính trung bình điểm đánh giá sản phẩm nhận được

	x_1	x_2	x_3	x_4	x_5
w_1	5	5	2	0	?
w_2	2	4	0	?	?
w_3	0	1	3	4	5
w_4	5	?	?	?	1
w_5	?	?	3	2	4
Điểm TB	3	3.333	2	2	3.333

Bảng 2.5: Ví dụ ma trận tương tác

- Thực hiện trừ điểm đánh giá của sản phẩm với điểm đánh giá trung bình

	x_1	x_2	x_3	x_4	x_5
w_1	2	1.667	0	-2	?
w_2	-1	0.667	-2	?	?
w_3	-3	-2.333	1	2	1.667
w_4	2	?	?	?	-2.333
w_5	?	?	1	0	0.667
Điểm TB	3	3.333	2	2	3.333

Bảng 2.6: Ví dụ ma trận tương tác

- Các ô "?" điền giá trị 0

	x_1	x_2	x_3	x_4	x_5
w_1	2	1.667	0	-2	0
w_2	-1	0.667	-2	0	0
w_3	-3	-2.333	1	2	1.667
w_4	2	0	0	0	-2.333
w_5	0	0	1	0	0.667
Điểm TB	3	3.333	2	2	3.333

Bảng 2.7: Ví dụ ma trận tương tác

Tính toán độ tương đồng và dự đoán

Dự đoán độ quan tâm của w_2 lên w_5 sử dụng lọc cộng tác sản phẩm.

- Sản phẩm được w_2 đánh giá: $\{x_1, x_2, x_3\}$
- Độ tương tự giữa x_5 và $\{x_1, x_2, x_3\}$ lần lượt là: {-0.774, -0.449, 0.324}
- Xét K=2, ta có 2 sản phẩm giống x_5 nhất : $N(u,i) = \{x_2, x_3\}$ với điểm đánh giá chuẩn hóa là {0.667, -2}
- $\hat{y}_{(w_2,x_5)} = \frac{0.667 * -0.449 + (-2) * 0.324}{0.324 + |-0.449|} = -1.226$

- Đưa điểm đánh giá về thang đo ban đầu, ta cộng điểm đánh giá dự đoán với điểm đánh giá trung bình của sản phẩm: $-1.226 + 3.333 = 2.107$

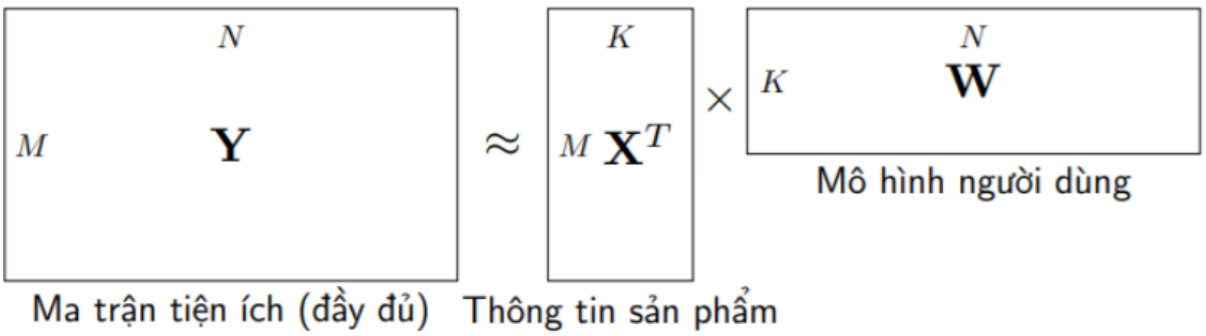
2.1.2 Lọc cộng tác phân tích ma trận

Giới thiệu

Ý tưởng chính của phương pháp này là tồn tại các đặc trưng ẩn mô tả sự liên quan giữa các sản phẩm và người dùng. Ví dụ với các bộ phim, các đặc trưng ẩn có thể rõ ràng như: hài, chính kịch, hành động, hoặc chúng là sự kết hợp của các đặc trưng ẩn rõ ràng, hoặc chúng là những đặc trưng chưa được đặt tên. Tương tự, mỗi người dùng cũng sẽ có xu hướng thích những đặc trưng ẩn nào đó của phim. Thay vì xây dựng ma trận của M sản phẩm X một cách độc lập, các đặc trưng ẩn này được huấn luyện đồng thời với dữ liệu của ma trận N người dùng Y.

