Hybrid-based hotel recommendation system

Nguyễn Hoàng Minh, Tạ Nhật Minh, Nguyễn Thiện Thuật, Huỳnh Văn Tín
Faculty of Information Science and Engineering, University of Information Technology
Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
{20521609, 20521614, 20521998}gm.uit.edu.vn
tinhv@uit.edu.vn

Abstract

Du lịch là hoạt động thiết yếu của con người nhằm đáp ứng nhu cầu tham quan, nghỉ dưỡng, tìm hiểu và khám phá. Trong một chuyến du lịch, chỗ ở là một trong những yếu tố quan trọng nhất, ảnh hưởng đến chất lượng và trải nghiệm của du khách. Chính vì vậy, để giúp người dùng tìm thấy khách sạn phù hợp với thông tin sở thích du lịch của họ, chúng tôi đã xây dựng một hệ khuyến nghị khách sạn (hotels recommendation system) bằng cách sử dụng phương pháp lai trên 2 phương pháp khuyến nghị chính là khuyến nghị dựa trên nội dung (content-based) và khuyến nghị dựa trên lọc cộng tác (colaborative-filtering) nhằm giúp tối ưu hóa trải nghiệm người dùng trên các trang web du lịch. Trong báo cáo này, chúng tôi thực hiện xây dựng bộ dữ liệu khách sạn bao gồm bộ dữ liệu với hơn về thông tin của các khách sạn ở Việt Nam và lịch sử feedback của người dùng trên 3 trang web về du lịch: traveloka, ivivu và booking.com. Bên cạnh đó, chúng tôi cũng đề xuất hệ thống khuyến nghị dựa trên phương pháp lọc dựa trên nội dung thông qua vector chú ý mong muốn của người dùng và đề xuất phương pháp lai với mô hình lọc cộng tác nhằm đưa ra nhiều kết quả đa dạng hơn cho người dùng. Chúng tôi cũng thực hiện đánh giá trên bộ dữ liệu bằng các độ đo MAP, NDCG@K, Precision@K và Recall@K. Kết quả tốt nhất của mô hình trên bộ dữ liệu sẽ được chúng tôi sử dụng để triển khai hệ thống demo. Nghiên cứu được thực hiện cho thấy cách tiếp cận của chúng tôi giúp mô hình khuyến nghị dự đoán đa dạng và có thể sử dụng để phát triển thêm trong tương lai. Chúng tôi tin rằng với ý tưởng này sẽ giới thiệu một Recommend System có thể được sử dụng để giải quyết những hạn chế còn tồn tại và hơn thế nữa là mở rộng khả năng ứng dụng của nó.

Keywords: Recommendation System, Hotel Bookings, Hybrid-based, Collaborative Filtering, Content-based.

1 Giới thiệu

Hệ thống gợi ý (Recommender System) là một hệ thống máy học sử dụng dữ liệu lịch sử để dự đoán các sản phẩm, dịch vụ hoặc nội dung mà người dùng có thể quan tâm. Chúng được sử dụng trong nhiều lĩnh vực khác nhau, bao gồm thương mại điện tử, truyền thông xã hội, giải trí, v.v. Có nhiều loại hệ thống gợi ý khác nhau, Một số loại hệ thống gợi ý phổ biến bao gồm:

- Hệ thống gợi ý dựa trên nội dung (Content-based recommender systems)(Herlocker et al., 2004): Hệ thống này dự đoán sở thích của người dùng dựa trên các sản phẩm, dịch vụ hoặc nội dung mà họ đã tương tác trong quá khứ.

- Hệ thống gợi ý dựa trên lọc cộng tác (Collaborative filtering recommender systems)(Sarwar et al., 2001a): Hệ thống này dự đoán sở thích của người dùng dựa trên sở thích của những người dùng khác có sở thích tương tự.

- Hệ thống gợi ý dựa trên đặt trưng (Factor-based recommender systems)(Koren et al., 2009a): Hệ thống này sử dụng các thuật toán học máy để học các mối quan hệ giữa các sản phẩm, dịch vụ hoặc nội dung và sở thích của người dùng.

Trong lĩnh vực du lịch, sự phát triển nhanh chóng của du lịch và internet đã dẫn đến sự tăng cường nhu cầu đặt phòng khách sạn trực tuyến. Ở phần Collaborative Filtering Based Hotel Recocommendations (Chaudhari and Thakkar, 2020) đã cho ta một cái nhìn tổng quát về các cách tiếp cận về xây dựng mô hình gợi ý các khách sạn cho người dùng như dựa trên sở thích người dùng, các blog, multi-criteria CF-based recommendations, adaptive neuro-fuzzy inference system (ANFIS), expectation maximization (EM), PCA, LDA... Với sự tiến bộ của internet, quy mô và độ phức tạp của các trang web đặt phòng khách sạn ngày càng gia tăng bởi vì lượng người dùng ngày càng tăng và sự thay đổi sở thích liên tục của người dùng thì các phương pháp đã được đề xuất sẽ phải chịu lượng tính toán

rất lớn. Điều này đặt ra thách thức cho người dùng khi cố gắng tìm kiếm thông tin về khách sạn và tốn nhiều thời gian.

Trong đề tài này, chúng tôi đã phát triển một hệ thống khuyến nghị khách sạn sử dụng các phương pháp lai từ hai kỹ thuật Content-based và Colaborative-filtering. Hệ thống này giúp người dùng dễ dàng tìm thấy khách sạn phù hợp với sở thích du lịch của họ, đồng thời hệ thống của chúng tôi giải quyết luôn vấn đề về hiện tượng người dùng lần đầu sử dụng trang web mà vẫn có thể có những gợi ý phù hợp cho họ thông qua thông tin mà họ đã cung cấp. Những công việc trong báo cáo này của chúng tôi có thể được tóm tắt như sau:

- Xây dựng bộ dữ liệu khách sạn ở Việt Nam từ 3 trang web: Ivivu, Traveloka, Booking.com.

- Đề xuất mô hình CB dựa trên vector chú ý mong muốn của người dùng.

- Chạy thực nghiệm các mô hình SOTA và Baselines của phương pháp CF trên bộ dữ liệu đã được thu thập.

- Xây dựng mô hình khuyến nghị lai từ 2 phương pháp CB và CF.

- Xây dựng demo hệ thống khuyến nghị khách sạn từ phương pháp đề xuất trên bộ dữ liệu đã thu thập.

