HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Lê Tuấn Anh

NGHIÊN CỨU PHƯƠNG PHÁP HỌC SÂU CHO HỆ TƯ VẤN

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

HÀ NỘI – 2024

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Lê Tuấn Anh

NGHIÊN CỨU PHƯƠNG PHÁP HỌC SÂU CHO HỆ TƯ VẤN

Chuyên ngành: Hệ thống thông tin
Mã số: 8.48.01.04

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

NGƯỜI HƯỚNG DẪN KHOA HỌC
TS. NGUYỄN DUY PHƯƠNG

HÀ NỘI - 2024

LỜI CAM ĐOAN

Tôi cam đoan rằng công trình nghiên cứu này là thành quả của công sức cá nhân của tôi và không sao chép từ bất kỳ nguồn nào khác. Tất cả thông tin được trình bày trong đề án này đều là sản phẩm của công việc cá nhân hoặc được tổng hợp từ nhiều nguồn tài liệu khác nhau. Mọi tài liệu tham khảo đều được trích dẫn một cách hợp pháp và có nguồn gốc rõ ràng. Các dữ liệu và kết quả được trình bày trong đề án đều là trung thực và chưa từng được công bố trong bất kỳ công trình nghiên cứu nào khác.

Tác giả đề án

Lê Tuấn Anh

LỜI CẢM ƠN

Lời đầu tiên, em xin gửi lời cảm ơn sâu sắc tới các thầy, cô giáo giảng viên khoa Công nghệ thông tin 1, khoa Đào tạo Sau đại học nói riêng và các thầy, cô giáo giảng viên Học viện Công nghệ Bưu chính Viễn Thông nói chung. Trong suốt quá trình học tập tại Học viện, các thầy cô đã chỉ bảo, giảng dạy cho em biết bao kiến thức, kinh nghiệm quý báu để em có hành trang vững bước trong tương lai.

Em cũng xin được gửi lời cảm ơn tới thầy hướng dẫn TS. Nguyễn Duy Phương, cảm ơn thầy đã luôn hướng dẫn chỉ bảo tận tình em trong suốt quá trình học tập, nghiên cứu và thực hiện đề án này. Những lời khuyên, sự chỉ bảo của các thầy đã giúp em hoàn thành đề án tốt nghiệp này cũng như có thêm rất nhiều kiến thức, kinh nghiệm trong việc học tập và nghiên cứu.

Dù đã nỗ lực hoàn thành đề án, em hiểu rằng có thể không tránh khỏi những sai sót. Kính mong được thầy cô và các bạn thông cảm và đóng góp ý kiến.

Em xin trân trọng cảm ơn.

MỤC LỤC

DANH MỤC TỪ VIẾT TẮT

STT	Từ viết tắt	Tiếng Anh	Tiếng Việt/Giải thích
1	RS	Recommender System	Hệ thống tư vấn
2	CF	Collaborative Filtering	Lọc cộng tác
3	MF	Matrix Factorization	Hệ số hóa ma trận
4	NCF	Neural Collaborative Filtering	Lọc cộng tác với mạng nơ-ron
5	MLP	Multi-layer Perceptron	Perceptron nhiều lớp
6	ML	Machine Learning	Học máy
7	DL	Deep Learning	Học sâu

DANH SÁCH CÁC BẢNG

DANH MỤC CÁC HÌNH

MỞ ĐẦU

Trong những năm gần đây, với sự phát triển nhanh chóng của công nghệ công nghệ, sự bùng nổ dữ liệu ngày càng trở nên mạnh mẽ. Khi khối lượng dữ liệu tăng lên, các cá nhân phải đối mặt với vấn đề thừa thông tin, khiến việc đưa ra quyết định đúng đắn trở nên khó khăn hơn. Hiện tượng này được gọi là quá tải thông tin. Sử dụng kỹ thuật trí tuệ nhân tạo để thu thập thông tin trừu tượng từ dữ liệu lớn (big data) và chuyển đổi nó thành tri thức hữu ích là một trong những vấn đề cốt lõi trong phân tích dữ liệu lớn. Để giải quyết vấn đề quá tải thông tin, hệ thống gợi ý ra đời theo yêu cầu của thời đại.

Tuy nhiên với sự phổ biến ngày càng lớn của dữ liệu và ứng dụng trực tuyến, hệ tư vấn đối diện với nhiều thách thức. Ngày càng khó khăn để xử lý và phân tích lượng lớn dữ liệu, đồng thời cần tìm ra cách cá nhân hóa đề xuất sao cho phù hợp và chính xác. Đây là một trong những lý do phương pháp học sâu nổi lên như một công cụ mạnh mẽ để giải quyết những thách thức này. Trong lĩnh vực học sâu, sử dụng mạng nơ-ron (neural) đã trở thành một phương pháp quan trọng để xử lý thông tin tương tác phức tạp giữa các thực thể trong hệ thống gợi ý. Học sâu mang lại khả năng học và rút trích đặc trưng từ dữ liệu lớn, xử lý thông tin phức tạp, và từ đó, cải thiện độ chính xác và khả năng đề xuất của hệ tư vấn.

Hướng nghiên cứu hiện tại của hệ thống tư vấn có thể được chia thành ba loại [5]: khuyến nghị dựa trên nội dung (Content-based Filtering Recommendation), khuyến nghị dựa trên lọc cộng tác (Collaborative Filtering Recommendation) và phương pháp khuyến nghị kết hợp (Hybrid Recommendation). Lọc thông tin theo nội dung khai thác những khía cạnh liên quan đến nội dung thông tin sản phẩm hoặc người dùng đã từng tương tác trong quá khứ để tạo nên tư vấn. Trái lại, lọc thông tin theo cộng tác khai thác những khía cạnh liên quan đến thói quen sở thích của người sử dụng sản phẩm để đưa ra dự đoán và phân bổ các sản phẩm cho người dùng này. Các phương pháp khuyến nghị kết hợp tìm cách đạt được kết quả khuyến nghị tốt nhất bằng cách kết hợp các phương pháp khuyến nghị dựa trên nội dung và các phương pháp khuyến nghị dựa trên lọc cộng tác. Trong bối cảnh của việc cải thiện hệ

thống gợi ý hiện nay, sự phát triển của các phương pháp lọc cộng tác (*Collaborative filtering*) đã đem lại một cách tiếp cận mạnh mẽ và linh hoạt hơn trong việc cung cấp gợi ý dựa trên hành vi và sở thích của người dùng.

Các phương pháp lọc cộng tác có thể được phân thành hai loại Dựa trên bộ nhớ (Memory-Based) và Dựa trên mô hình (Model-Based) [6]. Lọc cộng tác dựa trên mô hình cho kết quả tốt hơn lọc cộng tác dựa vào bộ nhớ.

Chính vì những ưu điểm đã nêu ra ở trên của các phương pháp, em đã lựa chọn đề tài: "*Nghiên cứu phương pháp học sâu cho hệ tư vấn*", phương pháp sẽ xây dựng hệ tư vấn lọc cộng tác dựa trên mạng nơ-ron lọc cộng tác (Neural Collaborative Filtering). Hy vọng rằng nghiên cứu này sẽ mang lại những hiểu biết sâu hơn và khám phá mới trong lĩnh vực này, tạo ra những đóng góp ý nghĩa cho cộng đồng nghiên cứu và mang lại giá trị thực tế cho người dùng.

Nội dung đề án được trình bày thành ba chương theo cấu trúc sau:

Chương 1: Tổng quan về hệ tư vấn

Trình bày tổng quan về các khái niệm liên quan cơ bản tới hệ tư vấn và đưa ra các phương pháp tiếp cận chính, các ưu và nhược điểm của từng phương pháp này được sử dụng để xây dựng hệ thống tư vấn.

Chương 2: Học sâu cho hệ tư vấn lọc cộng tác

Giới thiệu tổng quan về học sâu, các khái niệm đặc trưng của phương pháp phân rã ma trận, ưu và nhược điểm của phương pháp này. Trình bày chi tiết cách xây dựng mô hình Neural Collaborative Filtering (NCF) để giải quyết bài toán.

Chương 3: Thực nghiệm và kết quả

Xây dựng bộ dữ liệu từ dữ liệu thực tế, trình bày quá trình cài đặt thử nghiệm, so sánh hiệu suất của phương pháp Neural Collaborative Filtering (NCF) với một số phương pháp hiện có.

Kết luận và hướng phát triển.

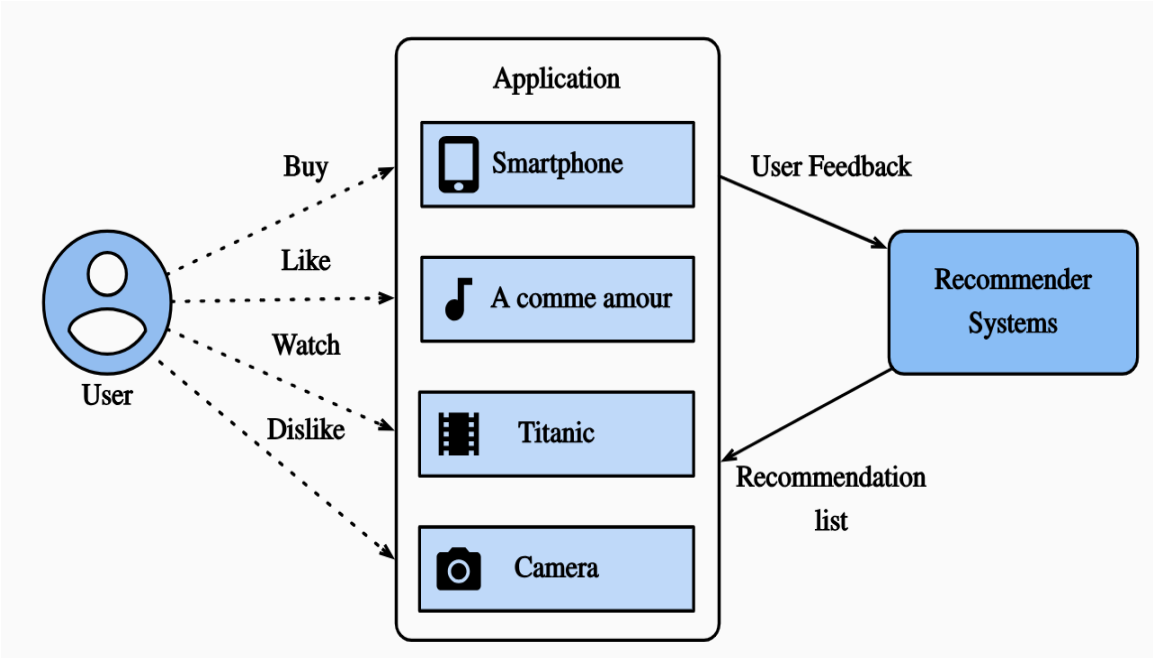
Trình bày tóm tắt những kết quả đã đạt và chưa đạt được. Qua đó đề xuất mục tiêu, hướng nghiên cứu, cũng như hướng phát triển tiếp theo.

CHƯƠNG 1. TỔNG QUAN VỀ HỆ TƯ VẤN

Chương này sẽ trình bày tổng quan về các khái niệm liên quan cơ bản tới hệ tư vấn và đưa ra các phương pháp tiếp cận chính, các ưu và nhược điểm của từng phương pháp này được sử dụng để xây dựng hệ thống tư vấn. Cuối cùng sẽ đưa ra các tiêu chuẩn để đánh giá hiệu quả của một hệ tư vấn.

1.1. Giới thiệu về hệ tư vấn

Trong xã hội ngày nay, vai trò của hệ tư vấn trở nên quan trọng không thể thiếu trong cuộc sống hàng ngày. Internet phổ biến và sự phát triển mạnh mẽ của công nghệ mang lại một lượng lớn thông tin và tài nguyên, làm tăng cường sự cần thiết của các hệ thống tư vấn. Những hệ thống này giúp người dùng tìm kiếm, khám phá và lựa chọn thông tin một cách hiệu quả. Hệ tư vấn trở thành giải pháp linh hoạt và hiệu quả để đối mặt với vấn đề bùng nổ thông tin do sự phát triển nhanh chóng của các dịch vụ Internet, và chúng được áp dụng rộng rãi trong nhiều lĩnh vực. Nó không chỉ giúp người dùng tiết kiệm thời gian và công sức trong việc tìm kiếm thông tin phù hợp với nhu cầu cá nhân mà còn mang đến trải nghiệm cá nhân hóa, tối ưu và thú vị.



Hình 1. 1: Hệ tư vấn

1.1.1. Giới thiệu bài toán tư vấn

Cho một tập hợp hữu hạn $U = \{u_1,\ u_2, \dots, u_n\}$ là tập gồm N người dùng (người sử dụng hệ thống), $I = \{i_1,\ i_2, \dots, i_k\}$ là tập gồm K sản phẩm (sản phẩm của hệ thống). Mỗi sản phẩm $i_k \in I$ có thể là sản phẩm hàng hóa, tài liệu, sách, báo, hoặc bất kể dạng thông tin nào mà người dùng quan tâm.

Ma trận đánh giá $A = \{a_{ij},\ i = 1, \dots,\ N,\ j = 1, \dots,\ K\}$ dùng để biểu diễn mối quan hệ giữa tập người dùng U và tập sản phẩm I. Mỗi giá trị $a_{ij} \in \{0,1,2, \dots, V\}$ thể hiện đánh giá của người dùng $u_i \in U$ đối với sản phẩm $i_j \in I$. Giá trị của a_{ij} có thể thu thập trực tiếp từ ý kiến của người dùng hoặc thu thập một cách gián tiếp thông qua các cơ chế phản hồi của người dùng. Giá trị $a_{ij} = 0$ có thể hiểu rằng người dùng u_i chưa bao giờ biết đến hoặc chưa đánh giá sản phẩm i_j.

Nhiệm vụ của hệ thống gợi ý là dựa trên những dữ liệu đã có, đưa ra những gợi ý về sản phẩm $i_j \in I$ mà người dùng $u_i \in U$ có khả năng sẽ quan tâm.

