



ĐỀ CƯƠNG KHOÁ LUẬN TỐT NGHIỆP
**XÂY DỰNG HỆ THỐNG GỢI Ý SẢN PHẨM
BẰNG PHƯƠNG PHÁP HỌC SÂU**
(Building Recommendation System using Deep Learning)

1 THÔNG TIN CHUNG

Người hướng dẫn:

– ThS. Trần Trung Kiên (Khoa Công nghệ Thông tin)

Nhóm sinh viên thực hiện:

1. Đào Đức Anh (MSSV: 1712270)
2. Nguyễn Thành Nhân (MSSV: 1712631)

Loại đề tài: Nghiên cứu

Thời gian thực hiện: Từ 1/2021 đến 6/2021

2 NỘI DUNG THỰC HIỆN

2.1 Giới thiệu về đề tài

Hệ thống gợi ý sản phẩm (recommendation system) là một lớp con của hệ thống lọc thông tin thường được tích hợp trên các trang web nhằm hỗ trợ người dùng (user) tìm kiếm được đúng sản phẩm (item) cần thiết và tìm cách dự đoán *xếp hạng* hoặc *sở thích* của họ cho một sản phẩm.

Các nội dung nói trên được gọi là các gợi ý, được tính toán thông qua dữ liệu của người dùng về các sản phẩm đó như lịch sử mua hàng, phản hồi, đánh giá của người dùng, ...

Có hai cách tiếp cận chính để xây dựng hệ thống gợi ý là *Collaborative filtering* và *Content-based filtering*. Trong đó, *Collaborative filtering* là hướng tiếp cận dựa trên các tương tác của người dùng với sản phẩm trong quá khứ để đưa ra các gợi ý. *Content-based filtering* là hướng tiếp cận dựa trên các thông tin của người dùng như: độ tuổi, giới tính, công việc, ... và các thông tin của sản phẩm như: phân loại, khối lượng, giá tiền, ... để đưa ra các gợi ý.

Collaborative filtering là hướng tiếp cận được sử dụng rộng rãi trong các hệ thống gợi ý. Mô hình gợi ý sản phẩm dựa trên Collaborative filtering được phát biểu như sau:

- Cho input là lịch sử tương tác của người dùng với các item: số lượt nhấp chuột, ấn thích, đánh giá, ...
- Yêu cầu: đưa ra tập các item (không có trong lịch sử) được dự đoán là phù hợp với người dùng (tự động làm bằng máy).

Trong thời gian gần đây, các nghiên cứu chỉ ra rằng việc áp dụng mạng nơ-ron nhân tạo (neural network) trong việc cài đặt cho Collaborative filtering mang lại những tín hiệu tích cực. Và đây cũng là hướng tiếp cận mà chúng em chọn để tìm hiểu.

2.2 Mục tiêu đề tài

- Hiểu rõ được tình hình nghiên cứu của bài toán xây dựng hệ thống gợi ý hiện nay (biết được các hướng tiếp cận phổ biến đồng thời cũng như là ý tưởng và ưu nhược điểm của các hướng tiếp cận đó; ngoài ra còn nắm được các thách thức và thuận lợi trong việc giải quyết bài toán gợi ý sản phẩm). Từ cơ sở đó có thể chọn ra một hướng tiếp cận phù hợp để tìm hiểu sâu và thực hiện cài đặt mô hình theo hướng đã chọn.
- Nắm rõ các kiến thức nền tảng bên dưới (toán học, xác suất thống kê, học

máy, ...) của mô hình đã chọn.

- Cài đặt lại mô hình để đạt được kết quả như trong bài báo tương ứng; có thể tiến hành thêm các thí nghiệm ngoài báo cáo để thấy rõ hơn về ưu nhược của mô hình.
- Trên cơ sở kiến thức nắm được từ mô hình có thể thực hiện các tối ưu (về kết quả, tốc độ huấn luyện, ...).
- Rèn luyện các kĩ năng: suy nghĩ rõ ràng, lên kế hoạch làm việc, làm việc theo nhóm, khả năng trình bày báo cáo, thuyết trình, ...