2 Các công trình liên quan

Trên thế giới đã có nhiều hệ thống đề xuất khách sạn được xây dựng dựa trên các phương pháp khác nhau như Collaborative Filtering, Content-Based, Domain-Specific. (Chaudhari and Thakkar, 2020) Những nghiên cứu này tập trung vào hai tác vụ chính: "tác vụ đề xuất" (Kim et al., 2011) và "tác vụ dự đoán đánh giá" (Sun et al., 2015)

Các tác giả đã sử dụng nhiều cách tiếp cận khác nhau để xây dựng các hệ khuyến nghị phù hợp nhất cho người dùng; Nghiên cứu của tác giả (Nilashi et al., 2015) tập trung vào bài toán "dự đoán đánh giá" dựa trên nhiều yếu tố khác nhau và đề xuất "multi-criteria CF-based recommendations" cho lĩnh vực du lịch. Với độ đo MAE, kết quả trên tập dữ liệu thu thập từ TripAdvisor.com dao động từ 0,86 đến 1,37 trên tập dữ liệu kiểm thử. Như trong nghiên cứu của (Shambour et al., 2022) đã có đề xuất kết hợp dựa trên cách tiếp cận enhanced user-based CF và enhanced item-based CF có tên là a fusion-based multi-criteria CF (FBMCCF), mô hình được đánh giá trên tập dữ liệu TripAdvisor MC dataset (Jannach et al., 2014) để so sánh với các mô hình Multi-Criteria Collaborative Filtering khác, với độ đo bằng MAE và RMSE cho kết quả vượt trội hơn các mô hình trước đó. Ở cách tiếp cận khác, (Kaya, 2020) phát triển một hệ thống đề xuất khách sạn mới dựa trên link prediction bằng cách sử dụng the customer-hotel bipartite network. Trong lĩnh vực rộng hơn liên quan đến du lịch, Các tác giả (Al-Ghobari et al., 2021) có ý tưởng sử dụng a location-aware traveler assistance (LAPTA) làm ngữ cảnh để đề xuất thông qua mô hình KNN Item-Based Collaborative Filtering.

Trong quá trình đọc các nghiên cứu về lĩnh vực này, chúng tôi nhận thấy rằng chưa có nhiều nghiên cứu về ngành du lịch ở Việt Nam và các mô hình trong lĩnh vực này chưa giải quyết được hiện tượng cold-start. Do đó, trong nghiên cứu này, chúng tôi đề xuất phát triển một hệ thống đề xuất khách sạn mới nhằm giải quyết khoảng trống đó.

3 Dữ liệu

Trong các mô hình khuyến nghị, chất lượng dữ liệu đóng vai trò then chốt trong việc xác định độ chính xác của kết quả mô hình. Do đó, ngay từ đầu cần thiết phải thu thập dữ liệu chất lượng cao. Chúng tôi đã phối hợp thu thập, xử lý dữ liệu và thực hiện dán nhãn dữ liệu, như hình 1 thể hiện, để thu được tập dữ liệu tối ưu cho việc xây dựng và phát triển Hybrid-based hotel recommender systems.

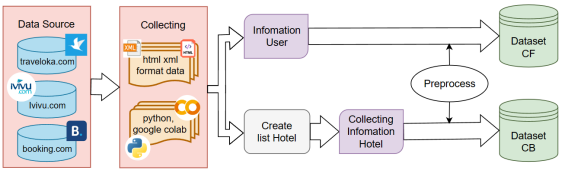


Figure 1: Quy trình thu thập và tiền xử lý dữ liệu.

3.1 Thu thập dữ liệu

Nghiên cứu này thu thập dữ liệu từ các trang web đặt phòng khách sạn uy tín trên thế giới trong đó có Việt Nam thông qua phương pháp kết hợp giữa trích xuất dữ liệu tự động bằng các thư viện Python chuyên dụng như Beautiful Soup và thu thập thủ công. Ngoài ra, nhóm nghiên cứu đã tìm kiếm trên nhiều trang web khác nhau để tìm nguồn đánh giá liên quan và thông tin đa dạng, hữu ích cho việc phát triển mô hình. Quá trình lựa chọn trang web đánh giá tin cậy và phù hợp cho hệ thống được thực hiện dựa trên các tiêu chí như xếp hạng uy tín, số lượng người dùng lớn. Dữ liệu thu thập được dưới dạng HTML và XML, bao gồm nhiều thông tin cần thiết để xây dựng và phát triển mô hình.

Đối với bộ dữ liệu cho Colaborative Filtering, chúng tôi gọi lệnh để truy cập đến API của danh sách đánh giá cho từng khách sạn để thu dữ liệu về dưới dạng Json với những trang web có API. Còn những trang web không có API, chung tôi thu thập bằng các gọi lệnh request thông thường đến trang web và tìm thông tin dữ liệu dưới định dạng HTML. Bộ dữ liệu được thu thập gồm 18274 mẫu dữ liệu với 7 thuộc tính chứa các thông tin của khách hàng đặc phòng khách sạn như ID khách hàng, tên khách hàng, ID khách sạn, tên khách sạn, địa điểm khách sạn, đánh giá của khách hạng, ngày đánh giá.

Đối với bộ dữ liệu cho Content-based Filtering, vì không có API về thông tin của các khách sạn, chúng tôi quyết định tiến hành thu thập dữ liệu thủ công dựa vào những khách sạn nằm trong bộ dữ liệu cho Colaborative Filtering. Bộ dữ liệu được thu thập gồm 309 mẫu dữ liệu với 11 thuộc tính chứa các thông tin chi tiết của từng khách sạn như Tên khách sạn, địa điểm, điểm đánh giá tổng thể của khách sạn, số lượng đánh giá của khách hàng cho khách sạn, giá cả, tiện tích, chất lượng khách sạn, khoảng cách so với chung tâm, các địa điểm xung quanh, các địa điểm lân cận và link đặt phòng khách sạn.