Với ý tưởng trên, thay vì xây dựng ma trận Y nghĩa là dự đoán các giá trị còn khuyết trong Y thì thuật toán sẽ cố gắng sấp xỉ ma trận người dùng W và ma trận sản phẩm X, sao cho tích của 2 ma trận này là $\hat{Y}$ xấp xỉ với Y.

$$Y \approx \hat{Y} = X^T W$$



Hình 2.3: Mô tả phân tích ma trận với K đặc trưng ẩn

Hình 2.3 mô tả phương pháp phân tích ma trận tiện ích Y thành 2 ma trận người dùng W và sản phẩm X. Trong đó K là số đặc trưng ẩn được

giả định của mỗi sản phẩm. Thông thường, K được chọn là một số nhỏ hơn M và N rất nhiều. Khi đó, hạng của X và Y không cao, giúp tiết kiệm tài nguyên.

Hàm mất mát

Hàm mất mát được xây dựng dựa trên các ô đã được điền của ma trận Y, được định nghĩa như sau:

$$L(X,W) = \frac{1}{2s}\sum_{n=1}^{N}\sum_{m:r_{mn}=1}(y_{mn} - x_m w_n)^2 + \frac{\lambda}{2}(||X||_F^2 + ||W||_F^2) \tag{2.3}$$

Trong đó $r_{mn} = 1$ nếu sản phẩm thứ m đã được đánh giá với người dùng thứ n, $||x||_2^F$ là căn bậc 2 của tổng bình phương tất cả các phần tử của ma trận, s là toàn bộ số đánh giá đã có. Trong công thức trên, thành phần thứ nhất chính là trung bình sai số của mô hình, thành phần thứ hai là $l2$ regularization, giúp tránh overfitting.

Việc tối ưu cả 2 ma trận X và W cùng lúc là tương đối phức tạp, vì vậy, phương pháp được sử dụng là tối ưu từng ma trận trong khi ma trận kia cố định đến khi hội tụ.

Tối ưu hàm mất mát

Gradient Descent là kỹ thuật được dùng để tối ưu 2 bài toán: cố định X tối ưu W và cố định W tối ưu X.

Với trường hợp cố định X tối ưu W, hàm mất mát được biểu diễn như sau:

$$L(W) = \frac{1}{2s}\sum_{n=1}^{N}\sum_{m:r_{mn}=1}(y_{mn} - x_m w_n)^2 + \frac{\lambda}{2}||W||_F^2 \tag{2.4}$$

Việc tối ưu công thức trên có thể được tách thành N bài toán nhỏ, mỗi bài toán ứng với việc đi tối ưu từng cột của ma trận W:

$$L(w_n) = \frac{1}{2s}\sum_{m:r_{mn}=1}(y_{mn} - x_m w_n)^2 + \frac{\lambda}{2}||w_n||_2^2 \tag{2.5}$$

Vì biểu thức chỉ phụ thuộc vào các sản phẩm đã được đánh giá bởi người dùng đang xét, công thức có thể được đơn giản bằng cách đặt $\hat{X}_n$ là ma trận được tạo bởi các hàng tương ứng với các sản phẩm đã được đánh giá đó, và $\hat{y}_n$ là các đánh giá tương ứng. Khi đó công thức trở thành:

$$L(w_n) = \frac{1}{2s}||\hat{y}_n - \hat{X}_n w_n||^2 + \frac{\lambda}{2}||w_n||_2^2 \tag{2.6}$$

và đạo hàm của nó:

$$\frac{\partial L(w_n)}{\partial w_n} = -\frac{1}{s}\hat{X}_n^T(\hat{y}_n - \hat{X}_n w_n) + \lambda w_n \tag{2.7}$$

