

Travel Recommendation System

Nguyễn Minh Tiên^{1,2,3}, Nguyễn Minh Tuệ^{1,2,3}, Huỳnh Văn Tín^{1,2,4}

¹Faculty of Information Science and Engineering, University of Information Technology,
Ho Chi Minh City, Vietnam

²Vietnam National University, Ho Chi Minh City, Vietnam

³{20522010, 20522129}@gm.uit.edu.vn

⁴tinhv@uit.edu.vn

Abstract

Trong thời đại internet phát triển, việc tìm kiếm thông tin và đặt phòng khách sạn, nhà hàng, và điểm thu hút trực tuyến ngày càng khó khăn. Điều này đặt ra thách thức cho người dùng, đòi hỏi họ phải dành nhiều thời gian và công sức để tìm kiếm những lựa chọn phù hợp. Để giải quyết vấn đề này, chúng tôi đã phát triển một hệ thống khuyến nghị du lịch đa dạng, giúp người dùng dễ dàng khám phá khách sạn, nhà hàng, và điểm thu hút phù hợp với sở thích cá nhân. Báo cáo trình bày quá trình xây dựng bộ dữ liệu Hotels, Restaurants, Attractions Booking Dataset. Chúng tôi thực hiện xử lý dữ liệu để tạo ra các tập training và testing cho các phương pháp content-based filtering, hybrid, graph-based, collaborative filtering, và model-based filtering. Bằng cách sử dụng ba độ đo là cosine, pearson, và jaccard, chúng tôi đo độ tương đồng giữa người dùng và giữa các địa điểm du lịch để đưa ra các khuyến nghị chính xác. Kết quả đạt được được đánh giá thông qua Precision@10, NDCG@10, MSE, và R2. Hiệu suất tốt nhất là 0.09 MSE, 0.93 R2, 0.88 Precision@10, và 0.92 NDCG@10 khi sử dụng phương pháp user-user cosine. Những kết quả này là minh chứng cho độ chính xác và hiệu suất xuất sắc của hệ thống khuyến nghị của chúng tôi.

Keywords: Du lịch trực tuyến, Hotels Dataset, Restaurants Dataset, Attractions Dataset, Content-based Filtering, Collaborative Filtering, Hybrid, Graph-based, Model-based Filtering.

1 Giới thiệu

Hệ thống Khuyến nghị (Recommender System hoặc Recommendation System) là một lĩnh vực quan trọng trong machine learning, được phát triển độc lập mạnh mẽ trong khoảng 10–15 năm gần đây, đồng thời là một phần không thể thiếu của hệ thống thông tin trên internet. Trong lĩnh vực này, có hai thực thể chính: người dùng (user) và sản phẩm (item). Người dùng biểu diễn các cá

nhân sử dụng, trong khi sản phẩm có thể là bất cứ thứ gì từ bộ phim, bài hát, sách đến người dùng khác trong trường hợp kết bạn. Mục tiêu chính của các hệ thống khuyến nghị là dự đoán mức độ quan tâm của người dùng đối với một sản phẩm cụ thể [3, 25, 5]. Thông qua việc phân tích hành vi người dùng và thu thập thông tin về sản phẩm, các hệ thống này có khả năng đưa ra các gợi ý chính xác và phù hợp, đồng thời tối ưu hóa trải nghiệm người dùng trên nền tảng trực tuyến.

Nghiên cứu trong lĩnh vực hệ thống khuyến nghị đã được đặc trưng bởi nhiều tác phẩm quan trọng như "Introduction to recommender systems handbook" của Ricci et al. (2011) [19], "Toward the next generation of recommender systems" của Adomavicius và Tuzhilin (2005) [1], và "Matrix factorization techniques for recommender systems" của Koren et al. (2009) [10]. Các nghiên cứu này đã đóng góp đáng kể vào việc hiểu biết và phát triển của lĩnh vực, cung cấp cơ sở lý thuyết cho nghiên cứu hiện tại.