3.2 Tiền xử lí dữ liệu

Bộ dữ liệu cho Colaborative Filtering

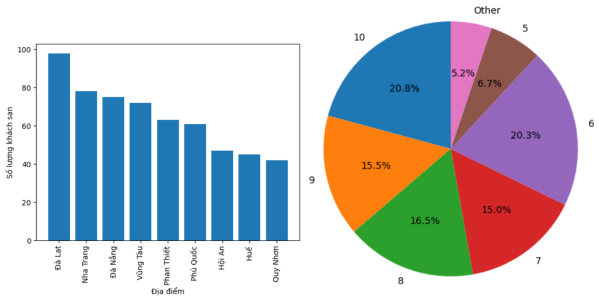
Để có được bộ dữ liệu cho lọc cộng tác, chúng tôi tiến hành các bước tiền xử lý dữ liệu cơ bản gồm việc loại bỏ những dòng dữ liệu có user chứa các ký tự đặt biệt, đồng bộ tên khách sạn của các trang web khác nhau, chuẩn hóa kiểu dữ liệu Date cho thuộc tính thời gian đánh giá. Sau khi xử lý sơ bộ các vấn đề cần phải giải quyết, chúng tôi tiến hành chia bộ dữ liệu ra thành 3 loại bao gồm: Thông tin user, thông tin khách sạn và lịch sử đánh giá. Đối với thông tin user và khách sạn, chúng tôi tiến hành tạo ra 2 dataframe để chứa thông tin cần thiết của từng loại, sau đó loại bỏ những mẫu dữ liệu. Kết quả thu được đối với thông tin user có 10875 user với 2 thuộc tính. Dữ liệu thông tin khách sạn gồm 597 mẫu dữ liệu với 3 thuộc tính. Đối với dữ liệu history, chúng tôi sử dụng bộ dữ liệu ban đầu, loại bỏ các thuộc tính có trong 2 bộ dữ liệu thông tin user và khách san ngoại từ ID user và ID Holel.

Bộ dữ liệu cho Content-based Filtering

Để tạo ra bộ dữ liệu cho lọc dựa trên nội dung, chúng tôi tiến hành lấy dữ liệu của các người dùng có tham gia đánh giá từ một lần trở lên, thu được bộ dữ liệu gồm lịch sử đánh giá của 9276 người dùng. Tiếp theo, tiến hành lấy danh sách các người dùng đánh giá từ 15 khách sạn trở lên để lấy làm tập dữ liệu kiểm thử. Kết quả thu được 18270 dòng dữ liệu cho tập huấn luyện và 447 dòng dữ liệu cho tập kiểm thử. Dữ liệu lịch sử người dùng dùng để huấn luyện gồm ID của người dùng và ID khách sạn, và sẽ được nối với bộ dữ liệu thông tin các khách sạn được thu thập trước đó. Gộp tất cả các thông tin về Tiện nghi(Facilities) của khách sạn mà mỗi người dùng đã đi thành một cho người dùng. Xử lí tương tự cho các thông tin về Xung quanh(Around), Lân cận(Vicinity) và Giá cả(Price category). Kết quả thu được 9276 dòng dữ liệu tương ứng với 9267 người dùng và được sử dụng để học các đặc trưng cho người dùng. Đối với dữ liệu dùng để trích xuất đặc trưng cho khách sạn, chúng tôi lấy các thuộc tính Tiện nghi, Xung quanh, Lân cận, Giá cả của từng khách sạn. Hoàn tất quá trình thu được 310 dòng dữ liệu ứng với 310 khách sạn.

Sau khi trích xuất được tập dữ liệu để huấn luyện và kiểm thử, chúng tôi tiến hành tiền xử lí dữ liệu. Chúng tôi xóa bỏ kí tự dấu phẩy, chuyển sang kí tự viết thường sau đó tiến hành tokenize để phân tách văn bản thành các từ, cụm từ. Sau đó tiến hành tạo bộ từ điển cho từng đặc trưng từ bộ dữ liệu huấn luyện. Đưa vào mô hình Word2Vec để ánh xạ các đặc trưng sang không gian vectơ. Sau khi tiền xử lí thu được 4 vectơ đặc trưng cho 4 yếu tố của khách sạn và 4 vectơ đặc trưng cho người dùng.

3.3 Thống kê bộ dữ liệu



(a) Thống kê số lượng khách sạn theo từng địa điểm.

(b) Phân phối giá trị đáng giá khách sạn.

Figure 2: Thống kê đánh giá của khách sạn và khách hàng.

Hình 2a cho thấy sự đồng đều của bộ data về số lượng khách sạn tại các địa điểm. Đà Lạt là địa điểm có nhiều khách sạn nhất Việt Nam với gần 100 khách sạn. Bên cạnh đó các địa điểm như Đà Nẵng, Nha Trang, Vũng Tàu, Phú Quốc, Phan Thiết có số lượng khách sạn nhiều hơn các địa điểm khác

từ 70 đến 80 khách sạn. Hình 2b thể hiện sự phân phối khá đồng đều về giá trị đánh giá trong khoảng từ 6 đến 10. Cho thấy được các khách sạn trong bộ dữ liệu có chất lượng từ trung bình trở lên.

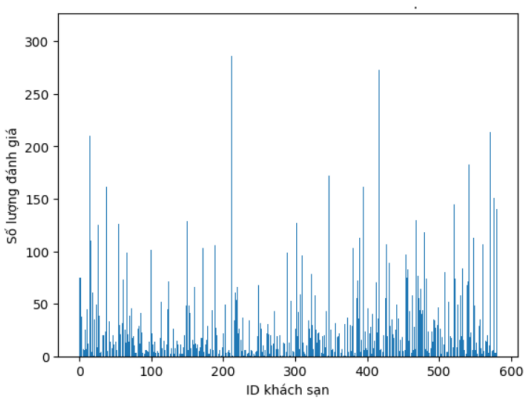


Figure 3: Thống kê số lượng đánh giá của từng khách sạn.

Hình 3 thể hiện sự đa dạng lớn về số lượng đánh giá của khách hàng cho từng khách sạn, các giá trị giao động chủ yếu từ 0 đến 50 lượt đánh giá. Bên cạnh đó, hình 4 thể hiện được sự đồng đều hơn về số lượng đánh của từng khách hàng, các giá trị giao động từ 10 đến 20 lượt đánh giá.

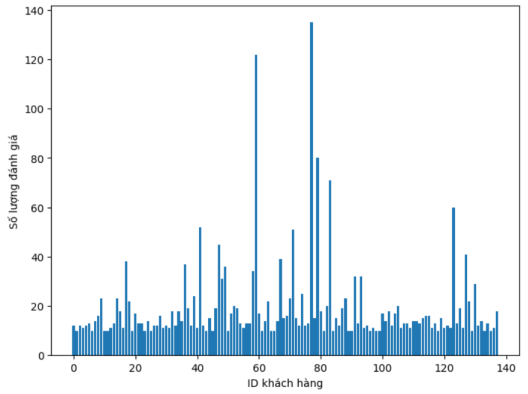


Figure 4: Thống kê số lượng đánh giá của từng khách hàng.