1.1.2. Một số khái niệm chung về hệ thống tư vấn

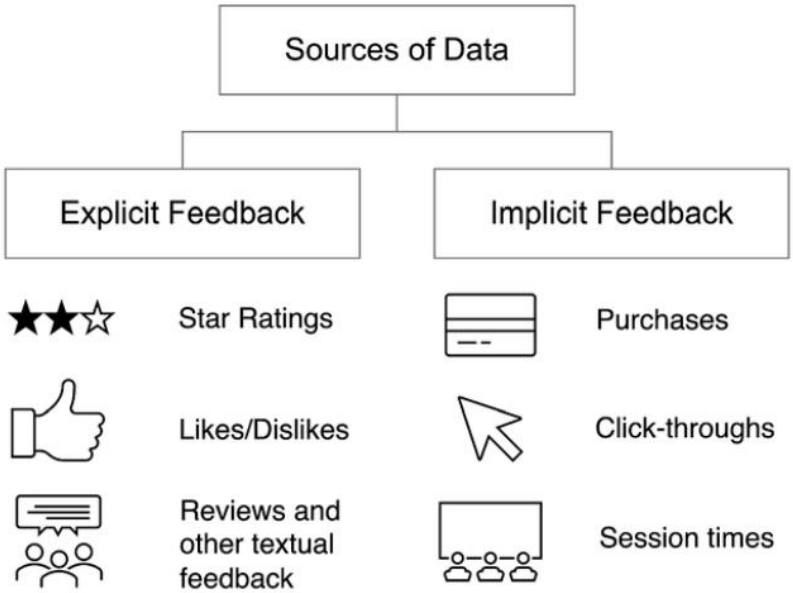
Hệ thống tư vấn, còn được gọi là Recommender System hoặc Recommendation System [15], là một loại công nghệ thông tin được thiết kế để tự động đề xuất các mục hoặc sản phẩm mà có thể phù hợp và được ưa thích nhất với mỗi người dùng cá nhân. Với sự phát triển mạnh mẽ của internet và dữ liệu, hệ thống tư vấn đã trở thành một phần quan trọng trong nhiều ứng dụng trực tuyến, từ thương mại điện tử đến giải trí và mạng xã hội. Mục tiêu chính của hệ thống tư vấn là cung cấp các gợi ý cá nhân hóa, giúp người dùng khám phá và tiêu thụ nội dung mới một cách hiệu quả. Để đạt được điều này, hệ thống tư vấn sử dụng một loạt các phương pháp và thuật toán, từ lọc dựa trên nội dung đến lọc dựa trên hành vi và kết hợp cả hai để tạo ra các gợi ý chính xác và phù hợp. Các hệ thống tư vấn thường sử dụng thông tin từ lịch sử hoạt động của người dùng, bao gồm các mục họ đã mua, xem, hoặc đánh giá, để đề xuất các mục tương tự hoặc phù hợp với sở thích của họ. Ngoài ra, hệ thống cũng có thể sử dụng thông tin về hồ sơ người dùng, như tuổi, giới tính, hoặc địa điểm, để tăng tính cá nhân hóa của các gợi ý. Với khả năng tùy chỉnh và cá

nhân hóa cao, hệ thống tư vấn không chỉ giúp cải thiện trải nghiệm của người dùng mà còn tạo ra cơ hội kinh doanh và tiếp thị cho các doanh nghiệp, bằng cách tối ưu hóa việc tiếp cận và tương tác với khách hàng.

Sản phẩm (*Item*) là thuật ngữ chung để chỉ những thứ mà người dùng có thể tương tác trong hệ thống tư vấn. Item có thể là sách, phim, truyện, tin tức... Thông thường, hệ thống tư vấn được thiết kế để phù hợp với một loại hình sản phẩm cụ thể, nhằm đảm bảo rằng quá trình tư vấn đối với mỗi sản phẩm đều đạt được hiệu suất cao và đáp ứng đúng nhu cầu của người dùng.

Trong thực tế, việc thu thập dữ liệu Hồ sơ người dùng thường được sử dụng thông qua hai phương pháp chính là phản hồi ẩn (implicit feedback) và phản hồi tường minh (explicit feedback). Đối với phương pháp phản hồi tường minh hệ thống yêu cầu người dùng thực hiện việc xếp hạng (rating) cụ thể cho mỗi sản phẩm để xây dựng Hồ sơ người dùng. Phương pháp này cung cấp dữ liệu người dùng trực tiếp cho hệ thống tư vấn mà không cần các bước biến đổi trung gian, và kết quả tư vấn từ đó được đánh giá là đáng tin cậy hơn [2]. Mặc dù được coi là mang lại kết quả tư vấn đáng tin cậy, nhưng phương pháp này đòi hỏi người dùng phải thực hiện thêm các thao tác với hệ thống, có thể làm giảm trải nghiệm người dùng trong một số trường hợp. Ngoài ra, tâm lý chung của người dùng không muốn chia sẻ quá nhiều thông tin cá nhân, điều này làm cho việc triển khai thực tế của phương pháp thu thập phản hồi tường minh thường gặp khó khăn.

Để khắc phục những mặt hạn chế của việc thu thập dữ liệu hồ sơ người dùng của phương pháp phản hồi tường minh, phương pháp thu thập phản hồi ẩn sử dụng/ghi nhận các dấu vết mà người dùng có thể để lại trên hệ thống như lịch sử truy cập vào website, lịch sử xem hoặc mua sản phẩm, thời gian truy cập trang web, số lần nhấp chuột, và các hoạt động khác tương tự để có thể suy luận các thông tin về sở thích của họ. Phương pháp này giúp cải thiện trải nghiệm của người dùng với hệ thống. Tuy nhiên, khả năng mô tả sở thích của người dùng bằng cách này được coi là không tốt bằng phương pháp phản hồi tường minh vì hệ thống bắt buộc phải thực hiện qua các bước biến đổi trung gian để có thể trích xuất thông tin từ hành vi của người dùng.



Hình 1. 2: Phản hồi ẩn và phản hồi tường minh

Ma trận tương tác Người dùng – Sản phẩm hay còn được gọi là *Utility Matrix* hoặc *User – Item matrix* là một cơ sở dữ liệu mô tả sở thích của mỗi Người dùng (User) với từng Sản phẩm (Item) trong hệ thống. Dữ liệu này có thể được biểu diễn dưới dạng ma trận, trong đó mỗi hàng đại diện cho một người dùng (User), mỗi cột đại diện cho một sản phẩm (Item), và giá trị tại mỗi ô của ma trận thể hiện giá trị đánh giá (rating) của người dùng đó cho sản phẩm tương ứng.

	Sản phẩm 1	Sản phẩm 2	Sản phẩm 3	...	Sản phẩm n
Người dùng 1	1	5	...	3	...
Người dùng 2	...		2	...	4
...	2	1	...	5	...
Người dùng m	3	...	1	...	1

R

Hình 1. 3: Ma trận tương tác Người dùng – Sản phẩm

Bắt nguồn từ yếu tố tâm lý, người dùng thường không đánh giá tất cả các sản phẩm (Item) họ đã trải nghiệm, thường chỉ đưa ra đánh giá tích cực (thích) hoặc tiêu cực (không thích) đối với các sản phẩm. Vì lý do trên nên thực tế ma trận tương tác Người dùng – Sản phẩm sẽ thường bị thiếu giá trị ở rất nhiều vị trí, vấn đề này được gọi là vấn đề dữ liệu thưa (*data sparsity*) [3]. Vấn đề dữ liệu thưa của ma trận tương tác (*User – Item matrix)* dẫn đến việc các thuật toán tư vấn dựa trên bộ nhớ (*memory-based recommendation algorithms*) [2] hoạt động một cách không hiệu quả. Bên cạnh đó, trong các trường hợp trong hệ thống xuất hiện những người dùng (User) hoặc sản phẩm (Item) mới chưa có bất kỳ tương tác nào thì ma trận User – Item sẽ xuất hiện các hàng/cột không có giá trị. Vấn đề nêu trên được gọi là vấn đề khởi động nguội (*cold-start problem*) [3], vấn đề này thường xảy ra đối với các hệ tư vấn dựa/khuyến nghị dựa trên cộng tác, khi mà hệ thống gặp phải những sản phẩm hoàn toàn không có bất kỳ tương tác nào, hệ tư vấn sẽ không thể khai thác được các thông tin như sở thích giống nhau giữa các nhóm người dùng/sản phẩm để có thể đưa ra tư vấn.

1.1.3. Các tiêu chí đánh giá hệ tư vấn

1.1.3.1. Phương pháp đánh giá hệ tư vấn

Để có thể đánh giá được độ chính xác của hệ thống tư vấn/khuyến nghị, đầu tiên từ ma trận đánh giá R, chúng ta tiến hành chia tập người dùng U (các hàng trong ma trận đánh giá R) thành hai phần, một phần ký hiệu là U_{train} được sử dụng làm dữ liệu huấn luyện (training), phần còn lại ký hiệu là U_{test} được sử dụng để kiểm tra (testing) sao cho $U_{train} \cup U_{test} = U$ và $U_{train} \cap U_{test} = \emptyset$ Tập dữ liệu huấn luyện U_{train} được sử dụng để xây dựng mô hình theo các thuật toán sử dụng trong hệ tư vấn/khuyến nghị. Tập dữ liệu kiểm tra U_{test} được sử dụng vào quá trình kiểm nghiệm thuật toán tư vấn. Chúng ta có thể biết đến một số cách tiếp cận thưởng được sử dụng để chia tập người dùng U thành 2 phần huấn luyện (U_{train})và kiểm tra (U_{test}) là:

- Lấy mẫu Bootstrap *(Bootstrap sampling*)
- Phân chia (*Splitting*)
- Kiểm thử chéo (*k-fold cross validation*).

1.1.3.2. Độ đo đánh giá độ chính xác của đánh giá dự đoán

Để đánh giá tính chính xác của các giá trị dự đoán từ hệ tư vấn, một trong những phương pháp phổ biến là sử dụng các độ đo dựa trên độ sai số giữa giá trị dự đoán và giá trị thực tế. Điều này giúp đo lường mức độ chính xác của dự đoán và đưa ra cái nhìn tổng quan về hiệu suất của hệ thống. Dưới đây là một số độ đo phổ biến được sử dụng để đánh giá sai số trong các bài toán phân loại:

- **Sai số trung bình tuyệt đối (Mean Absolute Error - MAE)**: Đây là độ đo tính toán trung bình của giá trị tuyệt đối của các sai số giữa dự đoán và giá trị thực tế. MAE cho biết mức độ lớn nhỏ của sai số mà hệ thống tư vấn gặp phải trong dự đoán.
- **Độ đo trung bình lỗi lấy căn (Root Mean Square Error - RMSE):** RMSE tính toán căn bậc hai của trung bình của bình phương của các sai số giữa dự đoán và giá trị thực tế. Điều này có ý nghĩa là sai số lớn hơn sẽ được đánh giá cao hơn và phản ánh mức độ biến động của sai số.
- **Sai số bình phương trung bình (Mean Squared Error - MSE):** MSE tính bình phương của sai số giữa giá trị dự đoán và giá trị thực tế, sau đó lấy trung bình của các bình phương sai số đó.

Cả hai độ đo này đều cung cấp thông tin quan trọng về mức độ chính xác của dự đoán, giúp người phát triển và người sử dụng hệ tư vấn có cái nhìn tổng quan về hiệu suất của hệ thống. Tùy thuộc vào bản chất của dữ liệu và yêu cầu cụ thể của bài toán, các độ đo này có thể được sử dụng một cách linh hoạt để đánh giá và cải thiện hiệu suất của hệ tư vấn.

1.1.3.3. Độ đo đánh giá độ chính xác của danh sách sản phẩm tư vấn

Các độ đo phổ biến được sử dụng để đánh giá độ chính xác của danh sách sản phẩm tư vấn cung cấp một cái nhìn chi tiết về hiệu suất của hệ thống, đặc biệt là trong các hệ thống tư vấn sản phẩm hoặc nội dung. Dưới đây là một số độ đo quan trọng:

- **Độ chính xác (Precision)**: Đây là tỷ lệ giữa số lượng sản phẩm được đề xuất chính xác đến số lượng sản phẩm được đề xuất tổng cộng. Độ chính xác cao chỉ ra rằng số lượng sản phẩm đề xuất là chính xác và thích hợp.
- **Độ nhạy (Recall)**: Được tính bằng tỷ lệ giữa số lượng sản phẩm được đề xuất chính xác đến số lượng sản phẩm thực sự liên quan. Recall cao chỉ ra rằng hệ thống có khả năng đề xuất ra tất cả các sản phẩm liên quan một cách toàn diện.
- **E-measure và F-measure**: Đây là hai độ đo kết hợp giữa precision và recall, cung cấp một cái nhìn tổng quan về hiệu suất của hệ thống. E-measure và F-measure thường được sử dụng khi cần cân nhắc cả precision và recall một cách cân bằng.
- **Độ chính xác trung bình tuyệt đối MAP (Mean Average Precision)**: Đây là độ đo phổ biến được sử dụng trong các bài toán đề xuất hàng loạt (recommendation) để đánh giá hiệu suất của hệ thống. MAP tính toán precision cho mỗi ngưỡng của danh sách đề xuất và sau đó tính trung bình các precision này. Điều này giúp đánh giá độ chính xác của hệ thống trên toàn bộ danh sách đề xuất.

Bằng cách kết hợp các độ đo này, người phát triển và người sử dụng có thể có cái nhìn tổng quan và chi tiết về hiệu suất của hệ thống tư vấn, từ đó đưa ra các điều chỉnh và cải thiện để tăng cường trải nghiệm người dùng và chất lượng của các đề xuất sản phẩm.

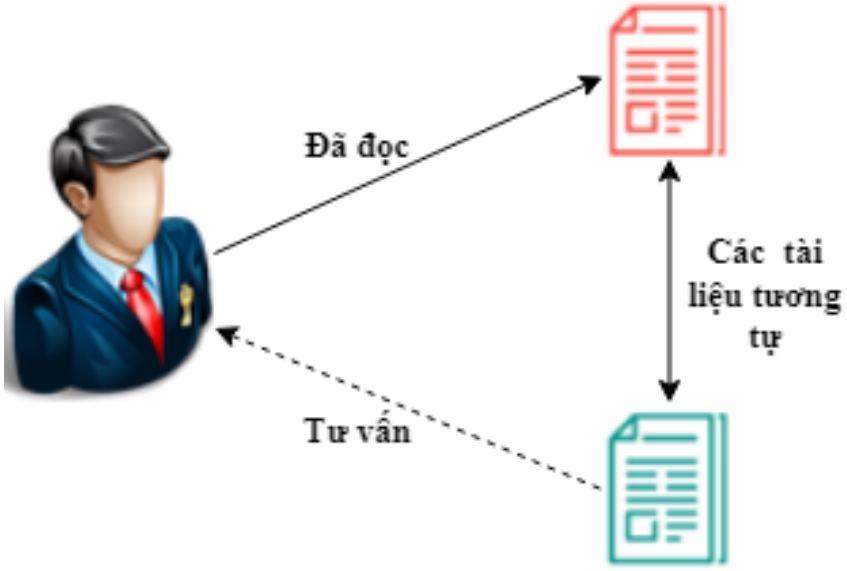
1.2. Các hướng tiếp cận trong hệ tư vấn

Hướng nghiên cứu hiện tại của hệ thống tư vấn có thể được chia thành ba loại [5]: tư vấn dựa trên lọc theo nội dung (*Content-Base Filtering*), tư vấn dựa trên lọc cộng tác (*Collaborative Filtering*) và tư vấn dựa trên lọc kết hợp (*Hybrid Filtering*).