Đối mặt với sự phát triển nhanh chóng của ngành du lịch và sự gia tăng không ngừng của thông tin trên internet, việc tìm kiếm và đề xuất thông tin du lịch phù hợp với mong muốn của người dùng ngày càng trở nên phức tạp. Trong bối cảnh này, hệ thống khuyến nghị đang trở thành một công cụ quan trọng để giúp người dùng khám phá và trải nghiệm những điểm đến mới một cách hiệu quả. Đồng thời, ngành du lịch không chỉ đòi hỏi sự linh hoạt trong việc khuyến nghị khách sạn và nhà hàng, mà còn yêu cầu khả năng dự đoán đánh giá từ phía người dùng để cung cấp những đề xuất chất lượng và phản ánh chính xác về trải nghiệm du lịch. Chính vì vậy, nghiên cứu này đặt ra mục tiêu xây dựng một hệ thống khuyến nghị toàn diện, kết hợp nhiều phương pháp và mô hình máy học để tối ưu hóa trải nghiệm du lịch cho người dùng. Hệ thống khuyến nghị của chúng tôi tập trung vào các khía cạnh chính của du lịch, bao gồm khách sạn, nhà hàng và điểm thu hút, và sử dụng nhiều

kỹ thuật khác nhau như content-based, user-based, item-based, hybrid, và graph-based để cung cấp các đề xuất phong phú và đa dạng. Điều này giúp đáp ứng đa dạng nhu cầu của người dùng và mang lại trải nghiệm cá nhân hóa.

Ngoài ra, chúng tôi tích hợp các mô hình máy học và neural network để dự đoán đánh giá từ phía người dùng. Việc này giúp chúng tôi cung cấp đánh giá chính xác hơn và tối ưu hóa quá trình khuyến nghị.

Trong quá trình nghiên cứu, chúng tôi sử dụng các phép đo tương đồng như cosine similarity, pearson correlation và jaccard similarity để đánh giá hiệu suất của hệ thống khuyến nghị trong việc so sánh và lọc thông tin. Qua đó, chúng tôi hy vọng nghiên cứu của mình sẽ đóng góp một phần nhỏ vào sự phát triển của lĩnh vực hệ thống khuyến nghị và ứng dụng của chúng trong ngành du lịch.

Với những mục tiêu và phương pháp nghiên cứu đều đặn, chúng tôi tin rằng nghiên cứu này sẽ mang lại những thông điệp quan trọng và ý nghĩa cho cộng đồng nghiên cứu và ngành công nghiệp du lịch.

Trong báo cáo này, chúng tôi bắt đầu bằng việc giới thiệu các nghiên cứu liên quan trong Phần 2. Sau đó, chúng tôi trình bày về quá trình thu thập và xử lý dữ liệu để tạo ra bộ dữ liệu Hotels, Restaurants, Attractions Dataset, đó là cơ sở dữ liệu chúng tôi sử dụng cho bài toán Travel Recommendation System. Phần 3 tập trung vào phác thảo các bước xử lý dữ liệu tùy thuộc vào phương pháp recommendation cụ thể, từ đó tạo ra các tập dữ liệu training và testing. Phần 4 mô tả chi tiết về hướng tiếp cận bài toán. Sau khi đã chuẩn bị dữ liệu, Phần 5 tập trung vào quá trình thực nghiệm và phân tích kết quả của các phương pháp recommendation system trên nhiều tập dữ liệu khác nhau. Cuối cùng, trong Phần 6, chúng tôi đưa ra kết luận và tổng kết cho toàn bộ nghiên cứu.

2 Các công trình liên quan

Trong phần này, chúng tôi sẽ tiến hành giới thiệu một số công trình liên quan đến hệ thống khuyến nghị du lịch ở 2.1 và một số phương pháp cũng như mô hình liên quan ở 2.2.

2.1 Công trình liên quan đến hệ thống khuyến nghị du lịch

Trong nghiên cứu của Van Canneyt và đồng nghiệp (2012) [24], họ đề xuất một chiến lược sử dụng mạng xã hội để khám phá các địa điểm độc đáo.

Phương pháp này kết hợp dữ liệu từ mạng xã hội được gắn kết địa lý để bổ sung cho cơ sở dữ liệu địa điểm hiện tại. Nghiên cứu của Yin et al. (2016) giới thiệu một đặc điểm độc đáo với hệ số xếp hạng dựa trên ba yếu tố: xếp hạng không gian cho các mục không gian và ngược lại. Họ triển khai các mô hình như LALDA và ULA-LDA để dự đoán vị trí của người dùng. Một góc nhìn mới là tích hợp thông tin từ geotagging để cá nhân hóa gợi ý [14].