4 Mô hình khuyến nghị

4.1 Mô hình khuyến nghị dựa trên nội dung

Trong báo cáo này, chúng tôi trình bày về mô hình khuyến nghị dựa trên nội dung. Mục tiêu của chúng tôi là xây dựng một mô hình khuyến nghị khách sạn, có khả năng đề xuất các nội dung tương tự dựa trên đặc trưng của chúng. Mô hình khuyến nghị dựa trên nội dung là một phương pháp trong lĩnh vực khuyến nghị, trong đó các đề xuất được tạo ra dựa trên sự tương đồng về nội dung giữa các mục

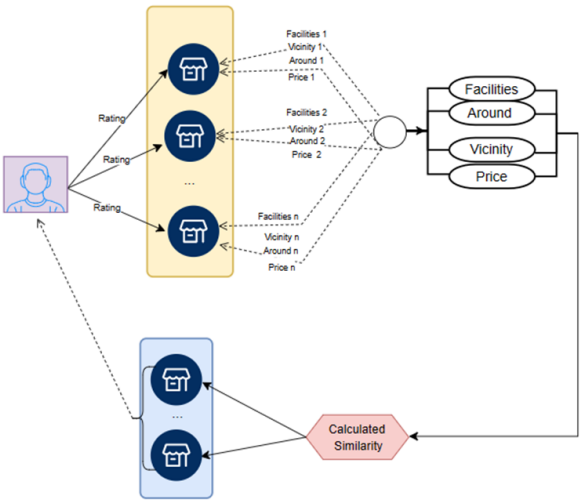


Figure 5: Mô hình Content based Filtering.

tiêu. Mô hình này không phụ thuộc vào thông tin về người dùng, mà chỉ tập trung vào các đặc trưng của nội dung để tạo ra các đề xuất phù hợp.

Phương pháp khuyến nghị dựa trên nội dung của chúng tôi được thể hiện qua Figure 5. Với mỗi người dùng đã thực hiện đánh giá khách sạn, chúng tôi thu thập các thông tin về 4 đặc đưng của khách sạn được đánh giá và tổng hợp lại thông tin 4 đặc trưng đó cho từng người dùng. Các thông tin được chuyển về không gian vectơ thông qua mô hình Word2Vec, kết thúc quá trình thu được các vectơ đặc trưng cho 4 thuộc tính của người dùng và khách sạn. Tiến hành tổng hợp các vectơ thành phần thành vectơ tổng hợp đại diện cho người dùng và khách sạn theo công thức (1) và (2). Độ tương đồng được chúng tôi tính toán dựa trên độ đo Cosine similarity (3).

$$e_h = 0.25 \times (h_F + h_A + h_V + h_P) \quad (1)$$

$$e_u = \sum_{i=1}^{4} \lambda_i h_i \quad (2)$$

$$cos(e_u, e_h) = \frac{e_u \cdot e_h}{|e_u| * |e_h|} \quad (3)$$

Với e_h và e_u là vectơ đặc trưng của người dùng U và khách sạn H. λ_i là các tham số thể hiện mức độ quan tâm của user với đặc trưng h_i. h_F, h_A, h_V, h_P lần lượt là các vectơ thể hiện đặc trưng về Tiện nghi, Xung quanh, Lân cận, Giá cả được trích xuất từ Word2Vec.

4.2 Mô hình khuyến nghị dựa trên Lọc cộng tác

Ở phần này, chúng tôi tiến hành thực nghiệm trên bộ dữ liệu đã thu thập[1] các mô hình baselines và SOTA của phương pháp lọc cộng tác (CF) trong lĩnh vực recommendation. Sau đó, kết quả tốt nhất của mô hình trên bộ dữ liệu sẽ được chúng tôi sử dụng để triển khai cho hệ thống.

Recomenmder: Chúng tôi thực nghiệm các mô hình sau: Most popular (MostPop) là một cách tiếp cận không cá nhân hóa bằng việc gợi ý những mặt hàng phổ biến nhất cho tất cả người dùng. RBM(Salakhutdinov et al., 2007) sử dụng a class of two-layer undirected graphical models trên dữ liệu rating, NCF(He et al., 2017a) sử dụng Neural network based cho lọc cộng tác, LightGCN(He et al., 2020) đơn giản hóa thiết kế của Graph Convolution Network để phù hợp cho việc đề xuất.

Cornac: Chúng tôi thực nghiệm các mô hình sau: Bayesian Personalized Ranking (BPR)(Mnih and Salakhutdinov, 2007) Probabilistic Matrix Factorization (PMF)(Salah et al., 2016) là 2 mô hình tương tác giữa user-item theo xác xuất, Spherical K-means (SKM)(He et al., 2017b) là phương pháp phân cụm dữ liệu trong không gian nhiều chiều để đưa ra khuyến nghị. NeuMF, MLP, GMF, NGCF(Liang et al., 2018), IBPR(Wang et al., 2019), VAECF(Steck, 2019), EASER(Truong et al., 2021), and BiVAE(Isinkaye et al., 2015) là các phương pháp gần đây được sử dụng để phân tích so sánh trong lọc cộng tác.

Memory-based Methods: UserKNN(Sarwar et al., 2001b) và ItemKNN(Koren et al., 2009b) là các kỹ thuật tận dụng tính tương tự của người dùng và tính tương tự của item để dự đoán xếp hạng trong lọc cộng tác. Ứng với mỗi mô hình, chúng tôi sử dụng 2 độ đo tương đồng cosine và pearson để tính toán khoảng cách.

Chúng tôi triển khai các phương pháp thông qua code và cài đặt các tham số trên Cornac [2] và Recommender [3] như Bảng 1.

Về độ đo đánh giá, theo nghiên cứu trước (Chen and Liu, 2017), đối với việc đánh giá hiệu quả đề xuất, chúng tôi dùng các độ đo như MAP, NDCG@5, NDCG@10, Precision@5, Precision@10, Recall@5, Recall@10.

[1] Về việc xử lý dữ liệu để huấn luyện mô hình CF, trong quá trình thực nghiệm, chúng tôi nhận thất rằng các người dùng đi ít khách sạn (dưới 4 lần) sẽ làm giảm độ chính xác rất nhiều cho mô hình khi tiến hành dự đoán những user có ít lịch sử như vậy. Chính vì vậy nên chúng tôi đã loại bỏ các lịch sử của người dùng này trước khi huấn luyện mô hình.