Từ đó, công thức cập nhật cho mỗi cột của W được định nghĩa như sau:

$$w_n = w_n - \eta\left(-\frac{1}{s}\hat{X}_n^T(\hat{y}_n - \hat{X}_n w_n) + \lambda w_n\right) \tag{2.8}$$

Khi thực hiện cố định W tối ưu X, hàm mất mát được biểu diễn:

$$L(X) = \frac{1}{2s}\sum_{n=1}^{N}\sum_{m:r_{mn}=1}(y_{mn} - x_m w_n)^2 + \frac{\lambda}{2}||X||_F^2 \tag{2.9}$$

Việc tối ưu công thức trên có thể được tách thành M bài toán nhỏ, mỗi bài toán ứng với việc đi tối ưu một cột của ma trận X:

$$L(x_m) = \frac{1}{2s}\sum_{n:r_{mn}=1}(y_{mn} - x_m w_n)^2 + \frac{\lambda}{2}||x_m||_2^2 \tag{2.10}$$

Vì biểu thức chỉ phụ thuộc vào các sản phẩm đã được đánh giá bởi người dùng đang hàng xét, công thức có thể được đơn giản bằng cách đặt $\hat{W}_m$ là ma trận được tạo bởi các hàng tương ứng với các sản phẩm đã được đánh giá đó, và $\hat{y}_m$ là các đánh giá tương ứng. Khi đó, công thức trở thành:

$$L(x_m) = \frac{1}{2s}||\hat{y}_m - x_m\hat{W}_m||_2^2 + \frac{\lambda}{2}||x_m||_2^2 \tag{2.11}$$

và đạo hàm của nó:

$$\frac{\partial L(x_m)}{\partial x_m} = -\frac{1}{s}\left(\hat{y}_m - x_m\hat{W}_m\right)\hat{W}_m^T + \lambda x_m \tag{2.12}$$

Từ đó, công thức cập nhật cho mỗi cột của W được định nghĩa như sau:

$$x_m = x_m - \eta\left(-\frac{1}{s}(\hat{y}_m - x_m\hat{W}_m)\hat{W}_m^T + \lambda x_m\right) \tag{2.13}$$

2.2 Phân loại quan điểm người dùng

2.2.1 Tiền xử lý dữ liệu

Quá trình tiền xử lý dữ liệu gồm 4 bước:(1)Chuẩn hóa văn bản: Bước này, văn bản được đưa về chữ thường, các biểu tượng cảm xúc, đường dẫn bị loại bỏ. (2) Tách từ và loại bỏ dấu câu: Tách từ là đưa câu bình luận về dạng 1 danh sách các từ cũng với đó là loại bỏ các dấu câu. Các dấu câu không có ý nghĩa cho việc phân loại quan điểm. (3) Loại bỏ Stopword: Stopword là những từ xuất hiện nhiều nhưng không có ý nghĩa trong quá trình phân loại quan điểm. Ví dụ: “is”, “a”, “the”, ... (4) Chuyển về dạng chuẩn: Ví dụ: “rooms”=>”room”, “person”=>”people”, các từ được đưa về dạng nguyên bản.

2.2.2 Trích chọn đặc trưng

TF-IDF: là một phương pháp thống kê, nhằm phản ánh độ quan trọng của mỗi từ hoặc 1 cụm N-grams đối với văn bản trong phạm vi toàn bộ tài liệu đầu vào. Cho một kho gồm p văn bản khác nhau $D = D_1, D_2, ..., D_N$, mỗi văn bản D_i được tạo bởi các từ $D_i = d_{i1}, ..., d_{in}$. Cho $T = t_1, t_2, ..., t_q$ là tập hợp những từ xuất hiện trong kho văn bản.

$$\text{tf-idf}(t_j, D_i) = tf(t_j, D_i) \times idf(t_j, D) \tag{2.14}$$

Trong đó, $tf(t, d)$ của từ t trong văn bản d được định nghĩa theo:

$$tf(t, d) = \frac{\text{số lần t suất hiện trong d}}{\text{số từ trong d}} \tag{2.15}$$