2.3 Phạm vi của đề tài

Đề tài làm với dữ liệu có các phản hồi ẩn của người dùng cho một sản phẩm; cụ thể, chúng em dự kiến sẽ làm với bộ dữ liệu thường được sử dụng trong lĩnh vực xây dựng hệ thống gợi ý là MovieLens. Về cơ bản, đề tài chỉ tìm hiểu và cài đặt lại mô hình của một bài báo uy tín, ngoài ra có thể thêm các thí nghiệm khác cũng như huấn luyện trên bộ dữ liệu khác ngoài bài báo để thấy rõ hơn về ưu, nhược điểm của mô hình. Lý do chúng em giới hạn đề tài như vậy là vì: (i) chúng em thấy riêng việc hiểu rõ mô hình (và các kiến thức nền tảng bên dưới) và có thể tự cài đặt lại đã tốn khá nhiều thời gian, và (ii) chúng em xác định là chỉ trên cơ sở hiểu rõ mô hình (và các kiến thức nền tảng bên dưới) thì mới có thể có được các cải tiến thật sự trong tương lai, cũng như là có thể vận dụng mô hình được cho các bài toán khác. Tất nhiên, trong khóa luận, nếu có đủ thời gian thì chúng em sẽ thử đề xuất và cài đặt các cải tiến; tuy nhiên, chúng em xác định đây không phải là mục đích chính.

2.4 Cách tiếp cận dự kiến

kiến. Sau đây sẽ là trình bày một số mô hình theo hướng tiếp cận sử dụng các mô hình học sâu để giải quyết bài toán xây dựng hệ thống gợi ý sản phẩm mà chúng em đã tìm hiểu được cho đến thời điểm hiện tại, cũng như là mô hình mà chúng em dự kiến sẽ chọn để tập trung tìm hiểu sâu.

- “Neural collaborative filtering” của Xiangnan He cùng các cộng sự được công bố tại hội nghị WWW’] [?] (số lần trích dẫn là 1967 tính cho đến thời điểm hiện tại là ngày 28/2/2020) là một trong những bài báo nổi bật được coi như là bước đặt nền móng cho việc áp dụng mạng nơ ron kết hợp với phương pháp truyền thống collaborative để xây dựng một hệ thống gợi ý sản phẩm. Điểm nổi bật của mô hình này là tận dụng sức mạnh của các mạng học sâu để đánh bật các thuật đoán tại thời điểm đó như Matrix Factorization, ... đã trích xuất được các đặc trưng phi tuyến từ đó mô hình có khả năng mô hình hóa dữ liệu tương tác trong quá khứ của người dùng từ đó trích xuất các đặc trưng ẩn giữa các user và các item để cải thiện độ chính xác cho hệ thống gợi ý. Mô hình được đề xuất từ nhóm tác giả gồm 2 phần:

- Generalized Matrix factorization viết tắt GMF là việc kết hợp nhân linear kernel cho với phương pháp Matrix factorization để học những các tương tác ẩn giữa tập người dùng và user (latent interaction). Trả về các embedding vector thể hiện cho người dùng và các item
- Mạng nơ ron nhiều tầng MLP (Multi layer Perceptron) để đưa ra kết quả từ embedding vector có được từ GMF

Mô hình “NeuMF” này được huấn luyện với bộ dữ liệu implicit feedback bao gồm các vector one hot coding thể hiện tương tác giữa user và các item. Mô hình này tuy đơn giản sử dụng MLP nhưng đã học được các đặc trưng phi tuyến thông qua mạng nơ ron chứ không cần trích xuất bằng tay như những công trình nghiên cứu trước đó. Và tại thời điểm ra mắt thì bài báo đã đạt được kết quả tốt nhất cho đến thời điểm đó.

- Tiếp theo đó nhiều bài báo lần lượt áp dụng mạng nơ ron và mạng hồi quy để giải quyết bài toán này như . Trong đó bài báo “Variational autoencoder for collaborative” được công bố tại hội nghị WWW2018[?] bởi nhóm nghiên cứu ở Netflix và Google, là một bài báo nổi bật khi mà sử dụng “Variational autoencoder” để phát sinh các latent variables cho việc xây dựng hệ thống gợi ý sản phẩm. Variational Autoencoder là một mô hình Autoencoder tuy