[2] https://github.com/PreferredAI/cornac

[3] https://github.com/recommenders-team/recommenders

4.3 Mô hình lai

Trên thực tế, khi người dùng lần đầu tiên đến trang web thì các hệ thống đề xuất sẽ không thể để xuất được vì hệ thống chưa có dữ liệu lịch sử của họ, vì vậy cho nên người dùng sẽ không nhận được các đề xuất phù hợp. Thêm vào đó, hệ thống đề xuất theo phương pháp lọc cộng tác (CF) sẽ bị vấn đề về cold-start, tức là cần phải có lịch sử của người dùng để đề xuất, chính vì vậy nên cũng sẽ không có kết quả đề xuất cho người dùng, nên phương pháp lọc cộng tác sẽ không giải quyết được trong tình huống này.

Để giải quyết vấn đề trên, chúng tôi đã đề xuất kết hợp 2 mô hình CB và CF nhằm tránh trường hợp hệ thống recommend không có kết quả từ việc cho người dùng điền vào "interested form" các vấn đề mà họ quan tâm khi đặt phòng khách sạn để làm dữ liệu đầu vào ban đầu cho hệ thống làm cơ sở để đề xuất.

Kết quả của hệ thống đề xuất lai được biểu diễn theo công thức như sau:

$$reslt_{CB} = rec_{CB}(Cos\,(e_u\,, e_h)) \quad (4)$$

$$reslt_{CF} = rec_{CF}(u) \quad (5)$$

$$reslt_{Hybrid} = (1-\lambda)reslt_{CB} + \lambda reslt_{CF} \quad (6)$$

Với rec_{CB} là hệ thống đề xuất dựa trên nội dung (CB) và rec_{CF} là hệ thống đề xuất dựa trên lọc cộng tác (CF), reslt là các kết quả đề xuất của các hệ thống đề xuất trả về. λ = 0.5 nếu rec_{CF} có kết quả và λ = 0 nếu rec_{CF} không có kết quả.

Phương pháp lai này cũng đồng thời làm cho gợi ý của hệ thống cho người dùng thêm đa dạng hơn với những người đã có lịch sử và có điền form nhờ vào việc mô hình quan tâm tới cả form và lịch sử của họ.

5 Thực nghiệm

5.1 Độ đo đánh giá

Mean Average Precision at K (mAP@K) là một độ đo đánh giá quan trọng trong việc đánh giá chất lượng của mô hình khuyến nghị. Nó tính toán độ chính xác trung bình của các mục tiêu đúng trong K đề xuất đầu tiên. Mô hình được đánh giá dựa trên khả năng đưa ra các đề xuất đúng và sắp xếp chúng theo thứ tự chính xác.

$$mAP@K = \frac{1}{n}\sum_{k=1}^{n} AP_k \quad (7)$$

Mô hình	Cài đặt
Recommender	
MostPop(2009)	TOP_K=10
RBM(2012)	hidden_units=100, training_epoch=200, minibatch_size=30, keep_prob=0.9, with_metrics=True
NCF(2017)	model_type="NeuMF", n_factors=4, layer_sizes=[16,8,4], n_epochs=100, batch_size=256, learning_rate=1e-3
LightGCN(2020)	n_layers=3, batch_size=1024, epochs=25, learning_rate=0.005, eval_epoch=5, top_k=10
Cornac	
PMF(2007)	k=40, max_iter=100, learning_rate=0.001, lambda_reg=0.001
NGCF(2019)	num_epochs=100, emb_size=64, layer_sizes=[64, 64, 64], dropout_rates=[0.1, 0.1, 0.1], early_stopping={"min_delta": 1e-4, "patience": 50}, batch_size=1024, learning_rate=0.001, lambda_reg=1e-5
IBPR(2017)	k=40
BiVAECF(2021)	k=50, encoder_structure=[100], act_fn="tanh", likelihood="pois", n_epochs=100, batch_size=32, learning_rate=0.001
EASE^R b>0(2019)	lamb=500, posB=True
EASE^R(2019)	lamb=500, posB=False
BPR(2012)	N/A
VAECF(2018)	k=10, autoencoder_structure=[20], act_fn="tanh", likelihood="mult", n_epochs=100, batch_size=100, learning_rate=0.001, beta=1.0, @seed=123, use_gpu=True
GMF(2017)	num_factors=8, num_epochs=50, learner="adam", batch_size=256, lr=0.001, num_neg=50, seed=123
NeuMF(2017)	num_factors=8, layers=[64, 32, 16, 8], act_fn="tanh", learner="adam", num_epochs=50, batch_size=256, lr=0.001, num_neg=50, seed=123
SKMeans(2016)	k=5, max_iter=50, tol=1e-10, seed=123
MLP(2017)	layers=[64, 32, 16, 8], act_fn="tanh", learner="adam", num_epochs=50, batch_size=256, lr=0.001, num_neg=50, seed=123
Memory-based	
UserKNN-cosine(1998)	k=200
UserKNN-pearson(1998)	k=200
ItemKNN-cosine(2001)	k=100
ItemKNN-pearson(2001)	k=100

Table 1: Cài đặt các siêu tham số của tất cả các mô hình thực nghiệm.

Precision at K (precision@K) tính toán tỷ lệ giữa số lượng các mục tiêu đúng trong K đề xuất đầu tiên và K. Độ đo này đo lường khả năng của mô hình trong việc đưa ra các đề xuất chính xác trong một tập hợp K mục tiêu.

$$Precision@K = \frac{TP@K}{K} \tag{8}$$

Recall at K (recall@K)tính toán tỷ lệ giữa số lượng các mục tiêu đúng trong K đề xuất đầu tiên và tổng số lượng mục tiêu thực tế N. Độ đo này đo lường khả năng của mô hình trong việc tìm ra tất cả các mục tiêu quan trọng trong một tập hợp K đề xuất.

$$Recall@K = \frac{TP@K}{N} \tag{9}$$

Normalized Discounted Cumulative Gain at K (nDCG@K) đo lường chất lượng của mô hình khuyến nghị bằng cách xem xét sự xếp hạng chính xác của các mục tiêu quan trọng đối với người dùng. Độ đo này tính toán giá trị DCG@K (Discounted Cumulative Gain) và chuẩn hóa nó bằng giá trị tối đa của DCG@K.

$$nDCG@k = \frac{DCG@K}{IDCG@K} \tag{10}$$

Trong đó:

$$DCG@K = \sum_{1=1}^{k} \frac{rel_i}{\log_2(i+1)} \tag{11}$$

IDCG@K là giá trị tối đa của DCG@K, được tính bằng cách sắp xếp tất cả mục tiêu thực tế theo thứ tự giảm dần của rel_i và tính DCG@K cho sắp xếp đó.