N-grams: Một cụm N-grams là một dãy gồm N ký tự hoặc từ liên tiếp nhau trong một văn bản cho trước. Số phần tử trong một cụm N-grams được gọi là bậc của N-grams. Thông thường, bậc của N-grams thường nằm trong khoảng (1,3), với các tên gọi tương ứng là unigram (bậc 1), bigram (bậc 2) và trigram (bậc 3). N-grams được dùng để tính tần suất xuất hiện của 1 cụm N-grams có trong kho văn bản. Bảng **??** là ví dụ biểu diễn N-grams với bậc 1, 2, 3 cho câu: “That picture is beautiful.”.

Bậc	Cụm từ
1 gram	That, picture, is, beautiful
2 gram	That picture, picture is, is beautiful
3 gram	That picture is, picture is beautiful

Bảng 2.8: Biểu diễn N-grams cho 1 câu

2.2.3 Mô hình học máy có giám sát cho bài toán phân loại quan điểm người dùng

Naive Bayes Classifier: Bộ phân loại Bayes là một giải thuật thuộc lớp giải thuật phân lớp thống kê, nó có thể dự đoán xác suất của một phần tử dữ liệu thuộc vào một lớp là bao nhiêu. Bộ phân loại Bayes được dựa trên định lý Bayes [7]. Định lý Bayes cho phép tính xác suất xảy ra của 1 sự kiện ngẫu nhiên A khi biết sự kiện liên quan B đã xảy ra. Xác suất này được ký hiệu là $P(A|B)$, và đọc là "xác suất của A nếu có B". Đại lượng này được gọi là xác suất có điều kiện hay xác suất hậu nghiệm vì nó được rút ra từ giá trị được cho của B hoặc phụ thuộc vào giá trị nào đó. Theo định lý Bayes, xác suất xảy ra A khi biết B phụ thuộc vào 3 yếu tố:

- Xác suất xảy ra A của riêng nó, không quan tâm tới B, ký hiệu là $P(A)$, đọc là xác suất của A. Đây là xác xuất biên duyên hay xác suất tiên nghiệm, nó là “tiên nghiệm” nghĩa rằng nó không quan tâm tới bất cứ thông tin nào của B.

- Xác suất xảy ra B của riêng nó, không quan tâm đến A, ký hiệu là $P(B)$ và đọc là "xác suất của B". Đại lượng này còn gọi là hằng số chuẩn hóa (Normalising Constant), vì nó luôn giống nhau, không phụ thuộc vào sự kiện A đang muốn biết.

- Xác suất xảy ra B khi biết A xảy ra. Ký hiệu là P(B|A) và đọc là "xác suất của B nếu có A". Đại lượng này gọi là khả năng (Likelihood) xảy ra B khi biết A đã xảy ra.

Khi đó, xác suất của A khi biết B được định nghĩa bởi công thức:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.16}$$

Từ định lý trên, bộ phân loại Naive Bayes được phát triển và hoạt động như sau [1]:

1. Gọi D là tập dữ liệu huấn luyện, trong đó mỗi phần tử dữ liệu X được biểu diễn bằng một vector chứa n giá trị thuộc tính $A_1, A_2, ..., A_n$, $X = x_1, x_2, ..., x_n$

2. Giả sử có m lớp $C_1, C_2, ..., C_n$, cho 1 phần từ dữ liệu X, bộ phân loại sẽ gán nhãn cho X là lớp có xác suất hậu nghiệp lớn nhất. Cụ thể, bộ phân loại Bayes sẽ dự đoán X thuộc vào lớp C_i khi và chỉ khi:

$$P(C_i|X) > P(C_j|X) \quad (1 \leq i \leq m, i \neq j) \tag{2.17}$$

3. Để tìm giá trị xác suất lớn nhất, ta nhận thấy trong công thức 2.17 giá trị $P(X)$ là giống nhau với mọi lớp nên không cần tính. Do đó chỉ cần tìm giá trị lớn nhất của $P(X|C_i) \times P(C_i)$. Trong đó, $P(C_i)$ được ước lượng bằng công thức $P(C_i) = \frac{|D_i|}{D}$ với D_i là tập các phần tử dữ liệu thuộc lớp C_i. Nếu xác suất tiền nghiệm $P(C_i)$ cũng không xác định được thì ta coi chúng bằng nhau, khi đó chỉ cần tìm giá trị $P(X|C_i)$ lớn nhất.