Tác giả của tài liệu tham khảo [14] tập trung vào việc sử dụng địa điểm du lịch tích hợp với hoạt động thời gian thực của người dùng từ mạng xã hội. Phương pháp này cung cấp một phương tiện để tạo ra cơ sở dữ liệu đầy đủ về các địa điểm đáng quan tâm. Martinkus và Madiraju (2013) mô tả cách dữ liệu từ Twitter có thể được sử dụng để cá nhân hóa gợi ý địa điểm độc đáo thông qua việc phân tích văn bản tweet và siêu dữ liệu. Thông tin này được sử dụng để đề xuất các địa điểm đáng quan tâm khi người dùng tìm kiếm trong hệ thống. Trong tài liệu tham khảo [15], mô hình được mở rộng bằng cách sử dụng các đặc trưng tweet bổ sung như số lượng URL, số lượng phương tiện, sở thích của bạn bè và người theo dõi để cải thiện chất lượng gợi ý.

Mô hình đề xuất sử dụng phương pháp lọc cộng tác và lọc dựa trên nội dung từ tweet của người dùng và bạn bè để đánh giá điểm số cho các loại địa điểm đáng quan tâm. Nó cũng sử dụng Dịch vụ Google để thu thập danh sách các địa điểm thăm trong một thành phố cụ thể. Mô hình này giải quyết các vấn đề liên quan đến niềm tin bằng cách chỉ cung cấp danh sách mà không đưa ra quyết định thay mặt người dùng. Điểm độc đáo của mô hình là khả năng điều chỉnh địa điểm quan tâm của người dùng dựa trên cả các lựa chọn di chuyển hiện tại và ổn định của họ liên quan đến các địa điểm du lịch.

2.2 Những phương pháp và mô hình liên quan

Trong tuyển tập "Machine Learning in Recommender Systems" của Dietmar Jannach và đồng nghiệp (2010) [9], các số liệu thống kê nhấn mạnh sức mạnh của machine learning trong lĩnh vực hệ thống khuyến nghị. Việc áp dụng machine learning đã mang lại những tiến triển quan trọng, đặc biệt là trong các mô hình như collaborative filtering, content-based recommendation, và hybrid recommendation. Dữ liệu thống kê chi tiết từ tuyển tập này làm rõ ảnh hưởng tích cực của machine learning đối với hiệu suất và linh hoạt của các hệ thống khuyến nghị. Mô hình "Neural Collaborative Filtering" (Xiangnan He et al., 2017) [8] đã đạt được

kết quả ẩn tượng trong việc kết hợp mạng nơ-ron và collaborative filtering. Sự kết hợp này không chỉ nâng cao độ chính xác của đề xuất mà còn giải quyết một cách hiệu quả vấn đề khám phá, nổi lên từ collaborative filtering. Dữ liệu thống kê chi tiết từ nghiên cứu này cho thấy sự xuất sắc của mô hình trong nhiều bộ dữ liệu thử nghiệm, đặc biệt là trong việc cải thiện đáng kể trải nghiệm người dùng.

Phương pháp lọc dựa trên nội dung đề xuất các mục cho người dùng dựa trên sự so sánh giữa nội dung của các mục hoặc thực thể và nội dung được người dùng cung cấp trên hồ sơ cá nhân của họ [21]. Ví dụ, một hồ sơ phim có thể bao gồm các yếu tố khác nhau như thể loại, các diễn viên tham gia, doanh thu tại phòng vé, và nhiều hơn nữa. Các hồ sơ được tạo ra sẽ cho phép hệ thống kết nối người dùng với các sản phẩm phù hợp. Tuy nhiên, chiến lược dựa trên nội dung đòi hỏi thu thập thông tin cá nhân từ người dùng, điều này có thể không có sẵn hoặc khó thu thập.