1.2.1. Lọc theo nội dung (Content-Base Filtering)

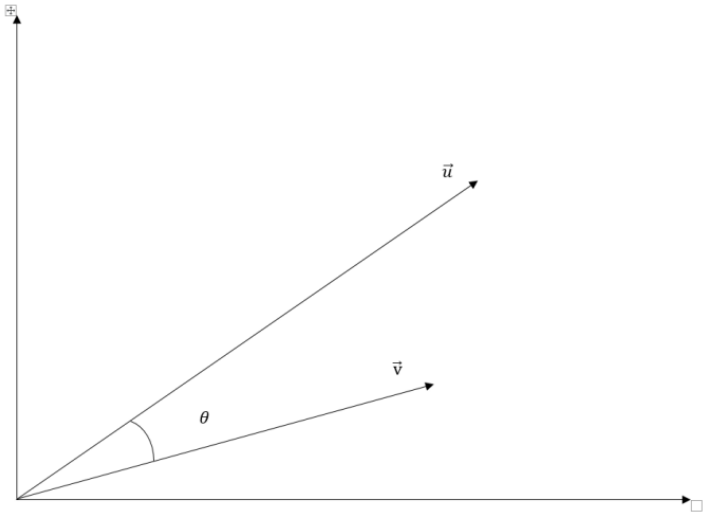
Phương pháp khuyến nghị dựa trên nội dung *(Content-Base Filtering)* sử dụng nội dung của các sản phẩm và tìm ra điểm tương đồng giữa chúng. Sau khi phân tích

đủ số lượng sản phẩm mà một người dùng đã thể hiện sự yêu thích, hồ sơ sở thích của người dùng sẽ được thiết lập. Sau đó, hệ thống tư vấn có thể tìm kiếm cơ sở dữ liệu và chọn các sản phẩm phù hợp theo hồ sơ [4]. Khó khăn của các thuật toán này nằm ở cách tìm kiếm sở thích của người dùng dựa trên nội dung của các sản phẩm. Khi các vấn đề về quyền riêng tư cá nhân ngày càng nhận được nhiều sự quan tâm của người dùng Internet, việc thu thập hồ sơ người dùng cho các phương pháp khuyến nghị dựa trên nội dung ngày càng trở nên khó khăn hơn.



Hình 1. 4: Cơ chế hoạt động lọc theo nội dung

Hệ thống tư vấn ghi nhận Hồ sơ người dùng (User Profile) dưới dạng vector $\vec{u} = (u_1, u_2, \ldots, u_n)$, trong đó u_i là trọng số thể hiện mức độ quan tâm của người dùng đối với từng thuộc tính của sản phẩm. Vector Hồ sơ sản phẩm (Item Profile) $v = (v_1, v_2, \ldots, v_n)$ biểu diễn thông tin sản phẩm thông qua các thuộc tính v_i. Mức độ phù hợp giữa sở thích của người dùng và sản phẩm được đánh giá dựa trên góc lệch giữa hai vector Hồ sơ người dùng (u) và vector Hồ sơ sản phẩm (v).



Hình 1. 5: Độ tương tự giữa hai vector

Để đánh giá mức độ tương đồng giữa 2 vector $\vec{u}$ và $\vec{v}$ để có thể đưa ra tư vấn, hệ thống thực hiện so sánh bằng cosine góc lệch giữa 2 vector:

$$\text{similarity } = \cos(\theta) = \frac{\vec{u}\vec{v}}{\|\vec{u}\| \cdot \|\vec{v}\|} = \frac{\sum_{i=1}^{n} u_i \cdot v_i}{\sqrt{\sum_{i=1}^{n} u_i^2} \cdot \sqrt{\sum_{i=1}^{n} v_i^2}} \quad (1)$$

Những ưu điểm của phương pháp tư vấn dựa trên nội dung:

- Hệ thống không cần yêu cầu phải có quá nhiều dữ liệu từ những người dùng khác để đạt được độ chính xác tư vấn chấp nhận được. Đối với phương pháp lọc dựa trên cộng tác hệ thống tư vấn cần phải phân tích toàn bộ dữ liệu tương tác để tìm ra được các quy luật (như sản phẩm được người dùng ưa thích hoặc để ý đến) trước khi có thể đưa ra gợi ý/đề xuất. Ngược lại với phương pháp lọc dựa trên nội dung, hệ thống chỉ sử dụng thông tin về nội dung của sản phẩm và lịch sử tương tác của từng người dùng cụ thể để tạo ra các gợi ý.
- Dựa trên cơ chế hoạt động của hệ thống lọc nội dung, có thể cung cấp gợi ý về các sản phẩm mới dựa trên thông tin mô tả sẵn có, mà không cần phải dựa vào dữ liệu tương tác từ người dùng. Quá trình này dựa trên việc phân tích nội dung của các sản phẩm để tạo ra các gợi ý. Khi một sản phẩm mới xuất hiện, nếu nội dung của nó "tương tự" với các sản phẩm trước đó mà người dùng ưa thích, hệ thống sẽ tự động gợi ý cho người dùng về sản phẩm này.

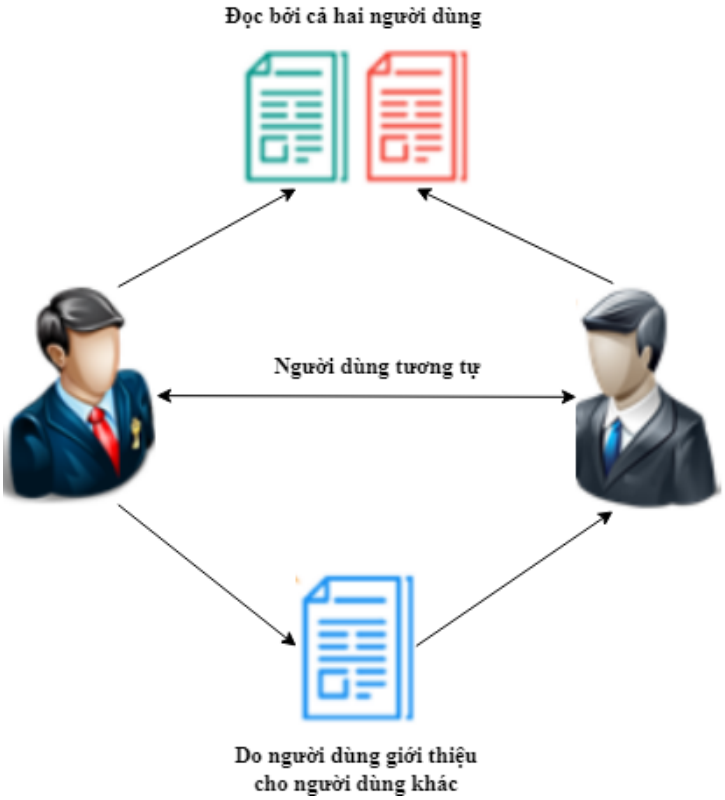
- Trong trường hợp người dùng có những sở thích đặc biệt dựa trên nội dung của sản phẩm, nhưng các sản phẩm này không phổ biến trong ma trận tương tác, hệ thống vẫn có khả năng khám phá được thông tin này thông qua nội dung.
- Mô hình này có khả năng cung cấp giải thích chi tiết về các yếu tố ảnh hưởng đến sở thích của người dùng thông qua các giá trị trọng số trong vector hồ sơ của họ. Các giá trị trọng số lớn hơn sẽ phản ánh mức độ quan tâm của người dùng đối với các yếu tố nội dung của sản phẩm. Điều này giúp tăng cường các đặc tính/nội dung phù hợp của nội dung sản phẩm với người dùng dựa trên thông tin này

Một số nhược điểm của phương pháp tư vấn dựa trên nội dung:

- *Tính đa dạng hạn chế:* Phương pháp này có thể gặp khó khăn trong việc đề xuất các mục có tính đa dạng cao. Do việc dựa vào nội dung để đề xuất, hệ thống thường chỉ đề xuất các mục tương tự về nội dung, dẫn đến việc hạn chế sự đa dạng trong các gợi ý.
- *Vấn đề Cold start:* Một khi một mục mới được thêm vào hệ thống, hoặc một người dùng mới tham gia, việc tạo ra gợi ý chính xác dựa trên nội dung có thể gặp khó khăn. Điều này là do không có đủ dữ liệu nào về mục đó hoặc người dùng đó để xây dựng một hồ sơ nội dung hoặc sở thích rõ ràng.
- *Hiện tượng "cứng nhắc" (Over-specialization):* Hệ thống dựa vào nội dung có thể dễ dàng bị mắc kẹt trong việc đề xuất các mục tương tự nhau, làm giảm trải nghiệm của người dùng. Nếu một người dùng đã tương tác với một số loại nội dung nhất định, hệ thống có thể chỉ đề xuất các mục tương tự mà không khám phá ra các sở thích mới của người dùng.
- *Vấn đề về mức độ chất lượng của dữ liệu*: Để áp dụng phương pháp lọc theo nội dung, cần có dữ liệu về nội dung của các sản phẩm, và đôi khi việc thu thập và đánh giá dữ liệu này có thể tốn kém và phức tạp. Điều này có thể dẫn đến việc sử dụng dữ liệu không chính xác hoặc không đủ để tạo ra các gợi ý chất lượng.

1.2.2. Lọc cộng tác (Collaborative Filtering)

Các phương pháp khuyến nghị dựa trên lọc cộng tác *(Collaborative Filtering)* [6] tận dụng tối đa thông tin hành vi và thông tin tùy chọn do người dùng tạo trước đây mà không sử dụng thông tin cá nhân của người dùng và thông tin mô tả sản phẩm, chẳng hạn như đánh giá của người dùng về sản phẩm để tạo sản phẩm được khuyến nghị. Sử dụng các kỹ thuật thống kê để tìm ra sự tương đồng giữa vectơ người dùng hoặc sản phẩm. Các phương pháp Collaborative Filtering có thể được phân thành hai loại Dựa trên bộ nhớ (Memory-Based) và Dựa trên mô hình (Model-Based) [7]. Các phương pháp khuyến nghị dựa trên lọc cộng tác cũng có một số hạn chế. Khi ma trận đánh giá rất thưa thớt, độ chính xác của khuyến nghị thường giảm xuống rất rõ ràng. Quan trọng hơn, nó không thể tạo khuyến nghị cho một sản phẩm mới chưa nhận được đánh giá của người dùng.



Hình 1. 6: Cơ chế hoạt động lọc cộng tác

1.2.2.1. Lọc cộng tác dựa trên bộ nhớ

Dựa trên giá trị đánh giá của người dùng trong ma trận Người dùng – Sản phẩm, độ tương đồng giữa người dùng hiện tại với những người dùng tương tự được tính theo hai bước như sau:

Bước 1: Hệ thống tính toán độ tương tự giữa các người dùng/sản phẩm.

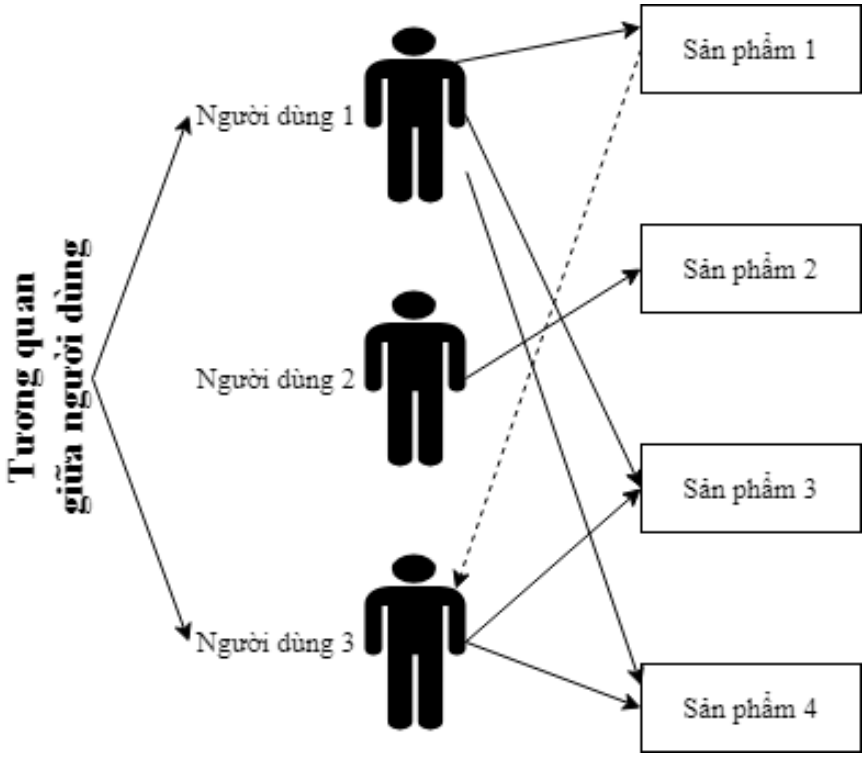
$$\text{similarity}(x,y) = \cos(\vec{x},\vec{y}) = \frac{\vec{x}\cdot\vec{y}}{\|\vec{x}\|_2 \times \|\vec{y}\|_2} = \frac{\Sigma r_{x,i}\cdot r_{y,i}}{\sqrt{\Sigma r_{x,i}^2}\sqrt{\Sigma r_{y,i}^2}} \quad (2)$$

Bước 2: Tính toán giá trị đánh giá dự đoán theo công thức:

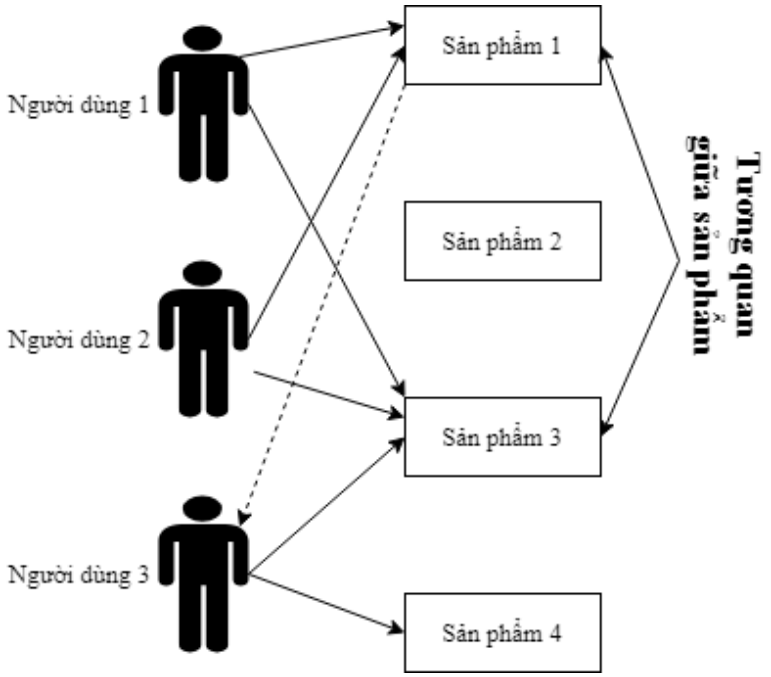
$$r(a,i) = \bar{r_a} + \frac{\sum_{i=1}^{n}\left(r_{u,i} - \bar{r_u}\right)\times \text{similarity}(x,y)}{\sum_{i=1}^{n}\text{similarity}(x,y)} \quad (3)$$

Kỹ thuật lọc cộng tác dựa trên bộ nhớ có thể thực hiện dựa trên người dùng và dựa trên đối tượng:

- *Kỹ thuật dựa trên người dùng:* Phương pháp này đo độ tương đồng giữa các người dùng bằng cách so sánh đánh giá (rating) của họ trên các sản phẩm tương tự. Đánh giá dự đoán cho sản phẩm hiện đang xem xét của người dùng được tính bằng cách lấy giá trị đánh giá (rating) trung bình của sản phẩm đó, nhân với trọng số phản ánh mức độ tương đồng của người dùng hiện tại với những người dùng có sở thích tương tự.
- *Kỹ thuật lọc dựa trên đối tượng*: Phương pháp này tính toán dự đoán đánh giá (rating) dựa trên sự tương đồng giữa các sản phẩm. Nó xây dựng một mô hình các sản phẩm tương tự bằng cách xem xét tất cả các sản phẩm đã được đánh giá bởi người dùng đang hoạt động từ ma trận tương tác. Sau đó, nó xác định mức độ tương đồng của các sản phẩm truy xuất đối với sản phẩm đích, chọn ra k sản phẩm tương tự nhất và tính toán dự đoán bằng cách lấy trung bình có trọng số của đánh giá người dùng trên các sản phẩm tương tự k.