4. Khi số lượng các thuộc tính mô tả dữ liệu là lớn thì chi phí tính toán $P(X|C_i)$ là rất lớn, do đó để làm giảm độ phức tạp, giải thuật Naive Bayes giả thiết các thuộc tính là độc lập nhau hay không có sự phụ thuộc nào giữa các thuộc tính. Khi đó ta có:

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times ...P(x_n|C_i) \tag{2.18}$$

Naive Bayes là một giải thuật đơn giản, dễ cài đặt, thời gian huấn luyện nhanh, thực hiện phân loại khá tốt với các bài toán đa nhãn và không cần

quá nhiều dữ liệu huấn luyện. Tuy nhiên, giả định về sự độc lập giữa các đặc trưng của dữ liệu thường khó xảy ra trong thế giới thực. Với những đặc điểm trên, Naïve Bayes thường được sử dụng trong các hệ thống dự đoán thời gian thực, các bài toán dự đoán đa nhãn, phân loại văn bản, lọc thư rác, ...

Support Vector Machines: là kỹ thuật học có giám sát được đề xuất lần đầu tiên vào năm 1992 cho bài toán phân loại nhị phân [5]. Hiện nay thuật toán này được mở rộng cho các bài toán phân loại đa lớp. SVM hỗ trợ xây dựng một siêu phẳng hoặc một tập hợp các siêu phẳng trong một không gian nhiều chiều hoặc vô hạn chiều, có thể được sử dụng cho phân loại, hồi quy hoặc các nhiệm vụ khác. Để phân loại tốt nhất thì các siêu phẳng nằm càng xa các điểm dữ liệu của tất cả các lớp (gọi là lề) càng tốt. Trong nhiều trường hợp, không thể phân chia các lớp dữ liệu một cách tuyến tính trong một không gian ban đầu vì vậy, cần phải ánh xạ các điểm dữ liệu trong không gian ban đầu vào một không gian mới nhiều chiều hơn, để việc phân tách chúng trở nên dễ dàng hơn.

Để việc tính toán được hiệu quả, phép ánh xạ sử dụng trong thuật toán SVM chỉ ràng buộc tích vô hướng của các véc-tơ dữ liệu trong không gian mới có thể được tính dễ dàng từ các tọa độ trong không gian cũ.

$$K(a,b) =< a,b > \tag{2.19}$$

Sử dụng hàm đối ngẫu Lagrange, bài toán tìm lệ cực đại của SVM được đưa về bài toán tìm véc-tơ hệ số $\vec{\alpha} = (\alpha_1, ..., \alpha_n)$ cho phép cực tiểu hóa hàm mục tiêu

$$W(\vec{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j K(x_i, x_j) \tag{2.20}$$

đồng thời thỏa mãn:

$$\sum_{i=1}^{n} y_i \alpha_i = 0 \qquad 0 \leq \alpha_i \leq C \tag{2.21}$$

Trong đó, x_i và y_i tương ứng là dữ liệu và nhãn của nó, α_i là hệ số cần xác định, C là số điểm dữ liệu tối đa được phân loại sai. Quá trình huấn luyện