nhien thay vì encoder thực hiện encode dữ liệu thành một điểm thì mô hình generative này sẽ phát sinh phân phối xác suất để có thể phát sinh dữ liệu. Do đó dữ liệu được phát sinh từ các phân phối xác suất sẽ đảm bảo được tính chất là liên tục và hoàn chỉnh có nghĩa là 2 điểm dữ liệu gần nhau sau khi được giải mã bởi decoder vẫn sẽ gần nhau và với một điểm dữ liệu được phát sinh ngẫu nhiên từ phân phối xác suất mà được tạo ra từ encoder sẽ “mang ý nghĩa” sau khi được giải mã. Từ cơ sở này thì ta có thể thấy được việc áp dụng VAEs cho bài toán gợi ý sản phẩm và cụ thể là áp dụng cho việc mở rộng collaborative filtering khi ta có thể phát hiện và khai phá các đặc trưng giữa việc tương tác với các item của các user từ các tương tác trong quá khứ của họ. Ngoài ra thuật toán học được sử dụng trong mô hình này còn có liên kết với các khái niệm trong lĩnh vực lý thuyết thông tin để có thể tối ưu được kết quả của mô hình. Với cách tiếp cận này, kết quả từ bài báo mang lại đã đánh bại State-of-the-art tại thời điểm đó trên các tập dữ liệu lớn từ thế giới thật.

- Cho đến hiện nay nhiều cải tiến trong việc áp dụng các phương pháp học sâu để xây dựng hệ thống gợi ý sản phẩm đã được cộng đồng nghiên cứu quan tâm và phát triển nhiều hơn khi nay đã có nhiều hướng tiếp cận hiện đại như sử dụng cơ chế attention trong bài báo [?] , đồ thị tri thức [?] , ... hơn và có nhiều cải tiến từ những phương pháp trước đó như: [?] , ... đã đạt được kết quả vượt trội đáng kể.

Với những thông tin mà nhóm đã tìm hiểu ở trên thì nhóm em dự định sẽ tập trung tìm hiểu model được đề xuất ở bài báo [5]. Mặc dù không phải là mô hình đạt được kết quả tốt nhất hiện nay trong việc xây dựng hệ thống gợi ý sản phẩm tuy nhiên kiến thức nền tảng để xây dựng mô hình này bao phủ về mạng nơ ron, mô hình phát sinh (deep generative model) và kiến thức về mô hình xác suất. Ngoài ra mô hình này còn thể hiện mối liên hệ với kiến thức trong lý thuyết thông tin (information theory) do đó việc hiểu rõ được mô hình này sẽ là bước đệm cho các cải tiến sau này.

2.5 Kết quả dự kiến của đề tài

- Cài đặt lại được mô hình đề xuất trong bài báo[?]>.
- Có được kết quả thí nghiệm cho thấy mô hình tự cài đặt ra được các kết quả như trong bài báo.
- Có được các kết quả thí nghiệm để thấy rõ về ưu, nhược điểm của mô hình.
- Nếu có thời gian thì có thể cài đặt và thí nghiệm thêm các cải tiến.

2.6 Kế hoạch thực hiện

Công Việc	Thời Gian	Người Thực hiện
Tìm hiểu tình hình nghiên cứu của bài toán xây dựng hệ thống gợi ý sản phẩm, chọn ra mô hình để tập trung tìm hiểu sâu	Tháng 01/2021 - Tháng 02/2021	Anh, Nhân
Tìm hiểu lý thuyết của mô hình đã chọn (bao gồm cả việc tìm hiểu lý thuyết nền tảng bên dưới)	Tháng 03/2021	Anh, Nhân
Cài Đặt lại từ đầu mô hình để ra được kết quả giống như trong bài báo	Tháng 4/2021	Anh, Nhân
Tiến hành thí nghiệm để thấy rõ về ưu/nhược điểm của mô hình; xem xét cải tiến nếu có thể	Tháng 05/2021	Anh, Nhân
Viết cuốn và slide	Tháng 05/2021 - Tháng 06/2021	Anh, Nhân

XÁC NHẬN
CỦA NGƯỜI HƯỚNG DẪN
(Ký và ghi rõ họ tên)

TP. Hồ Chí Minh, ngày 01 tháng 03 năm 2021
NHÓM SINH VIÊN THỰC HIỆN
(Ký và ghi rõ họ tên)

ThS. Trần Trung Kiên

Đào Đức Anh

Nguyễn Thành Nhân