fit@hcmus

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

ĐỀ CƯƠNG KHOÁ LUẬN TỐT NGHIỆP

XÂY DỰNG HỆ THỐNG GỢI Ý SẢN PHẨM DỰA TRÊN MÔ HÌNH AUTOENCODER

(Building Recommendation System using Autoencoder model)

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

Bài toán xây dựng hệ thống gợi ý sản phẩm (recommendation system) được phát biểu như sau:

- Cho input là lịch sử tương tác của người dùng (user) với các sản phẩm (item) (đối với bài toán *Collaborative filtering*) hoặc có thêm các mô tả của người dùng và sản phẩm. (Các sản phẩm có thể là: quảng cáo, bộ phim, văn bản để đọc, ... tùy thuộc vào ngành nghề) (đối với bài toán *Content-base filtering*).
- Yêu cầu: đưa ra tập các items (không có trong lịch sử) được dự đoán là phù hợp với người dùng (tự động làm bằng máy).

Nếu giải quyết được bài toán này thì một ứng dụng có thể có là xây dựng hệ thống hỗ trợ các dịch vụ thương mại điện tử, quảng cáo trực tuyến: giúp các nhà cung cấp dịch vụ đưa ra sản phẩm/quảng cáo phù hợp với thị hiếu người dùng. Một ứng dụng khác có thể có khác là hệ thống gợi ý các khóa học hoặc bài giảng trong lĩnh vực giáo dục.

Trong khóa luận này, chúng em sẽ tập trung vào bài toán *Collaborative filtering* vì nó giải quyết tốt 2 nhược điểm chính của *Content-base filtering* là: (i) tận dụng được các thông tin từ các user khác và (ii) trường hợp sản phẩm thiếu hoặc ít các mô tả chi tiết.

Trong thời gian gần đây, các nghiên cứu chỉ ra rằng việc áp dụng các mô hình phát sinh (generative model) dựa trên phương pháp học sâu (deep learning), cụ thể hơn là mô hình Autoencoder trong việc cài đặt cho *Collaborative filtering* mang lại các kết quả tốt. Và đây cũng là hướng tiếp cận 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 (implicit feedback); implicit feedback là phản hồi được suy ra từ hành vi của người dùng, cụ thể, nếu người dùng xem bộ phim A, chúng ta có thể suy ra họ thích nó. 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 việc sử dụng dữ liệu implicit feedback sẽ xây dựng được một hệ thống gợi ý khách quan hơn, ngoài ra, (ii) 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 (iii) 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

Trong thời đại bùng nổ thông tin hiện nay thì việc xây dựng một hệ thống gợi ý sản phẩm hiệu quả có thể sẽ mang lại việc tăng lợi nhuận hay cải thiện trải nghiệm người dùng cho các doanh nghiệp. Do đó việc nghiên cứu xây dựng một

hệ thống gợi ý sản phẩm cũng là một chủ đề được quan tâm hiện nay. Cùng với sự phát triển mạnh mẽ của mạng học sâu (Deep Neuron Network) trong các lĩnh vực hình ảnh, âm thanh, văn bản, ... thì đã có nhiều nghiên cứu rộng rãi trong việc áp dụng mạng học sâu trong bài toán gợi ý sản phẩm. Một số hướng tiếp cận áp dụng mạng học sâu ta có thể điểm qua như: Mạng nơ ron nhiều tầng (Multi Layer Perceptron), mạng nơ ron hồi quy (Recurrent Neuron Network), Autoencoder, cơ chế attention, ... Trong khóa luận này, chúng em sẽ nghiên cứu và tìm hiểu hướng tiếp cận sử dụng Autoencoder trong việc xây dựng hệ thống gợi ý sản phẩm. Mô hình Autoencoder là một mạng nơ ron bao gồm 2 phần: (1) Encoder là một mạng nơ ron có nhiệm vụ biểu diễn dữ liệu ở chiều không gian ban đầu thành một điểm dữ liệu ở chiều không gian thấp hơn. Điểm dữ liệu ở chiều không gian mới này có thể dùng để biểu diễn các đặc trưng ẩn (latent variable) của dữ liệu ban đầu; (2) Decoder là một mạng nơ ron khác dùng để tái cấu trúc lại dữ liệu ban đầu từ các đặc trưng ẩn. Dưới đây sẽ trình bày một số hướng tiếp cận sử dụng mô hình Autoencoder để giải quyết bài toán 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:

- "AutoRec: Autoencoders Meet Collaborative Filtering"[1] được đề xuất bởi nhóm tác giả từ trường đại học quốc gia Australia ở hội nghị WWW2015, là một trong những bài báo đầu tiên áp dụng mô hình Autoencoder để xây dựng một hệ thống gợi ý sản phẩm. Mô hình AutoRec được đề xuất bởi nhóm tác giả nhậninput là vector tương tác giữa user và item, mô hình sẽ chiếu input lên một không gian có số chiều thấp hơn để biểu diễn các đặc trưng ẩn của dữ liệu, và sau đó chúng sẽ được tái cấu trúc lại với mục đích để dự đoán những tương tác bị trống trước đó. Mô hình học bằng cách tối thiểu độ lỗi trong việc tái cấu trúc dữ liệu ban đầu từ các đặc trưng ẩn và để tránh overfitting thì mô hình có áp dụng kĩ thuật chính quy hóa (regularization) trong việc huấn luyện mạng nơ ron trong Autoencoder. Trong bài báo này, tác giả đã thực nghiệm và cho kết quả đáng mong đợi khi mô hình đã đánh bại được mô hình nổi bật trong các hệ thống gợi ý sản phẩm theo hướng Collaborative Filtering

là Matrix Factorization. Bên cạnh đó, tác giả còn có các thử nghiệm mang lại những dấu hiệu tích cực trong việc áp dụng mạng nơ ron khi tăng số tầng ẩn hay kích thước của các tầng ẩn trong mạng thì kết quả được cải thiện theo. Điều này đã dẫn đến việc áp dụng các phương pháp học sâu trong việc xây dựng hệ thống gợi ý sản phẩm được quan tâm và nghiên cứu nhiều hơn trong cộng đồng nghiên cứu khoa học.

- Sau đó không lâu đã có nhiều bài báo được xuất bản dựa trên mô hình Autoencoder để xây dựng hệ thống gợi ý sản phẩm. Một trong những bài báo cáo khoa học nổi bật đó là "Hybrid Recommender System based on Autoencoders" [2] của Florian Strub cùng các cộng sự, nhóm tác giả đã đề xuất mô hình có tên là CFN, đây được xem như là một mô hình mở rộng của Autorec. Input của mô hình tương đồng với Autorec, tuy nhiên nhóm tác giả đề xuất 2 điểm cải thiện cho mô hình đó là: (1) áp dụng kĩ thuật khử nhiễu (denoising); (2) ngoài dữ liệu tương tác của user và item thì tác giả còn kết hợp với những thông tin khác như dữ liệu về user hay dữ liệu của item để giải quyết hai vấn đề thường gặp trong hệ thống gợi ý sản phẩm là dữ liệu thưa[1] và cold start[2]; Bên cạnh đó có thể kể đến "Collaborative Denoising Auto-Encoders for Top-N Recommender Systems" [3] nói về mô hình CDE được đề xuất bởi nhóm nghiên cứu tại trường đại học Beihang, Trung Quốc, cũng là một bài báo nổi bật trong việc áp dụng mô hình Autoencoder, mô hình này sử dụng một biến thể của Autoencoder là Denoising Autoencoder để xây dựng hệ thống gợi ý sản phẩm, điểm khác biệt của mô hình CDE là input của mô hình sẽ được thêm nhiễu nhằm giúp mô hình tránh tình trạng overfitting. Và theo bài báo thì CDE có được kết quả vượt trội hơn so với những state of the art tại thời điểm được đề xuất.