SVM là quá trình xác định α_i. Phương pháp hiệu quả và thông dụng nhất là tối ưu tuần tự SMO [8]. Sau khi phân loại xong, giá trị nhãn phân loại cho mẫu mới được tỉnh bởi:

$$f(x) = sign(\sum_{i=1}^{n} y_i \alpha_i K(x_i, x) + b) \tag{2.22}$$

với b được tính trong giai đoạn huấn luyện theo công thức:

$$b = y_i - \sum_{j=1}^{n} y_j \alpha_j K(x_i, x_j) \tag{2.23}$$

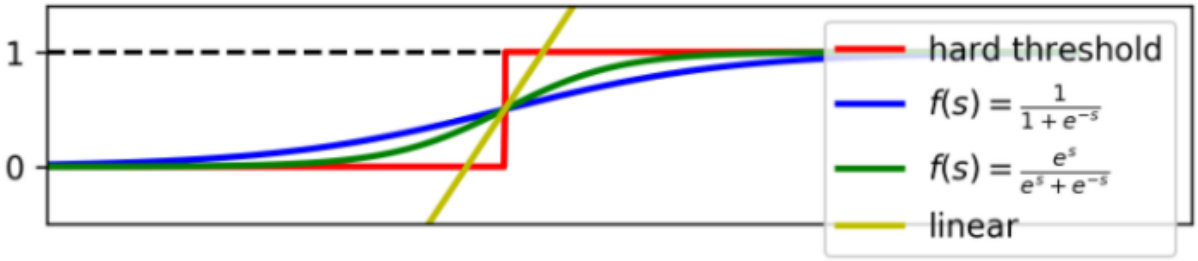
Trong đó, i là một hệ số thỏa mãn điều kiện $0 < \alpha_i < C$

Logistic Regression: có công thức biểu diễn như sau:

$$f(x) = \theta(w^T x) \tag{2.24}$$

Trong đó, $f(x)$ là xác suất sinh viên đỗ hay trượt, x là số giờ sinh viên ôn tập, w^T là hằng số được huấn luyện sao cho kết quả dự đoán là chính xác nhất, θ là hàm kính hoạt đưa kết quả về dạng xác suất. Tuy nhiên, các bài toán trong thực tế thường có dữ liệu có nhiều đặc trưng, cho nên $x = (x_1, x_2, ..., x_n)$ là một véc-tơ, w là ma trận các hằng số. Một số hàm kích hoạt cho mô hình tuyến tính được mô tả trong Hình 2.3. Đường màu đỏ và vàng không phù hợp với bài toán. Đường màu vàng không bị chặn ở 2 đầu. Ngoài ra, các điểm dữ liệu trong bài toán không hoàn toàn phân tách nên đường màu đỏ không phù hợp. Các đường màu xanh lam và xanh lục phù hợp với bài toán của đã nêu hơn. Chúng có một vài tính chất quan trọng sau:

- Là hàm số liên tục nhận giá trị thực, bị chặn trong khoảng (0, 1)
- Nếu coi điểm có tung độ là 1/2 làm điểm phân chia thì các điểm càng xa điểm này về phía bên trái có giá trị càng gần 0. Ngược lại, các điểm càng xa điểm này về phía phải có giá trị càng gần 1. Điều này khớp với nhận xét rằng học càng nhiều thì xác suất đỗ càng cao và ngược lại.



Hình 2.4: Một số hàm kích hoạt

- Là hàm liên tục nên có đạo hàm mọi nơi.

Sigmoid là hàm kích hoạt thường xuyên được sử dụng vì nó hoạt động trong khoảng [0, 1]. Hơn nữa, đạo hàm của hàm sigmoid rất đơn giản nên được sử dụng rộng rãi.

$$f(x) = \frac{1}{1+e^{-x}} \tag{2.25}$$

$$\frac{\partial f(x)}{\partial x} = f(x)(1-f(x)) \tag{2.26}$$

Công thức cập nhật cho Logistic Regression sử dụng hàm Sigmoid theo phương pháp Stochastic Gradient Descent với điểm dữ liệu (x_i, y_i) là:

$$w = w + \mu(y_i - z_i)x_i \tag{2.27}$$

Trong đó, $z_i = \theta(w^T x_i)$. Tuy có tên là Regression, nhưng thuật toán này thường sử dụng nhiều trong các bài toán phân loại. Mô hình này phân loại dữ liệu dựa trên phương trình siêu phẳng $w^T x$ có dạng tuyến tính. Do vậy, mô hình này chỉ phù hợp với các loại dữ liệu mà 2 lớp là phân biệt tuyến tính. Logistic Regression không phù hợp với các loại dữ liệu có lớp nằm bên ngoài đường tròn, lớp nằm trong đường tròn đó. Ngoài ra, các điểm dữ liệu nhiễu sẽ ảnh hưởng rất nhiều tới độ chính xác của mô hình.