5.2 Mô hình khuyến nghị dựa trên nội dung

Đối với mô hình khuyến nghị dựa trên nội dung, chúng tôi tiến hành thực nghiệm độ hiệu suất của mô hình bằng cách thực hiện trên dữ liệu lịch sử gần nhất với dữ liệu toàn bộ lịch sử người dùng. Với mội trường hợp thì sẽ tiến hành lấy kích thước chiều của vectơ đặc trưng lần lượt là 100, 200, 300. Các vectơ tổng hợp cho người dùng và khách sạn được lấy bằng cách lấy trung bình các vectơ thành phần. Đánh giá dựa trên hai độ đo Precision và Recall.

Bảng 2 thể hiện kết quả khuyến nghị của mô hình khuyến nghị dựa trên nội dung. Nhìn chung Khi so sánh giữa lịch sử gần nhất và toàn bộ lịch sử, ta thấy rằng kết quả Precision@5 và Precision@10 không có sự thay đổi đáng kể. Trong số các chiều (100, 200, 300), Precision@10 có xu hướng tăng dần từ Dim = 100 đến Dim = 300, tuy nhiên, sự khác biệt không lớn. Các độ đo Recall@5 và Recall@10 cũng thể hiện tương tự. Tổng quan, kết quả Precision và Recall trong bảng không cho thấy sự khác biệt đáng kể giữa lịch sử gần nhất và toàn bộ lịch sử. Tuy nhiên, kết quả Precision và Recall không cao, điều này có thể đòi hỏi cải thiện chất lượng của mô hình khuyến nghị để tăng độ chính xác và độ phủ của các đề xuất.

	P@5[1]	**P@10**[2]	**R@5**[3]	**R@10**[4]
Last his[5]				
Dim100	0.157	0.114	0.027	0.042
Dim200	0.157	0.129	0.028	0.043
Dim300	0.114	0.121	0.018	0.042
All his[6]				
Dim100	0.138	0.125	0.022	0.045
Dim200	0.138	0.131	0.023	0.043
Dim300	0.175	0.138	0.028	0.050

Note: [1]Precision@5, [2]Precision@10, [3]Recall@5, [4]Recall@10, [5]Lịch sử gần nhất, [6]Toàn bộ lịch sử

Table 2: Kết quả mô hình Content-based Filtering.

Kết quả thực nghiệm không đạt kết quả cao có thể bắt nguồn từ một vài lí do sau. Thứ nhất, do hạn chế về số lượng dữ liệu đưa vào mô hình, đồng thời các khách sạn được thu thập trong bộ dữ liệu thường có các đặc trưng về Tiện nghi, Xung quanh, Lân cận, Giá cả không quá khác biệt. Điều này dẫn đến số lượng từ đưa vào từ điển huấn luyện cho Word2Vec bị hạn chế, làm cho việc chuyển hóa các vectơ đặc trưng có độ chính xác không cao. Khi tính toán độ tương đồng dẫn đến không có sự khác biệt quá lớn giữa các khách sạn làm giảm độ chính xác của mô hình. Thứ hai, tập dữ liệu kiểm thử quá nhỏ, chỉ gồm 15 user và số lượng khách sạn mà mỗi người dùng đã đi trong tập kiểm thử là khá lớn (15 đến 65) làm cho độ phủ của các đề xuất trên tập thực tế bị giảm xuống.

5.3 Mô hình khuyến nghị dựa trên lọc cộng tác

Bảng 3 cung cấp toàn bộ kết quả của các phương pháp thực nghiệm trên bộ dữ liệu của chúng tôi. Kết quả in đậm và gạch chân cho biết kết quả tốt nhất; in đậm cho biết kết quả tốt nhất thứ hai, gạch chân cho biết kết quả tốt thứ ba.

Đầu tiên, nhìn chung, các phương pháp dựa trên mạng nơ-ron có hiệu suất tốt hơn các phương pháp dựa trên phép đo tương tự và dựa trên bộ nhớ. Bởi

Methods	MAP@10	NDCG@5	NDCG@10	Precision@5	Precision@10	Recall@5	Recall@10
Recommender							
MostPop(2009)	0.0363	0.0490	0.0688	0.0301	0.0261	0.0601	0.1154
RBM(2012)	0.0574	**0.0649**	**0.1018**	**0.0334**	**0.0332**	**0.0946**	**0.2036**
NCF(2017)	0.0330	0.0367	0.0552	0.0174	0.0169	0.0504	0.1021
LightGCN(2020)	0.0537	0.0637	0.0806	0.0285	0.0233	0.0813	0.1289
Cornac							
PMF(2007)	0.0460	0.0402	0.0470	0.0174	0.0126	0.0478	0.0671
NGCF(2019)	0.0502	0.0377	0.0465	0.0157	0.0122	0.0492	0.0738
IBPR(2017)	0.0588	0.0489	0.0606	0.0235	0.0174	0.0702	0.1021
BiVAECF(2021)	0.0651	0.0518	0.0740	0.0252	0.0222	0.0660	0.1306
EASE^R b>0(2019)	0.0652	0.0614	0.0740	0.0287	0.0209	0.0667	0.1089
EASE^R(2019)	0.0653	0.0541	0.0724	0.0252	0.0226	0.0642	0.1159
BPR(2012)	0.0656	0.0548	0.0801	0.0261	0.0243	0.0847	0.1550
VAECF(2018)	0.0615	0.0492	0.0733	0.0270	0.0230	0.0745	0.1430
GMF(2017)	0.0643	0.0539	0.0791	0.0261	0.0243	0.0847	0.1550
NeuMF(2017)	0.0667	0.0551	0.0793	0.0261	0.0230	0.0893	0.1586
Skmeans(2016)	**0.0716**	0.0579	0.0811	0.0261	0.0235	0.0803	0.1448
MLP(2017)	**0.0754**	**0.0642**	**0.0861**	0.0270	0.0235	0.0874	0.1463
Memory-based							
UserKNN-cosine(1998)	0.0192	0.0138	0.0102	0.0363	0.0193	0.1768	0.1780
UserKNN-pearson(1998)	0.0215	0.0176	0.0135	0.0368	0.0196	0.1971	0.1983
ItemKNN-cosine(2001)	0.0207	0.0136	0.0101	0.0390	0.0212	0.1837	0.1859
ItemKNN-pearson(2001)	0.0223	0.0159	0.0120	0.0394	0.0216	**0.1975**	**0.2004**

Table 3: Hiệu suất của các phương pháp khác nhau trên tập dữ liệu của chúng tôi.

vì các đặt trưng được trính xuất từ mạng nơ-ron thể hiện đặt điểm của người dùng tốt hơn việc quan sát rating của họ.