Hình 1. 7: Lọc cộng tác dựa trên bộ nhớ thông qua người dùng



Hình 1. 8: Lọc cộng tác dựa trên bộ nhớ thông qua đối tượng

Ưu điểm:

- Hệ thống Recommendation System không yêu cầu phải có tri thức chuyên biệt để xây dựng mô hình người dùng/sản phẩm. Bản chất của phương pháp lọc

cộng tác là khai thác thông tin về sở thích của một nhóm người dùng có cùng sở thích đối với một số sản phẩm cụ thể. Do đó, khi nhận biết các sản phẩm được ưa thích, hệ thống Recommendation System có thể cung cấp gợi ý trực tiếp đến những sản phẩm đó mà không cần quan tâm đến nội dung cụ thể của chúng. Điều này giúp hệ thống mở rộng khả năng tư vấn cho nhiều loại người dùng có đặc điểm khác nhau.

- Có khả năng sử dụng thông tin từ người dùng có các đặc điểm tương tự nhau để cung cấp các gợi ý và tư vấn (phân tích xu hướng và sở thích của các nhóm người dùng tương đồng)

Phương pháp lọc cộng tác dựa trên bộ nhớ vẫn có những hạn chế sau:

- Hệ thống thường không thể cung cấp các gợi ý cụ thể phù hợp với sở thích riêng của từng người dùng, thay vào đó, nó thường đưa ra các gợi ý dựa trên các sản phẩm phổ biến trong hệ thống.
- Chất lượng tư vấn dựa trên phương pháp lọc cộng tác dựa vào bộ nhớ giảm hiệu suất khi dữ liệu tương tác trở nên quá thưa (data sparsity). Thực tế, người dùng thường chỉ đánh giá sản phẩm khi họ cảm thấy rất tích cực hoặc rất tiêu cực đối với sản phẩm đó, dẫn đến việc dữ liệu ma trận tương tác giữa Người dùng – Sản phẩm thường thiếu giá trị ở nhiều vị trí do người dùng không đưa ra đánh giá cho sản phẩm đó. Trong trường hợp này, việc sử dụng độ tương tự giữa các người dùng thông qua đo lường cosine trở nên không hiệu quả.
- Phương pháp lọc cộng tác dựa trên bộ nhớ đòi hỏi thời gian tính toán lâu và yêu cầu bộ nhớ lớn. Phương pháp này thuộc nhóm instance-based, có nghĩa là hệ thống phải xây dựng một mô hình riêng cho mỗi người dùng u, dựa trên toàn bộ dữ liệu ma trận tương tác, để tìm ra giá trị đánh giá (rating) trung bình $\bar{r_a}$ của tập người dùng có cùng sở thích với u trước khi dự đoán giá trị đánh giá.

1.2.2.2. Lọc cộng tác dựa trên mô hình

Để giải quyết các hạn chế về thời gian tính toán và yêu cầu dung lượng bộ nhớ lớn, hai tác vụ chính trong phương pháp lọc cộng tác đã được phân biệt rõ ràng. Cụ thể, có hai nhiệm vụ quan trọng: xây dựng mô hình từ dữ liệu huấn luyện và tính toán kết quả tư vấn từ mô hình đã được xây dựng. Để huấn luyện mô hình và tạo ra một mô hình rút gọn mô phỏng sở thích của người dùng, hệ thống tư vấn sử dụng các thuật toán học máy, có thể là không giám sát hoặc có giám sát. Các thuật toán học máy có thể bao gồm cây quyết định, bộ phân loại Bayes, hồi quy, máy vector hỗ trợ (SVM), mạng nơ-ron và các thuật toán khác. Khi đã có mô hình từ quá trình huấn luyện, hệ thống RS sử dụng mô hình này trực tiếp để sản sinh ra kết quả tư vấn.

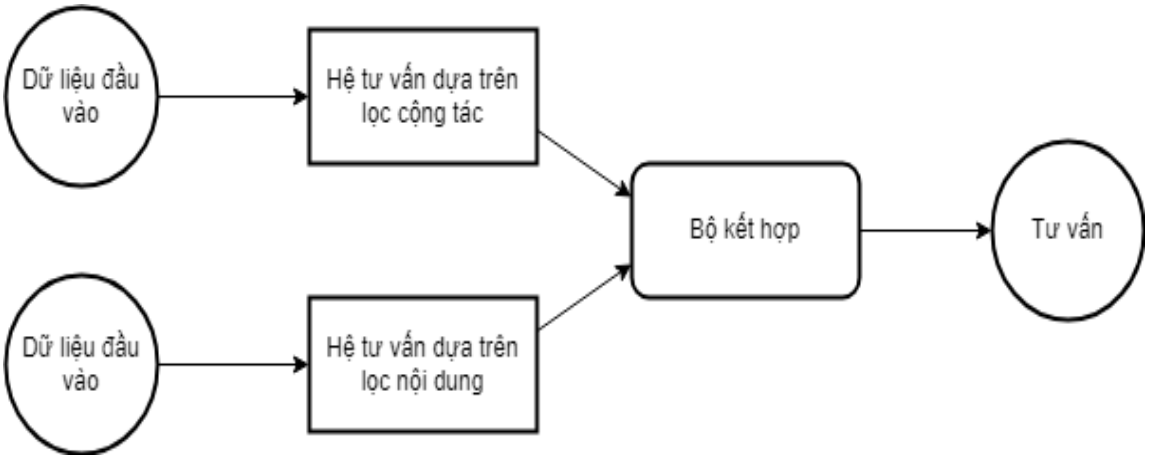
Ưu điểm so với phương pháp lọc cộng tác dựa trên bộ nhớ:

- *Độ linh hoạt cao*: Phương pháp lọc cộng tác dựa trên mô hình thường linh hoạt hơn trong việc xử lý các vấn đề như dữ liệu thưa (sparse data) hoặc dữ liệu lớn. Thay vì phụ thuộc hoàn toàn vào dữ liệu người dùng và sản phẩm, phương pháp này thường sử dụng các mô hình thống kê hoặc học máy để dự đoán các đánh giá hoặc sở thích của người dùng.
- *Giảm thiểu vấn đề lãng phí không gian:* Phương pháp dựa trên mô hình thường có thể giảm thiểu vấn đề lãng phí không gian (cold start problem) hơn so với phương pháp dựa trên bộ nhớ. Với phương pháp dựa trên mô hình, nếu có đủ dữ liệu về các đặc trưng của người dùng và sản phẩm, có thể dễ dàng tạo ra các dự đoán cho các người dùng mới hoặc sản phẩm mới dựa trên các đặc trưng đó.
- *Hiệu suất cao hơn với dữ liệu thưa:* Phương pháp dựa trên mô hình thường hoạt động tốt hơn với dữ liệu thưa hơn, trong đó có thể bao gồm các người dùng chưa đánh giá nhiều sản phẩm hoặc các sản phẩm ít được đánh giá.
- *Tính tổng quát hóa tốt hơn:* Phương pháp dựa trên mô hình thường có khả năng tổng quát hóa tốt hơn đối với các loại dữ liệu mới hoặc biến đổi, vì chúng dựa vào các mô hình toàn diện hơn về ngữ cảnh và tính chất của dữ liệu.

Tuy nhiên, phương pháp lọc cộng tác dựa trên mô hình thường đòi hỏi thời gian và tài nguyên tính toán lớn hơn để huấn luyện các mô hình và tạo ra dự đoán, đặc biệt là khi áp dụng cho dữ liệu lớn.

1.2.3. Lọc kết hợp (Hybrid Filtering)

Các phương pháp khuyến nghị kết hợp *(Hybrid Filtering)* [8] tìm cách đạt được kết quả khuyến nghị tốt nhất bằng cách kết hợp các phương pháp khuyến nghị dựa trên nội dung và các phương pháp khuyến nghị dựa trên lọc cộng tác. Các hệ thống khuyến nghị kết hợp được chia thành khuyến nghị kết hợp nguyên khối, khuyến nghị kết hợp song song và khuyến nghị kết hợp đường ống [9]. Các hệ thống tư vấn kết hợp được sử dụng để tận dụng sức mạnh của nhiều nguồn dữ liệu hoặc để cải thiện hiệu suất của các hệ thống tư vấn hiện có trong một phương thức dữ liệu cụ thể. Động lực quan trọng cho việc xây dựng các hệ thống tư vấn lai là các loại hệ thống tư vấn khác nhau, chẳng hạn như các phương pháp dựa trên lọc cộng tác, dựa trên nội dung, có những điểm mạnh và điểm yếu khác nhau. Một số hệ thống gợi ý hoạt động hiệu quả hơn khi khởi động nguội (cold-start), trong khi các hệ thống khác hoạt động hiệu quả hơn khi có đủ dữ liệu. Các hệ thống tư vấn kết hợp cố gắng tận dụng sức mạnh bổ sung của các hệ thống này để tạo ra một hệ thống có tổng thể lớn hơn.



Hình 1. 9: Cơ chế lọc kết hợp

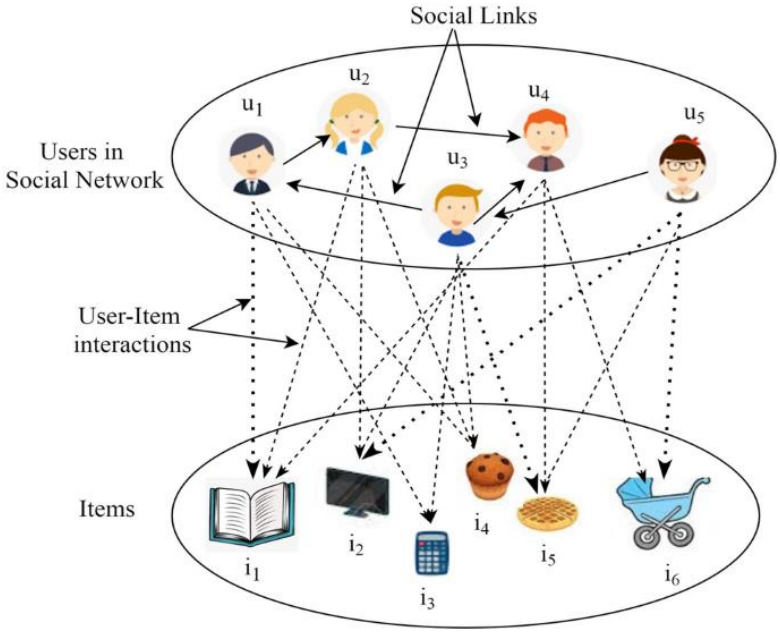
Một số phương pháp lọc kết hợp đã được nghiên cứu như sau:

- *Collaborative-Content Hybrid*: Phương pháp này kết hợp cả lọc dựa trên nội dung và lọc cộng tác, thông tin từ cả người dùng và sản phẩm được sử dụng để tạo ra các khuyến nghị. Phương pháp kết hợp này có thể khắc phục nhược điểm của từng phương pháp đơn lẻ và tận dụng các ưu điểm của cả hai để tạo nên tư vấn.
- *Model Combination*: Kết hợp các mô hình dự đoán khác nhau để tạo ra một mô hình tổng hợp. Ví dụ, chúng ta có thể tạo ra một mô hình bằng kết hợp một mô hình dự đoán dựa trên nội dung với một mô hình dự đoán dựa trên dữ liệu người dùng.
- *Feature Combination*: Kết hợp các đặc trưng (features) từ nhiều nguồn khác nhau để tạo ra một biểu diễn tổng hợp cho mỗi sản phẩm và người dùng. Điều này có thể bao gồm thông tin từ người dùng (ví dụ: hồ sơ người dùng, lịch sử tương tác) và từ sản phẩm (ví dụ: thông tin mô tả, từ khóa).
- *Cascade Hybrid*: Sử dụng các phương pháp khác nhau cho các giai đoạn khác nhau trong quá trình khuyến nghị. Ví dụ, có thể sử dụng collaborative filtering để tạo ra một danh sách ngắn các sản phẩm có thể được yêu thích, sau đó sử dụng content-based filtering để cải thiện các khuyến nghị trong danh sách này.
- *Meta-Level Hybrid*: Kết hợp các hệ thống khuyến nghị khác nhau tại một mức độ cao hơn, chẳng hạn bằng cách kết hợp các hệ thống khuyến nghị đã được huấn luyện trước.
- *Context-Aware Hybrid*: Kết hợp thông tin về ngữ cảnh (context) như vị trí địa lý, thời gian hoặc trạng thái người dùng để cải thiện các khuyến nghị.

1.2.4. Tư vấn xã hội (Social Recommendation)

Với sự bùng nổ phát triển của Internet, chúng ta có khả năng tiếp cận và khám phá toàn bộ thế giới thông qua những thiết bị thông minh như máy tính và điện thoại di động. Nhu cầu về kết nối và chia sẻ của con người đã điều chỉnh và phát triển theo chiều hướng này. Các ứng dụng web và mạng xã hội trở thành những sản phẩm không

thể thiếu, cùng với sự xuất hiện của các sàn thương mại điện tử trực tuyến, chúng là cầu nối giữa doanh nghiệp và người tiêu dùng. Khái niệm Web 2.0 (được giới thiệu bởi Dale Dougherty vào năm 2004) đưa ra những đặc trưng như sự hợp tác, chia sẻ, và sáng tạo giữa cộng đồng người dùng. Vì vậy rất để dàng để xây dựng các phần mềm cung cấp dưới dạng dịch vụ web, người dùng có thể sử dụng trên nhiều thiết bị khác nhau. Không chỉ vậy, phương tiện truyền thông xã hội cũng phát triển mạnh mẽ từ những xu hướng này.



Hình 1. 10: Tư vấn xã hội (Social Recommendation)

Từ khi xuất hiện, Web 2.0 đã tiếp tục phát triển tính chất sâu rộng trong cộng đồng, tạo điều kiện cho sự nổi lên của phương tiện truyền thông xã hội. Ví dụ, một người có thể tham gia vào một trang web chia sẻ hình ảnh, bài viết, video, âm nhạc, cũng như công trình nghiên cứu và ứng dụng khác. Những người khác có thể bình luận, tương tác, và phản hồi trên nền tảng này. Người dùng còn có thể là tác giả hoặc đóng góp tích cực vào nội dung trên phương tiện truyền thông xã hội, thường được tổ chức theo hệ thống phân cấp với các vai trò như quản trị viên, thành viên.