2.3 Kết luận

Chương 2 đã trình bày các kiến thức, nghiên cứu liên quan về tổng quan mô hình tư vấn và các phương pháp trích chọn đặc trưng, phân loại

quan điểm người dùng. Chương 3 tiếp theo sẽ trình bày về phương pháp đề xuất để giải quyết bài toán đã đặt ra.

Chương 3

Phương pháp giải quyết vấn đề

3.1 Hành vi người dùng trên mạng xã hội

3.1.1 Hành vi đánh giá

Các hệ tư vấn thường được xây dựng từ:

- Tập người dùng $W = w_1, w_2, ..., w_n$
- Tập sản phẩm $X = x_1, x_2, ..., x_n$

Hành vi đánh giá là hành động người dùng chấm điểm cho sản phẩm. Thông tin này được lưu trữ và thường được sử dụng làm cơ sở cho hệ thống thực hiện tư vấn. Điểm đánh giá từ người dùng w_i cho sản phẩm x_j được định nghĩa như sau:

$$rating_{ij} = y, \qquad y \in 1, 2, ..., t \tag{3.1}$$

Trong đó, t thường được chọn là 5 hoặc 10.

3.1.2 Hành vi bình luận

Hành vi bình luận là hành động của người dùng khi diễn đạt suy nghĩ, quan điểm của mình bằng văn bản. Người dùng thực hiện hành vi bình luận đối với sản phẩm thay vì hành vi chấm điểm sẽ mô tả rõ hơn trải nghiệm, suy nghĩ của họ đối với sản phẩm. Mỗi bình luận $comment_{ij}$ người

dùng w_i bày tỏ quan điểm đối với sản phẩm x_j. Bình luận mang nhãn 0 nếu người dùng thích hoặc khen khách sạn. Ngược lại, bình luận mang nhãn 1 nếu người dùng không thích hoặc chê khách sạn.

$$comment_{ij} = \begin{cases} 0 & \text{nếu } w_i \text{ thích } x_j \\ 1 & \text{nếu ngược lại} \end{cases} \tag{3.2}$$

3.1.3 Mô hình kết hợp hành vi đánh giá và hành vi bình luận

Để sử dụng dữ liệu hành vi đánh giá và bình luận cùng lúc cho tư vấn khách sạn thì cần có một phương pháp để kết hợp hai loại dữ liệu này. Như đã trình bày trong Phần 3.1.2: $rating_{ij}$: Điểm đánh giá từ người dùng w_i cho sản phẩm x_j và $comment_{ij}$: Bình luận bày tỏ quan điểm từ người dùng w_i cho sản phẩm x_j. Theo [9], các bình luận tiêu cực có ảnh hưởng không nhỏ tới quyết định mua hàng của người dùng. Tuy nhiên, nếu sản phẩm có nhiều phản hồi tích cực thì cũng làm tăng khả năng mua hàng của người dùng. Do đó, đồ án thực hiện kết hợp dựa trên ý tưởng: **"Nếu khách sạn có bình luận tiêu cực thì điểm đánh giá dành cho khách sạn này cần phải hạ xuống. Tuy nhiên, khách sạn có nhiều phản hồi tích cực thì điểm đánh giá cũng cần được tăng lên"**. Điều này có nghĩa là, nếu khách sạn có nhiều phản hồi tích cực thì các điểm đánh giá dành cho khách sạn này sẽ được thưởng thêm và ngược lại, nếu khách sạn có nhiều bình luận phàn nàn thì điểm đánh giá sẽ bị trừ đi.