Thứ hai, MAP@10: MAP@10 là một chỉ số đo lường độ presicion tổng thể của hệ thống gợi ý. Các phương pháp có hiệu suất tốt nhất về MAP@10 là những phương pháp có thể gợi ý các sản phẩm mà người dùng có khả năng thích thú với xác suất cao nhất. Các phương pháp có hiệu suất tốt nhất là MLP, Skmeans và NeuMF, với MAP@10 lần lượt là 0,0716, 0,0754 và 0,0667.

Thứ ba, NDCG@5 và NDCG10 là một chỉ số đo lường độ chính xác của hệ thống gợi ý đối với các sản phẩm nằm trong top 5 và 10. Các phương pháp có hiệu suất tốt nhất về NDCG@K là những phương pháp có thể gợi ý các sản phẩm có thứ hạng cao nhất với xác suất cao nhất. Các phương pháp có hiệu suất tốt nhất là RBM, MLP, EASE và Skmeans.

Thứ tư, Precision@5 và Precision@10 là một chỉ số đo lường độ chính xác của hệ thống gợi ý đối với các sản phẩm nằm trong top 5 và 10. Các phương pháp có hiệu suất tốt nhất về Precision@K là những phương pháp có thể gợi ý các sản phẩm mà người dùng thực sự đã thích với xác suất cao nhất. Các phương pháp có hiệu suất tốt nhất là MostPop, RBM, EASE và BPR.

Thứ năm, Recall@5 và Recall@10 là một chỉ số đo lường khả năng bao phủ của hệ thống gợi ý. Các phương pháp có hiệu suất tốt nhất về Recall@K là những phương pháp có thể gợi ý tất cả các sản phẩm mà người dùng thực sự đã thích. Các phương pháp có hiệu suất tốt nhất là RBM, MLP, NeuMF.

Từ bảng kết quả quan sát được, chúng tôi nhận thấy rằng mô hình RBM đạt hiệu quả tốt nhất trên bộ dữ liệu của chúng tôi. Chính vì vậy chúng tôi sử dụng mô hình RBM cho phương pháp CF để xây dựng hệ thống đề xuất

6 Triển khai hệ thống

Chúng tôi tiến hành triển khai hệ thống gợi ý khác sạn Hybrid-based bằng framework Streamlit của python. Tại đây, chung tôi xây dựng dự trên 3 tác vụ chính. Ngoài ra còn hỗ trợ hệ thống lọc bằng Filter sau khi mô hình đã đề xuất danh sách khách sạn dùng chung các tác vụ.

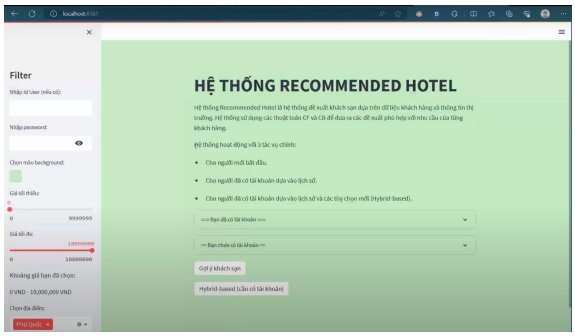


Figure 6: Triển khai hệ thông bằng framework Streamlit.

6.1 Đăng nhập và khuyến nghị dựa trên Lọc cộng tác

Đối với tác vụ này, người dùng cần phải đăng nhập bằng ID và mật khẩu để sử dụng. Khi người dùng nhập sai thông tin đăng nhập, hệ thống sẽ thông báo sai thông tin và yêu cầu đăng nhập lại. Bên cạnh đó thông báo về cách hướng sử dụng tác vụ thứ 2 nếu không có tài khoản. Khi người dùng đăng nhập thành công, hệ thống sẽ trả về thông tin của người dùng, lịch sử đánh giá khách sạn và danh sách tối đa 10 khách sạn được gợi ý dựa trên Lọc cộng tác có thể lọc bằng Filter.

6.2 Chọn thông tin và khuyến nghị dựa trên nội dung

Ở tác vụ này, người dùng không cần đăng nhập tài khoản. Tuy nhiên cần phải chọn đầy đủ thông tin yêu cầu của hệ thống. Nếu như không chọn đủ thông tin theo yêu cầu, hệ thống sẽ thông báo cho người dùng cần điền đầy đủ thông tin. Khi điều đầy đủ thông tin, hệ thống sẽ tiến hành chạy mô hình khuyến nghị dựa trên nội dung và xuất ra kết quả gồm danh sách tối đa 10 khách sạn. có thể lọc bằng Filter.

6.3 Khuyến nghị Hybrid-based

Tác vụ được xây dựng bằng cách kết hợp kết quả khuyến nghị dựa trên Lọc cộng tác với khuyến nghị dựa trên nội dung bằng cách cộng trọng số của hai mô hình. Để có thể dùng được tác vụ này, người dùng cần phải đăng nhập thông tin tài khoản và chọn đầy đủ thông tin theo yêu cầu. Nếu thiếu một trong hai điều kiện trên, hệ thống sẽ thông báo cần hoàn thành đúng theo từng yêu cầu để sử dụng. Khi đủ điều kiện, hệ thống sẽ hiển thị thông tin của người dùng, lịch sử đánh giá và danh sách tối đa 10 khách sạn có thể lọc bằng Filter.