Năm 1997, lần đầu tiên chúng ta biết đến khái niệm hệ tư vấn xã hội (social recommender systems). Nghiên cứu về hệ tư vấn xã hội cũng ngày càng phát triển và thu hút nhiều sự quan tâm, đầu tư từ nhiều phía. Sự phát triển lớn mạnh của các loại

hình phương tiện xã hội làm cho những nghiên cứu về tư vấn xã hội luôn phải tích cực đổi mới và sáng tạo. Nhưng cho đến nay, một khái niệm về tư vấn xã hội mà tất cả cùng công nhận vẫn chưa được đề xuất. Tác giả Jiliang Tang và các cộng sự [10] của mình đã đưa ra hai định nghĩa hẹp và rộng về tư vấn xã hội. Nội dung cơ bản của 2 định nghĩa có thể được trình bày như sau:

- Định nghĩa hẹp của tư vấn xã hội: Một đặc trưng quan trọng của các phương tiện xã hội là các mối quan hệ giữa người dùng với nhau, có thể đó là bạn bè cùng tham gia một trang mạng xã hội, quan hệ thứ bậc quản trị viên – thành viên, hay sự tác động về mặt tâm lý đến nhau là sự tin tưởng giữa các người dùng với nhau. Và hệ tư vấn sẽ sử dụng các mối quan hệ này như một đầu vào bổ sung để cải thiện hiệu quả hệ tư vấn cũng như cải thiện chính các mối quan hệ.
- Định nghĩa rộng của tư vấn xã hội: Trong một phạm vi rộng hơn, tư vấn xã hội bao gồm mọi hoạt động liên quan đến các nền tảng phương tiện xã hội. Đối tượng của tư vấn này có thể là bất kỳ thực thể nào trên các nền tảng xã hội, bao gồm người dùng, bài đăng, thẻ, cộng đồng, và nhiều hơn nữa. Ngoài ra, dữ liệu còn bao gồm các hành vi và tương tác và của người sử dụng trên các nền tảng mạng xã hội.

Các bài toán tư vấn xã hội phổ biến:

- *Tư vấn nội dung*: Hệ thống tư vấn tập trung vào việc sử dụng các dữ liệu từ nhiều nguồn như phim, ảnh, âm nhạc, tin tức trên các nền tảng xã hội mà người dùng thường xuyên tương tác. Bằng cách phân tích hành vi trước đó và sở thích của người dùng, hệ thống có khả năng đề xuất những nội dung cụ thể mà họ có thể quan tâm, bao gồm sách cần đọc, phim cần xem, hoặc âm nhạc cần nghe. Việc này giúp cá nhân hóa trải nghiệm của người dùng và tăng cơ hội tìm thấy những nội dung mới mà họ sẽ thích.
- *Tư vấn địa điểm*: Hệ thống tư vấn này tận dụng thông tin về địa lý và các địa điểm du lịch để đề xuất những điểm đến thú vị cho người dùng khám phá. Dựa trên vị trí hiện tại của họ hoặc các sở thích trước đó, hệ thống có thể gợi ý

những địa điểm phù hợp như nhà hàng, công viên, bảo tàng, hoặc điểm du lịch địa phương. Điều này giúp người dùng khám phá những điểm đến mới một cách dễ dàng và thuận tiện, đồng thời tạo ra trải nghiệm du lịch đa dạng và phong phú.

- *Tư vấn bạn bè*: Hệ thống này tận dụng thông tin về mối quan hệ của người dùng trên các nền tảng mạng xã hội, như danh sách bạn bè, theo dõi, được theo dõi và mối quan hệ khác, để đề xuất các hoạt động và nội dung liên quan. Bằng cách phân tích mối quan hệ và hoạt động trước đó của bạn bè, hệ thống có thể gợi ý những hoạt động, sự kiện hoặc nội dung mà người dùng có thể quan tâm. Điều này tạo ra một cách tiếp cận cá nhân hóa và xã hội hóa trong việc khám phá và chia sẻ trải nghiệm giữa các mối quan hệ của người dùng.

1.3. Phương pháp học sâu trong Collaborative Filtering

Các phương pháp học sâu đã mang lại sự đột phá trong lĩnh vực lọc cộng tác mở ra những khả năng mới trong việc cải thiện độ chính xác của các hệ thống đề xuất. Thay vì chỉ tập trung vào sự tương tác giữa người dùng và sản phẩm, các phương pháp này tích hợp sức mạnh của mô hình học sâu để hiểu biểu diễn phức tạp của dữ liệu.

Các phương pháp như Matrix Factorization, Neural Collaborative Filtering (NCF), AutoEncoders, DeepFM và Attention Mechanisms đều đại diện cho sự kết hợp giữa lọc thông tin truyền thống và ưu điểm của mô hình học sâu. Điều này giúp cải thiện khả năng dự đoán và tạo ra trải nghiệm người dùng cá nhân hóa hơn.

Sự linh hoạt của mô hình học sâu, khi tích hợp vào CF, giúp chúng ta vượt qua những giới hạn của các phương pháp truyền thống, đặc biệt là khi xử lý các tương tác phi tuyến tính và mối quan hệ phức tạp giữa người dùng và sản phẩm. Việc này tăng cường khả năng đề xuất sản phẩm chính xác và tối ưu hóa trải nghiệm người dùng.

Tuy nhiên, sự thành công của các phương pháp này cũng phụ thuộc vào việc có đủ dữ liệu đa dạng để huấn luyện mô hình. Điều này đặt ra thách thức về quản lý và thu thập dữ liệu lớn, cũng như vấn đề về tính toán đối với các mô hình phức tạp. Tuy nhiên, với sự tiện lợi mà chúng mang lại trong việc cải thiện độ chính xác và đề

xuất cá nhân hóa, các phương pháp học sâu trong Collaborative Filtering đang trở thành xu hướng quan trọng trong phát triển các hệ thống lọc thông tin hiện đại.

1.4. Kết luận chương

Nội dung chương 1 đã trình bày làm rõ một số khái niệm cơ bản của hệ tư vấn, các phương pháp tiếp cận để xây dựng hệ tư vấn. Chương 1 cũng đưa ra chi tiết những ưu điểm cũng như nhược điểm của từng phương pháp tiếp cận. Bên cạnh đó, chương 1 đã trình khái quát về phương pháp học sâu trong lọc cộng tác làm cơ sở để lựa chọn phương pháp học sâu cho hệ tư vấn được trình bày ở chương 2.

CHƯƠNG 2. HỌC SÂU CHO HỆ TƯ VẤN LỌC CỘNG TÁC

Giới thiệu về tổng quan về học sâu, khái niệm của phương pháp phân rã ma trận và lựa chọn phương pháp học sâu cho hệ tư vấn.

2.1. Giới thiệu về học sâu

Học sâu (Deep Learning - DL) không chỉ là một phần của học máy, mà còn là một lĩnh vực quan trọng và đầy tiềm năng trong lĩnh vực trí tuệ nhân tạo (AI). Nó chủ yếu là một phần mở rộng của học máy, sử dụng các mô hình mạng nơ-ron sâu để học từ dữ liệu. Với ít nhất ba lớp hoặc nhiều hơn, các mô hình học sâu cố gắng mô phỏng cách mà não người xử lý thông tin, từ việc nhận diện hình ảnh đến dự đoán ngôn ngữ tự nhiên.

Trong quá trình học, các mạng nơ-ron này tiến hành điều chỉnh các trọng số và các tham số khác của mạng để tối ưu hóa hiệu suất của chúng trên dữ liệu huấn luyện. Cách tiếp cận này cho phép học sâu "học" từ các mẫu dữ liệu lớn và phức tạp mà không cần phải biết trước cụ thể về cách hoạt động của dữ liệu.

Một trong những điểm mạnh của học sâu là khả năng tự động hóa các tác vụ phân tích dữ liệu và các tác vụ thông minh mà trước đây cần sự can thiệp của con người. Điều này có thể làm tăng hiệu suất và tiết kiệm thời gian cho các ứng dụng trong nhiều lĩnh vực, từ y tế đến tài chính và hàng không vũ trụ.

Ứng dụng của học sâu là rất đa dạng. Chúng không chỉ xuất hiện trong các hệ thống trợ lý kỹ thuật số như Siri và Alexa, mà còn trong các công nghệ tiên tiến như xe tự lái và robot hộp. Học sâu cũng đã có ảnh hưởng lớn đến nền kinh tế, với việc cải thiện tự động hóa trong quy trình sản xuất và dịch vụ, cũng như trong việc phát hiện gian lận và dự báo xu hướng thị trường.

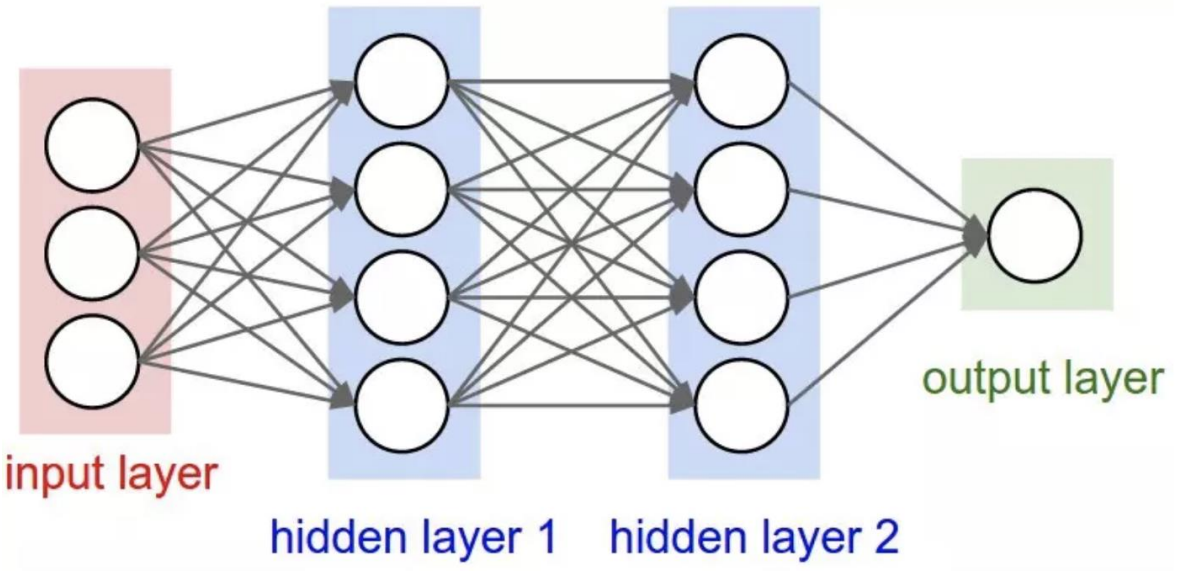
Học sâu cũng đóng một vai trò rất quan trọng trong các hệ thống tư vấn/khuyến nghị, nó giúp người dùng có thể có những trải nghiệm tốt nhất. Học sâu không chỉ là một công nghệ mạnh mẽ, mà còn là một lĩnh vực nghiên cứu đầy triển vọng, với tiềm năng để tạo ra những tiến bộ đáng kể trong nhiều lĩnh vực khác nhau trong tương lai.

2.1.1. Cách thức hoạt động của học sâu

Học sâu hoạt động bằng cách khám phá và tìm hiểu các cấu trúc phức tạp trong dữ liệu. Để thực hiện điều này, nó sử dụng các mô hình tính toán sâu có nhiều lớp xử lý thông tin. Nhờ vào sự kết hợp của các mạng nơ-ron (neural) sâu, học sâu có khả năng tạo ra nhiều mức độ trừu tượng khác nhau để biểu diễn và hiểu dữ liệu. Cũng giống động vật, bộ não của trí tuệ nhân tạo (AI) cũng có các nơ ron. Chúng được biểu diễn bằng các vòng tròn. Các nơ ron này đều đã được liên kết.

Các nơ ron được nhóm vào 3 loại layer khác nhau: Input layer, các hidden layer, Output layer

- *Input layer:* Nhận các dữ liệu đầu vào
- *Các hidden layer*: Thực hiện các phép tính toán trên các đầu vào. Một trong những thách thức lớn khi thiết kế mạng nơ-ron là quyết định về số lượng hidden layer và số nơ-ron trong mỗi layer.
- *Output layer*: trả về dữ liệu đầu ra



Hình 2. 1: Các layer của mạng nơ-ron

Ví dụ, trong hệ tư vấn, một mô hình học sâu có thể được sử dụng để đề xuất sản phẩm cho người dùng dựa trên lịch sử mua hàng của họ và thông tin về sở thích cá nhân. Mô hình này có thể được huấn luyện bằng cách sử dụng một lượng dữ liệu rất lớn, chẳng hạn như hàng triệu giao dịch mua hàng trước đó. Bằng cách học từ dữ

liệu, mô hình có thể tự động nhận diện các mẫu mua hàng, ưu tiên các sản phẩm mà người dùng có khả năng quan tâm và đề xuất chúng một cách hiệu quả. Trong nhiều trường hợp, các hệ thống tư vấn sử dụng học sâu có thể cung cấp đề xuất chính xác và cá nhân hóa hơn so với các phương pháp truyền thống của học máy.

Tuy nhiên, không có nghĩa là việc xây dựng các hệ thống học sâu dễ dàng hơn so với các hệ thống học máy thông thường. Điều chỉnh hàng nghìn siêu tham số là cần thiết để tối ưu hóa hiệu suất của mô hình học sâu.

2.1.2. Ưu điểm của học sâu so với phương pháp học máy cổ điển

Học sâu mang lại nhiều ưu điểm so với phương pháp học máy cổ điển, và điều này đã đóng vai trò quan trọng trong việc biến đổi cách chúng ta tiếp cận và giải quyết các bài toán máy học.

Một trong những ưu điểm đáng chú ý nhất của học sâu là khả năng học các đặc trưng phức tạp từ dữ liệu một cách tự động. Điều này làm cho việc xử lý các bài toán có tính phức tạp cao trở nên hiệu quả hơn và ít phụ thuộc vào sự hiểu biết chuyên môn của con người.

Không chỉ vậy, học sâu còn thể hiện hiệu suất tốt hơn trên dữ liệu lớn. Nhờ vào cấu trúc mạng nơ-ron sâu và khả năng tính toán song song trên các thiết bị có đồng nhất, học sâu có thể xử lý và học được từ lượng dữ liệu lớn một cách hiệu quả. Điều này đặc biệt quan trọng trong thời đại số hóa ngày nay, khi lượng dữ liệu được tạo ra và tích lũy ngày càng tăng lên.

Thêm vào đó, học sâu còn mang lại sự linh hoạt và đa dạng. Mạng nơ-ron sâu không chỉ học được từ nhiều loại dữ liệu khác nhau như hình ảnh, âm thanh, văn bản, mà còn có thể áp dụng cho nhiều lĩnh vực khác nhau như thị giác máy tính, xử lý ngôn ngữ tự nhiên, và dự đoán chuỗi thời gian. Điều này tạo ra một cơ sở cho ứng dụng rộng rãi của học sâu trong nhiều ngành công nghiệp và lĩnh vực nghiên cứu.

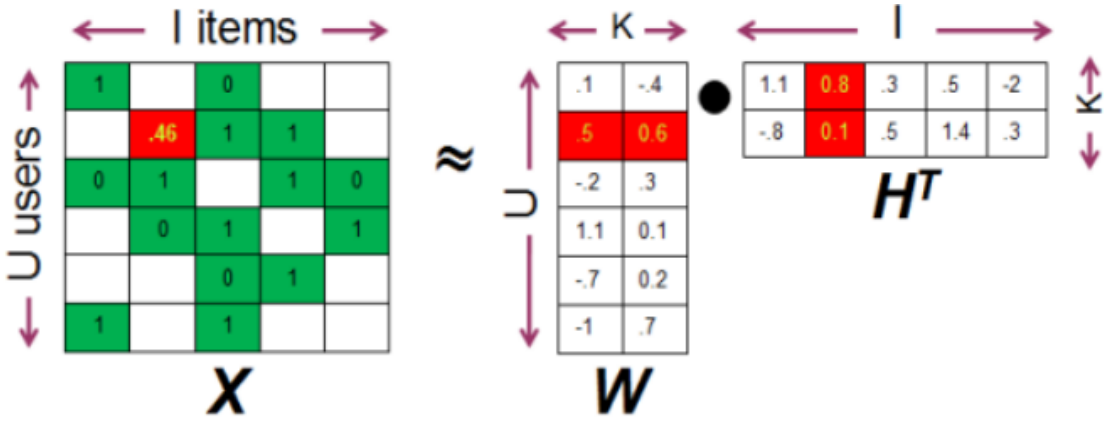
Ngoài ra, học sâu còn tích hợp phân tích cấp cao, cho phép nó hiểu được các mối quan hệ phức tạp giữa các yếu tố trong dữ liệu. Điều này làm cho học sâu trở thành một công cụ mạnh mẽ cho nhiều bài toán phức.