Với ý tưởng trên, điểm đánh giá dự đoán $rating_{ij}$ của người dùng w_i với khách sạn x_j , tỷ lệ số bình luận tích cực p_rate_j, tỷ lệ số bình luận tiêu cực n_rate_j của khách sạn x_j sẽ là 3 thành phần quyết định tới điểm đánh giá cuối cùng dành cho khách sạn. Coi điểm đánh giá cuối cùng là 100%, α và β là 2 trọng số tương ứng của $rating_{ij}$, p_rate_j và n_rate_j quyết định mức độ ảnh hưởng của 2 thành phần này lên điểm đánh giá cuối cùng. Như vậy, điểm đánh giá cuối cùng của người dùng có thể biểu

diễn theo công thức:

$$c\_rating(w_i, x_j) = \alpha \times rating_{ij} + \beta \times (p\_rate_j - n\_rate_j) \tag{3.3}$$

Trong đó:

- $c\_rating(w_i, x_j) \in [0; 5]$ là điểm đánh giá kết hợp, được sử dụng để làm dữ liệu thực hiện huấn luyện và đánh giá
- $rating_{ij} \in [0; 5]$ là điểm đánh giá được dự đoán thông qua thuật toán lọc cộng tác,
- $p\_rate_j \in [0; 1]$ là tỉ lệ số $comment_{ij} = 0$ trong tổng số các $comment_{ij}$ của sản phẩm x_j
- $n\_rate_j \in [0; 1]$ là tỉ lệ số $comment_{ij} = 1$ trong tổng số các $comment_{ij}$ của sản phẩm x_j
- Do $rating_{ij}$ nằm trong khoảng giá trị khác với p_rate_j và n_rate_j vì vậy, trước khi thực hiện kết hợp, đồ án thực hiện chuyển $rating_{ij}$ về cùng khoảng giá trị $[0; 1]$ với p_rate_j và n_rate_j bằng cách $rating_{ij} = \frac{rating_{ij}}{5}$
- Sau khi thực hiện tính toán, để đưa điểm đánh giá dự đoán về khoảng ban đầu, ta chỉ cần thực hiện $c\_rating(w_i, x_j) = c\_rating(w_i, x_j) \times 5$

Chương 4

Thực nghiệm, kết quả, so sánh và đánh giá

Thực nghiệm, kết quả, so sánh và đánh giá

Chương 5

Kết luận

Kết luận

Tài liệu tham khảo

[1] Thomas Bayes. "Naive bayes classifier". In: *Article Sources and Contributors* (1968), pp. 1–9.

[2] James Davidson et al. "The YouTube video recommendation system". In: *Proceedings of the fourth ACM conference on Recommender systems*. 2010, pp. 293–296.

[3] Angela Edmunds and Anne Morris. "The problem of information overload in business organisations: a review of the literature". In: *International journal of information management* 20.1 (2000), pp. 17–28.

[4] Mahesh Goyani and Neha Chaurasiya. "A review of movie recommendation system: Limitations, Survey and Challenges". In: *ELCVIA: electronic letters on computer vision and image analysis* 19.3 (2020), pp. 0018–37.

[5] Marti A. Hearst et al. "Support vector machines". In: *IEEE Intelligent Systems and their applications* 13.4 (1998), pp. 18–28.

[6] Folasade Olubusola Isinkaye, Yetunde O Folajimi, and Bolande Adefowoke Ojokoh. "Recommendation systems: Principles, methods and evaluation". In: *Egyptian informatics journal* 16.3 (2015), pp. 261–273.

[7] Thụy Hà Quang Nam Nguyễn Hà Thành Nguyễn Trí. "Giáo trình khai phá dữ liệu". In: *Giáo trình khai phá dữ liệu* (2012).

[8] John Platt. "Sequential minimal optimization: A fast algorithm for training support vector machines". In: (1998).

[9] Luming Yang, Min Xu, and Lin Xing. "Exploring the core factors of online purchase decisions by building an E-Commerce network evolution model". In: *Journal of Retailing and Consumer Services* 64 (2022), p. 102784.