Lọc cộng tác đánh giá các mối quan hệ khác nhau giữa người dùng và sự tương phản giữa các sản phẩm để xác định các mối quan hệ mới giữa user và item [27]. Lọc cộng tác có thể được thực hiện dưới dạng Lọc Cộng tác Dựa trên User hoặc Lọc Cộng tác Dựa trên Item; trong trường hợp này, các mục đề cập đến các bộ phim. Trong một hệ thống khuyến nghị dựa trên User, các gợi ý được tạo ra dựa trên xếp hạng của các nhóm người hoặc cụm người dùng [2]. Nó định vị các người dùng khác nhau có sở thích xem phim tương tự với người dùng hiện tại và tạo ra gợi ý cho người dùng dựa trên xếp hạng được cung cấp bởi những người đó.

Hệ thống khuyến nghị dựa trên lọc cộng tác gấp ván đè lớn khi nó chỉ đề xuất các bộ phim dựa trên một tiêu chí duy nhất, tức là xếp hạng được cung cấp bởi mọi người. Tuy nhiên, chỉ xếp hạng từ người dùng có thể không đủ để cung cấp cái nhìn toàn diện về sở thích thực sự của người dùng. Với sự phổ biến ngày càng tăng của Internet, người dùng đã trở nên thoải mái hơn trong việc tỏ ra bản thân và chia sẻ ý kiến của họ trên Internet bằng văn bản. Những đánh giá của người dùng có khả năng cung cấp thông tin chi tiết và nhất quán hơn về sở thích của người dùng. Nói cách khác, đánh giá văn bản của người dùng có thể được sử dụng để tạo ra xếp hạng về các đặc điểm của một bộ phim như Đạo diễn, Cốt truyện/ Kịch bản, Quay phim, Chỉnh sửa, Diễn xuất, Thiết kế sản xuất, Âm thanh và nhiều tính năng khác, kết hợp với xếp hạng số liệu, để tạo ra một quá trình khuyến nghị hiệu quả

hơn [16].

3 Dữ liệu

3.1 Thu Thập Dữ Liệu

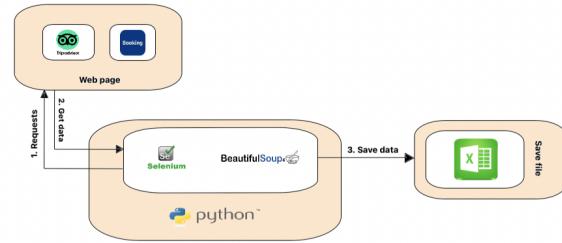


Figure 1: Mô tả quy trình thu thập dữ liệu cho hệ khuyến nghị.

Có 3 bộ dữ liệu được sử dụng trong báo cáo này có tên là Hotels Dataset, Restaurants Dataset và Attractions Dataset được chúng tôi thu thập từ trang web du lịch trực tuyến là Booking.com và TripAdvisor với quy trình được thể hiện ở hình 1. Sử dụng ngôn ngữ lập trình Python kết hợp với hai framework được hỗ trợ mạnh mẽ cho việc cào dữ liệu là Selenium và BeautifulSoup để thu thập thông tin về khách sạn và người dùng.

Dữ liệu thu thập từ tháng 10/2023 đến tháng 12/2023 đã được cung cấp và áp dụng đặc biệt cho mục đích nghiên cứu và học tập trong lĩnh vực Hệ Khuyến Nghị. Hình 2 minh họa một số điểm dữ liệu đại diện trong bộ dữ liệu này.

	Khách sạn	Nhà hàng	Điểm thu hút
Tỉnh thành	Gồm 13 tỉnh thành trên khắp Đất nước Việt Nam.	Gồm 11 tỉnh thành.	Gồm 6 tỉnh thành.
Số lượng mẫu	8475 mẫu.	1251 mẫu.	671 mẫu.
Thuộc tính	72 thuộc tính.	25 thuộc tính.	38 thuộc tính.

Figure 2: Thông kê bảng số liệu thu thập được.