Tuy nhiên, cũng cần lưu ý rằng việc huấn luyện mô hình học sâu thường đòi hỏi nhiều dữ liệu hơn và tài nguyên tính toán lớn hơn so với các phương pháp học máy truyền thống. Đồng thời, việc điều chỉnh các siêu tham số của mạng nơ-ron cũng có thể là một thách thức đối với học sâu. Tuy nhiên, với sự phát triển của công nghệ và sự gia tăng về khả năng tính toán, học sâu vẫn tiếp tục phát triển và trở thành một trong những công cụ quan trọng nhất trong lĩnh vực trí tuệ nhân tạo và máy học.

2.2. Phương pháp phân rã ma trận

Phân rã ma trận (Matrix Factorization - MF) [12] là một trong những thuật toán lâu đời sử dụng trong các hệ thống gợi ý. Mô hình đầu tiên được đề xuất bởi Simon Funk và phát triển mạnh mẽ, phổ biến sau cuộc thi Netflix tổ chức năm 2006.

Kỹ thuật phân rã ma trận được thực hiện thông qua việc chia một ma trận lớn X thành 2 ma trận có kích thước nhỏ hơn rất nhiều so với ma trận ban đầu là W và H, sao cho X có thể được xây dựng lại từ hai ma trận nhỏ hơn này càng chính xác càng tốt [11], điều này có nghĩa là $\mathbf{X} \approx \mathbf{W}\mathbf{H}^{\mathbf{T}}$ như minh hoạ trong hình 2.3. Trong đó, X là tập hợp tất cả các đánh giá (rating) của người dùng với mục tin, bao gồm cả những giá trị chưa biết cần được dự đoán tạo nên một ma trận gọi là Utility Matrix.



Hình 2. 2: Minh họa kỹ thuật phân rã ma trận

Trong đó:

- K là số nhân tố tiềm ẩn, và nhỏ hơn rất nhiều so với số người dùng và số mục dữ liệu (K << $|U|$ và K << $|I|$).

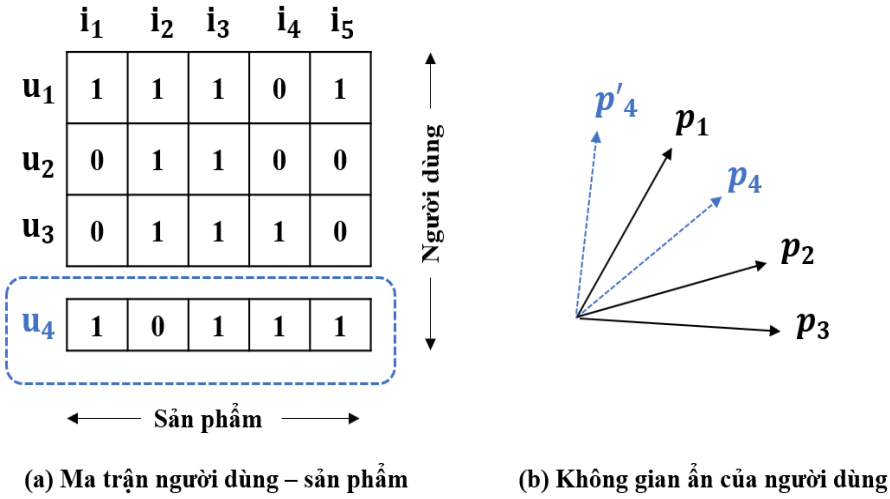
- W ∈ $R^{|U|\times K}$ là một ma trận mà mỗi dòng là một vector bao gồm K nhân tố tiềm ẩn (latent factors) mô tả user u.
- W ∈ $R^{|I|\times K}$ là một ma trận mà mỗi dòng là một vectors bao gồm K nhân tố tiềm ẩn mô tả cho item i.

Ý tưởng chính của MF là tồn tại các tính chất ẩn (latent features) mô tả sự liên quan giữa các người dùng và sản phẩm. Mỗi người dùng sẽ mang chất ẩn nào đó và được mô tả bởi các hệ số trong vector w_u. Tương tự, mỗi sản phẩm cũng sẽ có những tính chất ẩn ở một mức độ nào đó tương ứng với các hệ số trong vector h_i. Hệ số càng cao tương ứng với việc mang tính chất đó càng cao.

Gọi h_{ik}và w_{uk}là các phần tử tương ứng của hai ma trận H và W, khi đó xếp hạng (rating) của người dùng u trên mục dữ liệu i được dự đoán bởi công thức (4):

$$\hat{r}_{ui} = \sum_{k=1}^{K} w_{uk} h_{ik} = (WH^T)_{u,i} \quad (4)$$

Phương pháp MF mô hình hóa tương tác hai chiều giữa các thuộc tính ẩn của người dùng và sản phẩm. MF giả sử các chiều trong không gian ẩn đó là độc lập và tính ra giá trị tương tác bằng cách tổ hợp các chiều đó lại với cùng một trọng số. Như vậy, MF có thể coi là một mô hình tuyến tính của các thuộc tính ẩn.



(a) Ma trận người dùng – sản phẩm (b) Không gian ẩn của người dùng

Hình 2. 3: Ví dụ chứng minh hàm tích vô hướng có thể giới hạn chất lượng của MF

Hình 2.4 chứng minh rằng hàm tích vô hướng (inner product) có thể giới hạn chất lượng của MF. Do MF chiếu người dùng và sản phẩm lên cùng một không gian ẩn, độ tương đồng giữa 2 người dùng có thể được đánh giá bằng cách sử dụng phép nhân vô hướng hoặc có thể sử dụng độ đo Cosine giữa 2 vector ẩn được tạo. Để không làm mất tính tổng quát, ta cũng có thể sử dụng độ đo Jaccard để đo độ tương tự giữa 2 người dùng trong dữ liệu ban đầu mà sau đó MF cần có khả năng nắm bắt được.

Độ đo Jaccard: Đặt R_u là tập các items mà user **u** có tương tác, khi đó độ đo tương đồng **Jaccard** của hai user **i** và **j** được định nghĩa là:

$$s_{i,j} = \frac{|R_i| \cap |R_j|}{|R_i| \cup |R_j|} \quad (5)$$

Từ số liệu của ba dòng đầu tiên của hình 2.4 a, ta sẽ tính được $s_{2,3}(0.66) > s_{1,2}(0.5) > s_{1,3}(0.4)$. Như vậy, quan hệ trên không gian của p_1, p_2, p_3 có thể được vẽ như trong hình 2.4 b.

Xét user u_4 có $s_{4,1}(0.6) > s_{4,3}(0.4) > s_{4,2}(0.2)$ điều này có nghĩa là user u_4 tương tự nhất với u_1, tiếp theo là u_3 và u_2 . Tuy nhiên, do mô hình MF đưa vector người dùng lên cùng một không gian ẩn, nên với MF, có 2 cách đặt vector của user u_4 thỏa mãn gần u_1 nhất như hình vẽ trên (2 đường p'_4và p_4). Nhưng cả hai trường hợp này đều không thể thỏa mãn được tính chất thực tế, đó là u_4 gần u_3 hơn u_2. Từ đó, có thể thấy rằng MF không thể mô tả được độ đo Jaccard hay chính là độ đo sự tương tự giữa các người dùng trong trường hợp này.

2.3. Phương pháp học sâu cho hệ tư vấn

Do phương pháp phân rã ma trận còn có nhiều hạn chế như đã nêu ở mục 2.2 nên việc phải thiết kế một hàm tương tác tốt hơn để mô hình hóa việc tương tác giữa các thuộc tính ẩn của người dùng và sản phẩm là vô cùng cần thiết. Biểu diễn các người dùng và sản phẩm dựa vào tương tác sẽ làm giàu các thông tin cho người dùng và sản phẩm đó. Phép toán tích vô hướng (inner product) đơn giản chỉ là việc kết hợp bằng cách nhân các thuộc tính ẩn tuyến tính, như vậy có thể không đủ để nắm bắt cấu trúc dữ liệu tương tác phức tạp của người dùng. Vì vậy, đề án đề xuất phương pháp

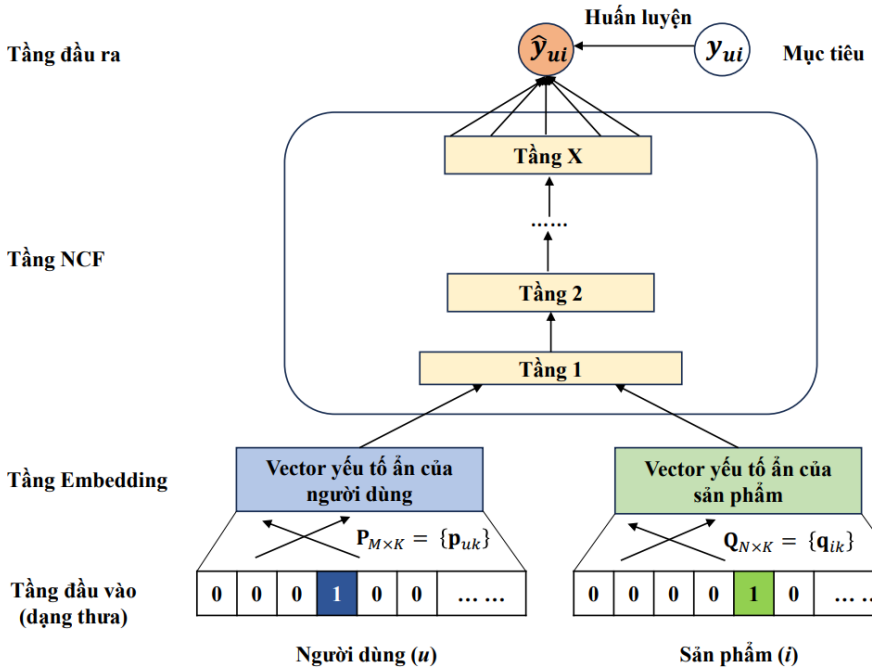
học sâu Neural Collaborative Filtering (NCF) cho hệ tư vấn. Mô hình NCF đi khai thác phương pháp MF theo hướng này và thực hiện nó bằng mạng nơ-ron. Thay vì việc chỉ sử dụng thứ tự của người dùng và sản phẩm chỉ có tác dụng phân biệt giữa người dùng này với người dùng khác, sản phẩm này với sản phẩm khác, việc biểu diễn thông qua tương tác sẽ có thể biểu diễn được sự tương đồng giữa người dùng với người dùng; sản phẩm với sản phẩm dựa vào các tương tác chung của các đối tượng này.

2.3.1. Phương pháp Neural Collaborative Filtering

Phương pháp Neural Collaborative Filtering (NCF) [13] là một kỹ thuật đặc biệt trong lĩnh vực học máy và khai phá dữ liệu, được sử dụng rộng rãi trong các ứng dụng đề xuất cá nhân và lọc cộng tác. NCF kết hợp cả hai tiếp cận truyền thống trong việc xây dựng hệ thống đề xuất: Collaborative Filtering (CF) và Neural Networks (NN), từ đó tạo ra một phương pháp mạnh mẽ và linh hoạt. Ở bản chất, CF tập trung vào việc dự đoán sở thích của một người dùng dựa trên lịch sử hoặc hành vi của người dùng đó cũng như hành vi của những người dùng khác. Trong khi đó, NN có khả năng học các biểu diễn phức tạp từ dữ liệu đầu vào và thực hiện các dự đoán dựa trên các biểu diễn đó.

NCF kết hợp cả hai mô hình này bằng cách sử dụng một mạng nơ-ron để học các biểu diễn người dùng và mặt hàng từ dữ liệu, sau đó sử dụng các biểu diễn này để dự đoán sở thích hoặc xếp hạng của người dùng cho các mặt hàng mà họ chưa xem hoặc đánh giá.

Mô hình NCF được nhóm tác giả Xiangnan He và cộng sự [13] đưa ra vào năm 2017. Nhóm tác giả đã đưa ra đề xuất sử dụng cấu trúc NCF tổng quát được thể hiện như trong hình vẽ dưới đây:



Hình 2. 4: Kiến trúc mô hình NCF

Trong hình vẽ trên, đi từ dưới lên, mô hình NCF bao gồm các tầng: (1) Tầng đầu vào, (2) Tầng Embedding, (3) Tầng NCF, (4) Tầng đầu ra. Sau đây, đề án sẽ trình bày nội dung chi tiết nội dung chi tiết của từng tầng trong mô hình NCF.

2.3.1.1. Tầng đầu vào

Tầng đầu vào bao gồm các thuộc tính vector của *người dùng* và *sản phẩm*. Như trong mô hình trên, các vector đó là biểu diễn dạng one-hot là vector với chỉ một trường có giá trị bằng 1, các trường còn lại có giá trị 0 dùng để biểu diễn định danh của *người dùng* (userID) và *sản phẩm* (itemID). Đây là biểu diễn thường thấy của các phương pháp cổ điển.

Giả sử v_k là vector biểu diễu dạng one-hot của user u, v_k được xác định bởi công thức:

$$v_k = \begin{cases} 1 & \text{nếu k bằng id của user u} \\ 0 & \text{ngược lại} \end{cases} \quad (6)$$

Sau khi thực hiện one-hot encoding, các vector này sẽ được kết hợp lại thành một vector lớn hơn để tạo thành đầu vào cho mô hình NCF, có thể được đưa vào mạng nơ-ron để thực hiện các phép tính và dự đoán. Điều này giúp mô hình NCF có

thể học được mối quan hệ giữa người dùng và sản phẩm dựa trên các tương tác giữa chúng.

2.3.1.2. Tầng Embedding

Dữ liệu phân loại (categorical data) là các loại dữ liệu mà các thuộc tính của chúng nhận giá trị từ một tập hữu hạn các danh mục khác nhau, thay vì nhận các giá trị liên tục. Ví dụ, trong tập dữ liệu về người dùng và bộ phim, các thuộc tính như thể loại phim hoặc đánh giá có thể được coi là dữ liệu phân loại. Tương tự, trong xử lý ngôn ngữ tự nhiên, từ vựng trong một tài liệu cũng là dữ liệu phân loại. Dữ liệu này sẽ được biểu diễn dưới dạng tensor một chiều hoặc vector $\vec{V}$ có các phần tử nhận giá trị như sau:

$$v_j = \begin{cases} 1 & \text{nếu nhận giá trị} \\ 0 & \text{nếu không nhận giá trị} \end{cases} \quad (7)$$

Thuật toán mã hóa được sử dụng ở đây được gọi là one-hot encoding, trong đó đầu ra của thuật toán là vector $\vec{V}$, một vector thưa có ít phần tử khác 0. Trong ngữ cảnh của hệ thống tư vấn, ma trận Người dùng – Sản phẩm (User – Item) có thể được hiểu như một tập hợp các tensor người dùng, trong đó các phần tử tương ứng với các sản phẩm mà người dùng đã đánh giá.