- Đặc biệt mô hình Multi-VAE được đề xuất bởi nhóm nghiên cứu của Netflix và Google tại hội nghị WWW2018 với tiêu đề "Variational Autoencoders for Collaborative Filtering" [4], là một trong những bài báo đáng chú ý trong việc

[1] Dữ liệu ban đầu có nhiều trường bị trống, hay tương tác giữa user và item quá ít

[2] khi một user mới hoặc item mới được thêm vào hệ thống thì sẽ có rất ít tương tác của user hay đối với item đó thì sẽ khó để tìm được sự tương đồng giữa các user trong việc tương tác với item đó

xây dựng hệ thống gợi ý sản phẩm, mô hình được cải thiện so với những mô hình trước đó bằng cách áp dụng Variational Autoencoders (VAEs), là một biến thể của Autoencoder. Ở Variational Autoencoder thì latent variables sẽ là phân phối xác suất thay vì một điểm dữ liệu trong chiều không gian thấp hơn như Autoencoder hay Denoising Autoencoder. Do đó dữ liệu sau khi tái cấu trúc sẽ đảm bảo được tính chất "liên tục" và "toàn vẹn" cho dữ liệu. Điều đó có nghĩa là 2 điểm dữ liệu gần nhau trong không gian dữ liệu ban đầu nếu gần nhau thì sau khi tái cấu trúc sẽ gần nhau (tính liên tục) và việc phát sinh ngẫu nhiên dữ liệu từ phân phối xác suất ẩn sẽ đảm bảo được tính toàn vẹn của dữ liệu (tính toàn vẹn). Từ cơ sở này ta có thể thấy được áp dụng VAEs cho bài toán gợi ý sản phẩm và cụ thể là theo hướng Collaborative Filtering là một lựa chọn phù hợp khi sự tương đồng giữa các user trong việc tương tác với các item sẽ được khám phá tốt hơn. Do đó ta có thể xây dựng một hệ thống gợi ý sản phẩm hiệu quả. Ngoài ra trong bài báo này, tác giả còn đề xuất việc sử dụng Multinomial Loglikelihood là phù hợp cho việc mô hình hóa dữ liệu implicit feedback thay vì những hàm phân phối phổ biến như logistic likelihood hay Gaussian Likelihood. Và mô hình đã mang lại một kết quả vượt bậc so với những mô hình trước đó.

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 . 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 [4].
- 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 slidess	Tháng 05/2021 - Tháng 06/2021	Anh, Nhân

Tài liệu

[1] S. Sedhain, A. K. Menon, S. Sanner, and L. Xie, "AutoRec: Autoencoders Meet Collaborative Filtering," *WWW '15 Companion: Proceedings of the 24th International Conference on World Wide Web*, p. 111–112, 2015.

[2] F. Strub, J. Mary, and R. Gaudel, "Hybrid recommender system based on autoencoders," *DLRS 2016: Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, pp. 11–16, 2016.

[3] Y. Wu, C. DuBois, A. X. Zheng, and M. Ester, "Collaborative denoising auto-encoders for top-n recommender systems," *WSDM '16: Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, p. 153–162, 2016.

[4] D. Liang, R. G. Krishnan, M. D. Hoffman, and T. Jebara, "Variational autoencoders for collaborative filtering," *WWW '18: Proceedings of the 2018 World Wide Web Conference*, p. 689–698, 2017.

[5] D. Kim and B. Suh, "Enhancing vaes for collaborative filtering: flexible priors and gating mechanisms," *RecSys '19: Proceedings of the 13th ACM Conference on Recommender Systems*, p. 403–407, 2019.

TP. Hồ Chí Minh, ngày 01 tháng 03 năm 2021

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

ThS. Trần Trung Kiên

NHÓM SINH VIÊN THỰC HIỆN
(Ký và ghi rõ họ tên)

Đào Đức Anh Nguyễn Thành Nhân