7 Kết luận

Trong báo cáo này, chúng tôi đã xây dựng được bộ dữ liệu phục vụ cho 2 hệ thống đề xuất dựa trên nội dung và lọc cộng tác, dữ liệu bao gồm các lịch sử feedback của người dùng và kèm theo các thông tin của khách sạn trên 3 trang web về du lịch nổi tiếng của Việt Nam là Traveloka, Ivivu và Booking.com. Bên cạnh đó, chúng tôi đề xuất một cách tiếp cận mới của phương pháp dựa trên nội dung (CB) và đồng thời cũng hoàn thành thực hiện đánh giá các mô hình SOTA và Baselines của phương pháp CF trên bộ dữ liệu. Thêm vào đó, nhằm giải quyết vấn đề cool-start của mô hình CF, chúng tôi đã xây dựng hệ thống đề xuất lai từ kết quả của 2 phương pháp CF và CB. Cuối cùng là thực hiện xây dựng web demo về hệ thống khuyến nghị khách sạn dựa trên phương pháp mà chúng tôi đã đề xuất thông qua thư viện streamlit.

Tuy nhiên, đề tài cũng còn gặp nhiều hạn chế khiến cho chất lượng bộ dữ liệu cũng như chứng minh mức độ hiệu quả của phương pháp chưa được chặt chẽ. Bộ dữ liệu của chúng tôi vẫn còn rất nhỏ so với các bộ dữ liệu hiện tại (gần 18 nghìn trong khi với movilens dataset ít nhất là 100 nghìn[4], amazon dataset ít nhất là 1triệu7 [5]) và chưa có giải pháp xử lý dữ liệu bị thưa. Chưa thực nghiệm và đánh giá được mức độ hiệu quả của Hybrid recommendation system và CB so với các phương pháp baslines hiện có. Các siêu tham số cũng được cài đặt ngẫu nhiên và chưa đánh giá mức độ hiệu quả của chúng. Chưa chứng minh được khả năng mở rộng của mô hình và nguồn tài nguyên tính toán là những lỗ hổng trong đề tài chưa thực hiện triệt để. Hạn chế về thời gian thực hiện đề tài cũng là một trong những lý do khiến cho bộ dữ liệu và độ chính xác của mô hình chưa thể đạt kết quả tốt hơn.

Trong tương lai, đề tài có thể được phát triển thêm để hoàn thiện tính chặt chẽ của báo cáo, ví dụ như giải quyết các vấn đề mà đề tài này gặp phải, thực nghiệm thêm để đánh giá mức độ hiệu quả của mô hình CB và hybrid được đề xuất so với các mô hình hiện có. Thực hiện các thực nghiệm nhằm chọn ra bộ tham số tối ưu cho mô hình. Thu thập thêm dữ liệu từ các trang web đặt phòng khách sạn du lịch trực tuyến và thu thập thêm các thuộc tính mới như: cảm xúc, loại item, bình luận,. . . để bộ dữ liệu phục vụ thực nghiệm được nhiều phương pháp của tác vụ recommend hơn.

Acknowledgements

Cảm ơn người dùng trên 3 website nổi tiếng: traveloka, ivivu, booking.com đã đóng góp cho báo cáo hoàn thiện bộ dữ liệu.

References

Mohanad Al-Ghobari, Amgad Muneer, and Suliman Mohamed Fati. 2021. Location-aware personalized traveler recommender system (lapta) using collaborative filtering knn. *Computers, Materials & Continua*, 69(2).

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

[5] https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2/

Kinjal Chaudhari and Ankit Thakkar. 2020. A comprehensive survey on travel recommender systems. *Archives of Computational Methods in Engineering*, 27:1545–1571.

Mingang Chen and Pan Liu. 2017. Performance evaluation of recommender systems. *International Journal of Performability Engineering*, 13(8):1246.

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. pages 639–648.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017a. Neural collaborative filtering. pages 173–182.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017b. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pages 173–182.

Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. 2004. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53.

Folasade Olubusola Isinkaye, Yetunde O Folajimi, and Bolande Adefowoke Ojokoh. 2015. Recommendation systems: Principles, methods and evaluation. *Egyptian informatics journal*, 16(3):261–273.

Dietmar Jannach, Markus Zanker, and Matthias Fuchs. 2014. Leveraging multi-criteria customer feedback for satisfaction analysis and improved recommendations. *Information Technology & Tourism*, 14:119–149.

Buket Kaya. 2020. A hotel recommendation system based on customer location: a link prediction approach. *Multimedia Tools and Applications*, 79:1745–1758.

Heung-Nam Kim, Abdulmajeed Alkhaldi, Abdulmotaleb El Saddik, and Geun-Sik Jo. 2011. Collaborative user modeling with user-generated tags for social recommender systems. *Expert Systems with Applications*, 38(7):8488–8496.

Yehuda Koren, Robert Bell, and Chris Volinsky. 2009a. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

Yehuda Koren, Robert Bell, and Chris Volinsky. 2009b. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

Dawen Liang, Rahul G Krishnan, Matthew D Hoffman, and Tony Jebara. 2018. Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 world wide web conference*, pages 689–698.

Andriy Mnih and Russ R Salakhutdinov. 2007. Probabilistic matrix factorization. *Advances in neural information processing systems*, 20.

Mehrbakhsh Nilashi, Othman bin Ibrahim, Norafida Ithnin, and Nor Haniza Sarmin. 2015. A multi-criteria collaborative filtering recommender system for the tourism domain using expectation maximization (em) and pca–anfis. *Electronic Commerce Research and Applications*, 14(6):542–562.

Aghiles Salah, Nicoleta Rogovschi, and Mohamed Nadif. 2016. A dynamic collaborative filtering system via a weighted clustering approach. *Neurocomputing*, 175:206–215.

Ruslan Salakhutdinov, Andriy Mnih, and Geoffrey Hinton. 2007. Restricted boltzmann machines for collaborative filtering. pages 791–798.

Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001a. Item-based collaborative filtering recommendation algorithms. pages 285–295.

Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001b. Item-based collaborative filtering recommendation algorithms. pages 285–295.

Qusai Shambour, Ahmad Adel Abu Shareha, and Mosleh M Abu-Alhaj. 2022. A hotel recommender system based on multi-criteria collaborative filtering. *Inf. Technol. Control.*, 51(2):390–402.

Harald Steck. 2019. Embarrassingly shallow autoencoders for sparse data. In *The World Wide Web Conference*, pages 3251–3257.

Jianshan Sun, Gang Wang, Xusen Cheng, and Yelin Fu. 2015. Mining affective text to improve social media item recommendation. *Information Processing & Management*, 51(4):444–457.

Quoc-Tuan Truong, Aghiles Salah, and Hady W Lauw. 2021. Bilateral variational autoencoder for collaborative filtering. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 292–300.

Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural graph collaborative filtering. In *Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval*, pages 165–174.