Trong *hình 2.6* mỗi hàng của ma trận ghi nhận các bộ phim mà người dùng tương ứng đã xem, mỗi người dùng tương ứng sẽ được biểu diễn bởi một tensor thưa. Nguyên nhân xuất phát từ việc người dùng thường chỉ xem một phần nhỏ các bộ phim trong toàn bộ tập hợp các bộ phim.

	MovieId				
	1	**2**	**3**	**...**	**10**
	X				
		X			
			X		
Người dùng u		X			X

Hình 2. 5: Vector thưa đại diện cho người dùng u

Ví dụ, người dùng u hiện tại trong hình 2.6 được đại diện bởi tensor [2, 10] trong đó giá trị 2 và 10 là các mã số (Movie Id) của các bộ phim đã xem.

Tuy nhiên, việc trực tiếp sử dụng biểu diễn dạng vector thưa này để huấn luyện mạng nơ ron có thể gây ra tình trạng bùng nổ các trọng số cần huấn luyện cho mạng neuron, với những hệ quả sau:

- Số lượng dữ liệu cần thiết để huấn luyện tăng lên: Mô hình sử dụng nhiều trọng số đòi hỏi một lượng dữ liệu lớn hơn để đạt được hiệu suất huấn luyện mong muốn.
- Khối lượng tính toán tăng lên: Với số lượng trọng số lớn hơn, khối lượng tính toán cần thiết để thực hiện các phép tính tăng lên đáng kể.

Có thể giải quyết những vấn đề nêu trên bằng cách sử dụng các thuật toán nhúng (embedding), trong đó vector thưa có số chiều lớn được ánh xạ vào một không gian vector có số chiều thấp hơn dày đặc hơn. Quá trình nhúng này giúp tạo ra các vector đặc tính ẩn (latent vector) mô tả về đối tượng (sản phẩm hoặc người dùng) một cách dày đặc và bảo toàn các mối quan hệ ngữ nghĩa tồn tại trong dữ liệu ban đầu.

Lớp embedding nhận 03 tham số đầu vào quan trọng như sau:

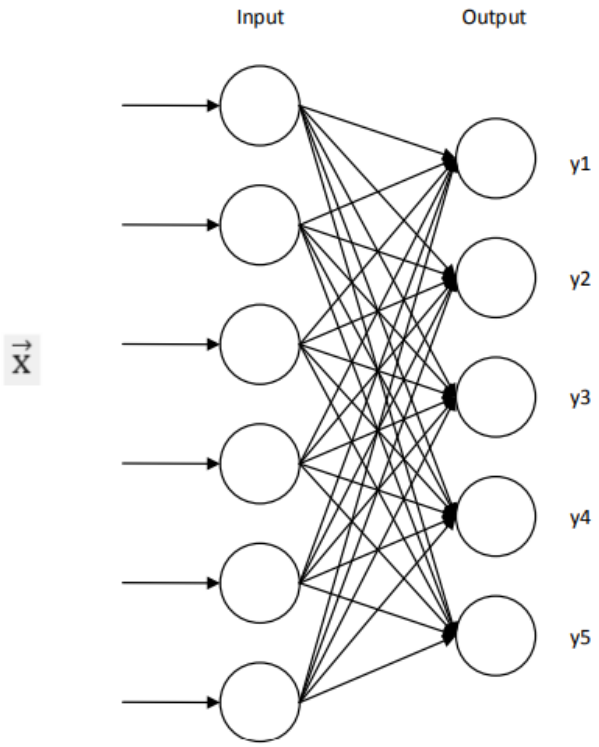
- data: mảng số nguyên chứa Id của đối tượng được embedding.
- input_dim: độ lớn của mảng data.
- output_dim: kích thước của vector đặc tính ẩn.

Các đặc tính ẩn thu được từ lớp embedding có thể là những đặc điểm về nội dung của bộ phim mà người dùng ưa thích, như thể loại phim, năm sản xuất, mức độ lời thoại, và nhiều yếu tố khác. Tuy nhiên, do đặc tính của lớp mô hình đặc tính ẩn, ngữ nghĩa của các vector đặc tính này chỉ có thể được giải thích một cách không hoàn toàn đầy đủ dựa trên tri thức miền. Kích thước của vector thuộc tính ẩn có thể được thay đổi trong quá trình thử nghiệm, để đạt được hiệu quả một cách tối đa thì việc lựa chọn kích thước phải hợp lý, nếu kích thước của vector thuộc tính ẩn quá nhỏ thì mạng nơ ron sẽ có ít khả năng học được các đặc tính của dữ liệu gốc. Ngược lại, nếu kích thước quá lớn, mạng neuron có nguy cơ bị quá khớp (overfitting).

2.3.1.3. Tầng Neural Collaborative Filtering

Các tầng thuộc mạng nơ-ron (tầng 1 đến N) - Tầng neural collaborative: các tầng này có nhiệm vụ ánh xạ từ các vector embedding đến vector đầu vào của tầng đầu ra (output). Mỗi tầng là một lớp ẩn.

Lớp ẩn trong mạng nơ-ron được sử dụng để mô hình hóa các quan hệ phi tuyến tính có thể tồn tại trong không gian vector đặc tính ẩn nằm ở đầu ra của lớp nối



Hình 2. 6: Hoạt động của lớp ẩn

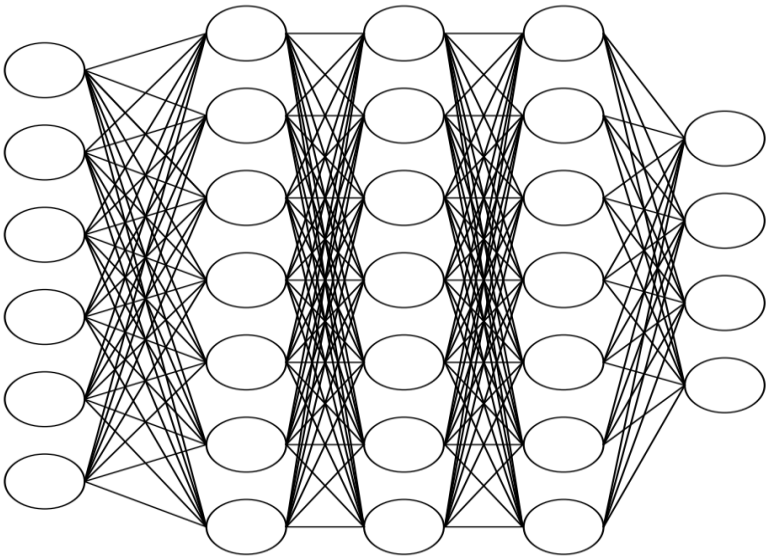
Cho $x \in \mathbb{R}^m$, $y_i \in \mathbb{R}$

$$y_i = \sigma(w_1 x_1 + \ldots + w_m x_m) \quad (8)$$

Vector y đầy đủ như sau:

$$y = \begin{pmatrix} \sigma(w_{1,1}x_1 + \ldots + w_{1,m}x_m) \\ \vdots \\ \sigma(w_{n,1}x_1 + \ldots + w_{n,m}x_m) \end{pmatrix} \quad (9)$$

Khi xếp chồng nhiều lớp mạng ẩn lên nhau, ta thu được mạng Multi-Level Perceptrons MLP như hình 2.8:



Hình 2. 7: Kiến trúc MLP

Đầu ra tại mỗi lớp bằng tích của vector đầu vào với ma trận trọng số và mỗi phần tử của vector đầu ra được áp dụng toán tử phi tuyến σ:

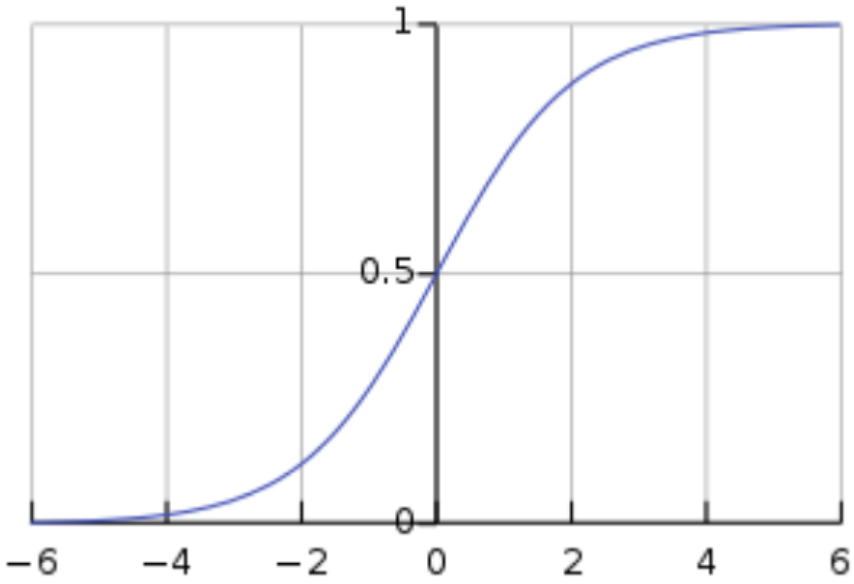
$$y = \sigma(wx) \qquad (10)$$

Trong đó: σ là ma trận không gian $R^{m \times n}$ có các phần tử được áp dụng toán tử phi tuyến σ.

Toán tử phi tuyến σ của các lớp là hàm sigmoidal được định nghĩa như sau:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (11)$$

Hàm sigmoid được sử dụng để chuyển đổi đầu vào từ một số thực thành một giá trị nằm trong khoảng (0, 1). Đối với đầu vào là số thực dương với giá trị rất lớn sẽ cho đầu ra tiệm cận 1, đầu vào là số thực âm với giá trị rất nhỏ sẽ cho đầu ra tiệm cận 0.



Hình 2. 8: Hàm Sigmoid

Hàm gặp phải một số hạn chế như sau:

- Hàm sigmoid là hàm bão hòa: Khi đầu vào của hàm sigmoid là một giá trị rất lớn hoặc rất nhỏ, đạo hàm của nó gần tiến về 0. Khi lan truyền ngược để cập nhật trọng số, việc này có thể dẫn đến việc gradient tiến về 0, làm chậm quá trình học của mạng neural (gây ra hiện tượng vanishing gradient).
- Không trung tâm ở 0 (Not zero-centered): Điều này có thể dẫn đến việc khó khăn trong việc huấn luyện mô hình.

Một trong những đặc điểm nổi bật của các mạng ẩn là khả năng học và ghi nhớ thông tin từ dữ liệu huấn luyện theo thời gian. Điều này thường được thể hiện qua việc giảm dần sai số trên tập huấn luyện sau mỗi lượt huấn luyện (epoch). Tuy nhiên, điều này không đảm bảo rằng mô hình sẽ hoạt động tốt trên dữ liệu kiểm thử. Thậm chí, có thể xảy ra tình trạng tăng sai số trên dữ liệu kiểm thử trong khi sai số trên dữ liệu huấn luyện giảm (gây ra hiện tượng quá khớp - overfitting). Vì vậy, việc mạng neural "hội tụ" trên tập dữ liệu huấn luyện không nhất thiết đồng nghĩa với việc nó hoạt động hiệu quả trên các dữ liệu thực tế.

Sau khi đi qua tất cả các lớp ẩn (Fully-connected layers), vector ẩn đầu ra tại lớp cuối cùng được tính toán theo công thức truy hồi như sau:

$$z_1 = \Phi_1(x),$$
$$\Phi_2(x_1) = \sigma_2(W_2^T x_1 + b_2), \qquad (12)$$
$$\dots$$
$$\Phi_L(x_{L-1}) = \sigma_L(W_L^T x_{L-1} + b_L),$$

Trong đó:

W_x : Ma trận trọng số tương ứng với lớp thứ x.

b_x: Vector hệ số thiên lệch (bias) tương ứng lớp thứ x.

σ: Hàm kích hoạt của perceptron thứ x.

Đầu ra của mạng MLP là một vector đặc tính ẩn mà trong đó mỗi phần tử mô tả một thuộc tính nào đó của người dùng, sản phẩm, thông tin bổ trợ.

2.3.1.4. Tầng đầu ra

Ở tầng này, điểm số dự đoán $\hat{y}_{ui}$ được tính toán và sử dụng lan truyền ngược để cập nhật các tham số sao cho giá trị $\hat{y}_{ui}$ đạt xấp xỉ y_{ui}

Việc huấn luyện NCF được thực hiện bằng cách tối thiểu hóa hàm lỗi giữa điểm số dự đoán với giá trị mục tiêu (target) y_{ui} tương ứng.

Điểm số dự đoán được tính như sau:

$$\hat{y}_{ui} = f\left(\mathbf{P}^{\mathrm{T}}\mathbf{v}_{\mathrm{u}}^{\mathrm{U}}, \mathbf{Q}^{\mathrm{T}}\mathbf{v}_{\mathrm{i}}^{\mathrm{I}} | \mathbf{P}, \mathbf{Q}, \boldsymbol{\theta}_{\mathrm{f}}\right) \qquad (13)$$

Trong đó:

- v_u^U, v_i^U lần lượt là các ma trận hàng (vector) one-hot biểu diễn cho user u và item i.
- $P \in R^{M \times K}$ và $Q \in R^{N \times K}$ lần lượt là các ma trận tầng embedding chiếu v_u^U, v_i^U vào không gian các thuộc tính ẩn. **P** và **Q** là ma trận trọng số giữa tầng input và tầng embedding.
- Θ_f là tập các tham số mô hình của hàm f. Do hàm f ở đây là ánh xạ đầu vào đầu ra của một mạng nơ-ron nên ta có thể viết f theo công thức sau:

$$f\left(\mathbf{P}^{\mathrm{T}}\mathbf{v}_{\mathrm{u}}^{\mathrm{U}}, \mathbf{Q}^{\mathrm{T}}\mathbf{v}_{\mathrm{i}}^{\mathrm{I}}\right) = \emptyset_{out}\left(\emptyset_X(\dots\emptyset_2(\emptyset_1(\boldsymbol{P}^T\boldsymbol{v}_u^U, \boldsymbol{Q}^T\boldsymbol{v}_i^I))\dots)\right) \qquad (14)$$

Trong đó:

- ϕ_{out} là hàm ánh xạ đầu vào tới đầu ra của tầng output.

- $\phi_n, \phi_{n-1}, \dots, \phi_2, \phi_1$ lần lượt là các hàm ánh xạ đầu vào tới đầu ra tại các tầng n, n-1, . . ., 2, 1 trong mạng nơ-ron.

2.4. Kết luận chương

Nội dung chương 2 đã trình bày khái niệm của phương pháp phân ra ma trận, ưu và nhược điểm của phương pháp này, lựa chọn phương pháp học sâu cho hệ tư vấn là phương pháp Neural Collaborative Filtering và trình bày kiến trúc mô hình của phương pháp này. Ở chương 3 tiếp theo, đề án sẽ trình bày về tập dữ liệu và tiến hành thực nghiệm và đánh giá kết quả phương pháp học sâu Neural Collaborative Filtering (NCF) đã lựa chọn.

CHƯƠNG 3: THỰC NGHIỆM VÀ KẾT QUẢ

Chương 3 trình bày về quá trình thực nghiệm, bao gồm về tập dữ liệu thực nghiệm, độ đo dùng để đánh giá mô hình, so sánh và đánh giá kết quả thực nghiệm trên mô hình đề xuất.

3.1. Tập dữ liệu thực nghiệm

3.1.1. Tập dữ liệu MovieLens-1M

Movielens [15] là một nhóm nghiên cứu cung cấp các bộ dữ liệu cho các bài toán xây dựng hệ thống gợi ý. Các bộ dữ liệu trong tập này bao gồm thông tin đánh giá xếp hạng (rating) của người dùng tới các bộ phim. Xếp hạng nhận giá trị từ 0.5 đến 5.

Tính đến thời điểm hiện tại, dự án Movielens cung cấp tổng cộng 06 bộ dữ liệu mẫu cụ thể như sau:

- Bộ dữ liệu ml-100k là bộ dữ liệu gồm 100.000 xếp hạng phim từ 943 người dùng cho 1.682 bộ phim được phát hành vào tháng 04/1998.
- Bộ dữ liệu ml-1M là bộ dữ liệu gồm 1.000.209 xếp hạng từ 6.040 người dùng cho 3.900 bộ phim được phát hành vào tháng 02/2003.
- Bộ dữ liệu ml-10M là bộ dữ liệu gồm 10.000.054 xếp hạng từ 71.567 người dùng cho 10.681 bộ phim được phát hành vào tháng 02/2003.
- Bộ dữ liệu ml-20M là bộ dữ liệu gồm 20.000.263 xếp hạng từ 138.493 người dùng cho 27.278 bộ phim được phát hành vào tháng 10/2016.
- Bộ dữ liệu ml-25M là bộ dữ liệu gồm 25.000.095 xếp hạng từ 162.541 người dùng cho 62.423 bộ phim được phát hành vào tháng 10/2019.
- Bộ dữ liệu MovieLens 1B Synthetic Dataset: Đây là bộ dữ liệu sử dụng tập dữ liệu ml-20M để sinh ra dữ liệu nhân tạo xấp xỉ gồm 1.223.962.043 xếp hạng cho dữ liệu huấn luyện và 12.709.557 xếp hạng cho dữ liệu kiểm tra.

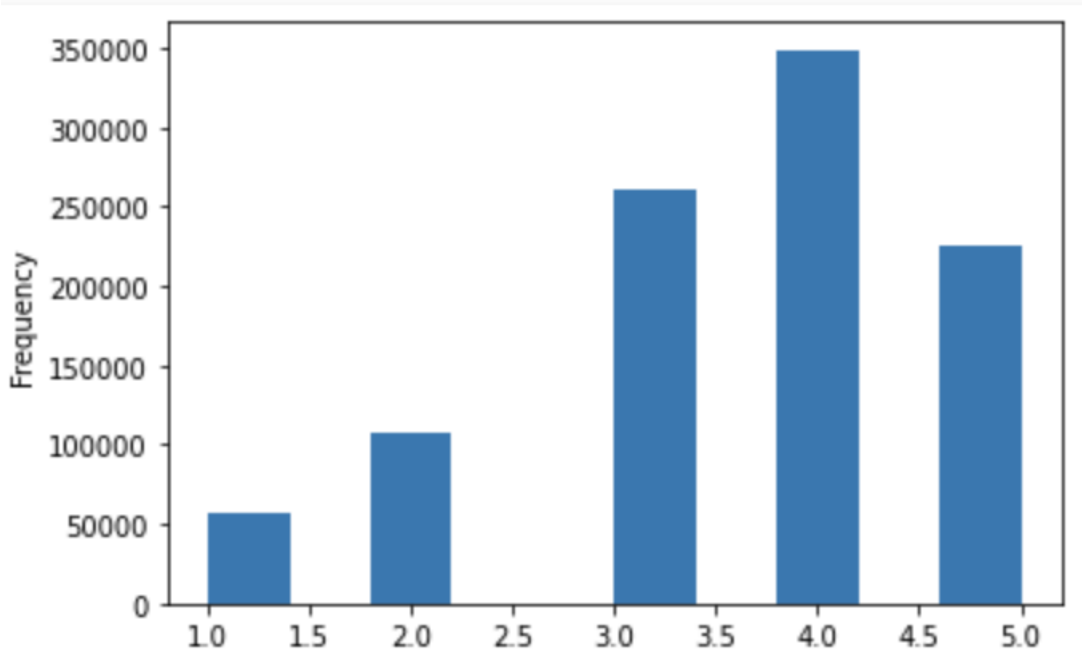
Bên cạnh dữ liệu về xếp hạng phim, dự án Movielens còn cung cấp một số dữ liệu bổ trợ như thông tin nhân khẩu học của người dùng, thông tin mô tả phim như thể loại phim, các thẻ (tags) được gán cho từng bộ phim.

Bộ dữ liệu MovieLens-1M bao gồm xấp xỉ 1 triệu bộ (user, movie, rating) từ khoảng 3900 bộ phim và 6040 người dùng. Số lượng bộ phim mà người dùng đánh giá nhiều nhất là 2314 bộ phim và 20 là số lượng bộ phim ít nhất mà mỗi người dùng đưa ra.

	UserID	MovieID	Rating
0	1	1193	5
1	1	661	3
2	1	914	3
3	1	3408	4
4	1	2355	5
5	1	1197	3
6	1	1287	5
7	1	2804	5
8	1	594	4
9	1	919	4

Hình 3. 1: Ví dụ 10 dòng dữ liệu đầu tiên của dataframe rating bộ dữ liệu MovieLens-1M

Phân phối của các điểm đánh giá (rating) ở biểu đồ hình 3.2 cho thấy rằng rằng điểm 4 được đánh giá nhiều nhất trong khi các điểm 1 và 2 có ít lượng đánh giá nhất. Điều này có thể được giải thích bằng sự thật là người dùng thường đánh giá khi họ rất thích một bộ phim; khi họ không thực sự thích, họ sẽ ít ra đánh giá hơn.



Hình 3. 2: Phân phối điểm đánh giá (rating) bộ dữ liệu MovieLens-1M

3.1.2. Tập dữ liệu Dlab

Dlab [1, 17] là cổng lập trình trực tuyến của Khoa Công nghệ thông tin 1 - Học viện Công nghệ Bưu chính Viễn thông. Dữ liệu Dlab được thu thập trực tiếp từ cổng lập trình. Tập dữ liệu thu thập từ 6120 người dùng. Trong đó, người dùng i lập trình đúng bài toán x cổng lập trình ghi lại giá trị 1, người dùng i lập trình chưa đúng bài toán x cổng lập trình ghi lại giá trị -1, người dùng i chưa giải bài toán x cổng lập trình ghi lại giá trị 0. Mỗi người lập trình có thể submit một bài nhiều lần và hệ thống chỉ ghi nhận giá trị 1, -1 cho kết quả cuối cùng. Trong số 6120 người dùng đã lọc ra được 5311 người dùng đã tham gia lập trình ít nhất 20 bài dù đúng hoặc sai để tiến hành thử nghiệm.

UserId	QuestionId	Result
1	1	-1
1	2	1
1	3	1
2	1	1
2	4	-1
3	5	-1
4	5	1
5	3	1
6	1	1

Hình 3. 3: *Mô tả tập dữ liệu Dlab*

3.2. Phương pháp thực nghiệm và kết quả

3.2.1. Phương pháp thực nghiệm

Đầu tiên, toàn bộ dữ liệu thực nghiệm được chia thành hai phần, một phần được sử dụng làm dữ liệu huấn luyện (training) ký hiệu là:U_{tr}, phần còn lại được sử dụng để kiểm tra (testing) ký hiệu là: U_{te}.

Tập Utr chứa 80% đánh giá và tập U_{te} chứa 20% đánh giá. Dữ liệu được sử dụng để thực nghiệm và so sánh đánh giá phương pháp Neural Collaborative Filtering (NCF) so với các phương pháp Collaborative Filtering (CF) và Matrix Factorization (MF).

3.2.2. Độ đo đánh giá

Mean Squared Error (MSE) là sai số bình phương trung bình, tính bình phương của sai số giữa giá trị dự đoán và giá trị thực tế, sau đó lấy trung bình của các bình phương sai số đó, theo công thức dưới đây:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(r_i - \hat{r}_i)^2 \quad (15)$$

Trong đó:

- N là số điểm dữ liệu
- r_i là giá trị thực
- $\hat{r}_i$ là giá trị dự đoán.

MSE càng thấp thì dự báo càng tốt.

3.2.3. Kết quả thực nghiệm

Đề án tiến hành thực nghiệm phương pháp Neural Collaborative Filtering (NCF) so với các phương pháp Collaborative Filtering (CF) và Matrix Factorization (MF). Các tham số được cài đặt ở mô hình NCF với các tập dữ liệu MovieLens-1M và Dlab cụ thể như sau:

Bộ dữ liệu MovieLens-1M:

- Số thuộc tính ẩn: 150
- Số lớp ẩn: 3
- Số nút ở mỗi lớp ẩn: 250
- Tỷ lệ dropout được áp dụng sau mỗi lớp ẩn: 0.2

Bộ dữ liệu Dlab

- Số thuộc tính ẩn: 64
- Số lớp ẩn: 3
- Số nút ở mỗi lớp ẩn: 150
- Tỷ lệ dropout được áp dụng sau mỗi lớp ẩn: 0.2

Các tập dữ liệu được chia ngẫu nhiên với mỗi lần thực nghiệm với dữ liệu huấn luyện chiếm 80%, dữ liệu kiểm tra chiếm 20%. Đề án sử dụng độ đo MSE để đánh giá và thu được kết quả ở bảng 3.1, kết quả này là kết quả tốt nhất của các lần thực nghiệm.

	MovieLens-1M	**Dlab**
Collaborative Filtering (CF)	0.9313	0.0742
Matrix Factorization (MF)	0.7870	0.0547
Neural Collaborative Filtering (NCF)	**0.7685**	**0.0515**

Bảng 3. 1: Kết quả thực nghiệm

Giá trị MSE trong bảng 3.1 cho thấy phương pháp đề xuất Neural Collaborative Filtering (NCF) cho giá trị MSE nhỏ hơn phương pháp Collaborative Filtering (CF) và Matrix Factorization (MF) trên hai tập dữ thực nghiệm là Movielen-1M và Dlab. Cụ thể, trong trường hợp dữ liệu Movielen-1M có giá trị đánh giá (rating) từ 0.5 đến 5 thì giá trị MSE của phương pháp CF, MF, NCF lần lượt là 0.9313, 0.7870, 0.7685. Giá trị MSE của các phương pháp CF, MF, NCF trên tập dữ liệu Dlab lần lượt là 0.0742, 0.0547, 0.0515. Kết quả thu được pháp Neural Collaborative Filtering (NCF) có giá trị MSE nhỏ hơn các phương pháp còn lại.

3.3. Kết luận chương

Chương 3 đã trình bày về dữ liệu và quá trình thực nghiệm trên tập dữ liệu thực tế áp dụng phương pháp đã đề xuất ở chương 2 để thực hiện so sánh và đánh giá. Kết quả thu được phương pháp đề xuất đạt kết quả tốt và có thể hỗ trợ cho việc đưa tư vấn.

KẾT LUẬN VÀ KIẾN NGHỊ

Đề án này tập trung nghiên cứu về phương pháp học sâu cho hệ tư vấn. Cụ thể đề án đã đạt được các kết quả sau:

- Nghiên cứu tổng quan về hệ tư vấn, các khái niệm và các phương pháp tiếp cận trong hệ tư vấn
- Nghiên cứu về phương pháp học sâu trong hệ tư vấn.
- Đề xuất phương pháp học sâu cho hệ tư vấn, tiến hành xây dựng dữ liệu và thực nghiệm so sánh đánh giá phương pháp đề xuất với các phương pháp khác trên tập dữ liệu xây dựng được.

Do thời gian thực hiện đề án không nhiều nên tác giả chưa có điều kiện nghiên cứu thêm nhiều phương pháp. Trong tương lai, nếu có điều kiện, tác giả sẽ tập trung nghiên cứu để xây dựng, cải tiến các phương pháp học sâu và tiến hành thực nghiệm thêm trên các tập dữ liệu thực tế khác để áp dụng vào các hệ tư vấn.

DANH MỤC TÀI LIỆU THAM KHẢO

[1]. Nguyễn Mạnh Sơn, Nguyễn Duy Phương. Một phương pháp tư vấn cộng tác cho các cổng lập trình trực tuyến. Kỷ yếu Hội nghị KHCN Quốc gia lần thứ XIV về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), TP. HCM, ngày 23-24/12/2021 DOI: 10.15625/vap.2021.0039

[2]. Isinkaye, F.O., Y.O. Folajimi, and B.A. Ojokoh, Recommendation systems: Principles, methods and evaluation. Egyptian Informatics Journal, 2015. 16(3): p. 261-273.

[3]. Isinkaye, F.O., Y.O. Folajimi, and B.A. Ojokoh, Recommendation systems: Principles, methods and evaluation. Egyptian Informatics Journal, 2015. 16(3): p. 261-273.

[4]. J. Ben Schafer, J. Konstan, and J. Riedl, ‘‘Recommender systems in ecommerce,’’ in Proc. 1st ACM Conf. Electron. Commerce, Denver, CO, USA, 1999, pp. 158–166.

[5]. E. Çano and M. Morisio, ‘‘Hybrid recommender systems: A systematic literature review,’’ Intell. Data Anal., vol. 21, no. 6, pp. 1487–1524, 2017.

[6]. R. Sharma, D. Gopalani, and Y. Meena, ‘‘Collaborative filtering-based recommender system: Approaches and research challenges,’’ in Proc. Int. Conf. Comput. Intell. Commun. Technol., 2017, pp. 1–6.

[7]. Aggarwal, C.C., Model-Based Collaborative Filtering, in Recommender Systems: The Textbook. 2016, Springer International Publishing: Cham. p. 71- 138.

[8]. Y. Li, L. Lu, and X. Li, ‘‘A hybrid collaborative filtering method for multiple-interests and multiple-content recommendation in ecommerce,’’ Expert Syst. Appl., vol. 28, no. 1, pp. 67–77, 2005.

[9]. R. Burke, ‘‘Hybrid recommender systems: Survey and experiments,’’ User Model. User-Adapted Interact., vol. 12, no. 4, pp. 331–370, 2002.

[10]. Jiliang Tang, Xia Hu, Huan Liu. Social Recommendation: A Review. Social Netw. Analys. Mining 3 (4): 1113-1133, 2013.

[11]. Koren Y., Bell R., 2011. Advances in Collaborative Filtering, in: Recommender Systems Handbook. Springer, Boston, MA, pp. 145-186. doi:10.1007/978-0-387-85820-3_5.

[12]. Koren, Y., Bell, R., Volinsky, C., 2009. Matrix Factorization Techniques for Recommender Systems. Computer 42, 30–37. doi:10.1109/MC.2009.263

[13]. Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, Tat-Seng Chua, arXiv:1708.05031v2 [cs.IR] 26 Aug 2017. Neural Collaborative Filtering

[14]. Ricci, F., Rokach, L., & Shapira, B. (2015). Recommender Systems Handbook (2nd ed.). Springer.

[15]. Steffen Rendle, Walid Krichene, Li Zhang, John Anderson. Neural Collaborative Filtering vs. Matrix Factorization Revisited. (2020). RecSys '20: Proceedings of the 14th ACM Conference on Recommender Systems

[16]. https://grouplens.org/datasets/movielens/

[17]. https://code.ptit.edu.vn/