Chúng tôi đã triển khai quá trình xây dựng bộ dữ liệu đa dạng cho hệ thống khuyến nghị của mình được thu thập từ nhiều tỉnh/ thành trên khắp Đất nước Việt Nam như: Côn Đảo, Đà Lạt, Hồ Chí Minh, Hà Nội, Vũng Tàu,... bao gồm 47 bộ dữ liệu cho khách sạn với tổng cộng 8,475 mẫu phục vụ cho 13 tỉnh/thành, mỗi bộ dữ liệu có 72 thuộc tính. Đối với nhà hàng, chúng tôi tạo ra 7 bộ dữ liệu với 1,251 mẫu cho 11 tỉnh/thành, mỗi bộ có 25 thuộc tính. Cuối cùng, với điểm thu hút, chúng tôi đã tạo 32 bộ dữ liệu với 671 mẫu cho 6 tỉnh/thành, mỗi bộ có 38 thuộc tính. Tổng cộng, có 86 bộ dữ liệu được

tạo ra, đảm bảo một quy trình nghiêm túc để đảm bảo chất lượng và hiệu suất của hệ thống.

STT	Tên thuộc tính	Thông tin	Kiểu dữ liệu
0	Hotel Name	8475	object
1	Image	8475 non-null	object
2	URL Hotel	8475 non-null	object
3	Hotel Location	8475 non-null	object
4	Show on map	8475 non-null	int64
5	Distance from centre (km)	8475 non-null	float64
6	Travel Sustainable Level	8475 non-null	int64
7	Ratings	8475 non-null	float64
8	Count Reviews	8475 non-null	int64
9	Address	8475 non-null	object
10	Details	8475 non-null	object
11	Services	8475 non-null	object
...
22	Facilities	8475	float64

Figure 3: Mô tả cho Hotels Dataset.

STT	Tên thuộc tính	Thông tin	Kiểu dữ liệu
1	Restaurant Name	1251 non-null	object
2	Count Reviews	1251 non-null	object
3	Address	1251 non-null	object
4	Mention	1251 non-null	object
5	Detail	1251 non-null	object
6	Country	1251 non-null	object
7	Time Operation	1251 non-null	object
8	Ratings	1251 non-null	float64
9	Food	1251 non-null	float64
10	Service	1251 non-null	float64
11	Value for money	1251 non-null	float64
12	Atmosphere	1251 non-null	float64
13	Price	1251 non-null	float64
14	Food1	1251 non-null	object

Figure 4: Mô tả cho Restaurants Dataset.

STT	Tên thuộc tính	Thông tin	Kiểu dữ liệu
0	Attraction Name	671	object
1	Title URL	671	object
2	Image URL	671	object
3	Location	671	object
4	Ratings	671	float64
5	Count Reviews	37	float64
6	Service1	512	object
7	Service2	512	object
8	Satisfaction	671	object
9	Detail	671	object
10	Du thuyền 45 phút	671	int64
11	Nhạc sống	671	int64
12	Nước đóng chai	671	int64
...

STT	Tên thuộc tính	Thông tin	Kiểu dữ liệu
23	Lớp học nấu ăn	671	int64
24	Đầu bếp nổi tiếng	671	int64
25	Dịch vụ hướng dẫn bằng tiếng Anh	671	int64
26	Cà phê	671	int64
27	Thú vị đặc biệt	671	int64
28	Hàn chè	37	object
29	Điểm tập trung	270	object
30	Thời lượng	670	object
31	Price	671	int64
32	Extra Cost	671	int64
...
33	User Name	671	object
34	User ID	671	int64
35	Attraction ID	671	int64

Figure 5: Mô tả cho Attractions Dataset.

Đối với các rating trong lĩnh vực khách sạn và nhà hàng, chúng được phân bố từ 1 đến 10, như thể hiện trong Hình 6. Trong khi đó, đối với bộ dữ liệu điểm thu hút, phân phối rating dao động từ 1 đến 5 và không đồng đều. Trong phạm vi bộ dữ liệu, rating thấp nhất là 1, với số lượng ít nhất (hơn 600 lượt đánh giá), trong khi rating cao nhất là 10, với gần 1,600 lượt. Các rating từ 2 đến 9 phân bố khá đồng đều, với số lượng đánh giá dao động từ 650 đến 900. Đáng chú ý, rating có giá trị 7 chiếm đa số với gần 2,500 lượt, tạo nên một phân phối đa dạng và cân bằng trong bộ dữ liệu. Quy trình xây dựng các bộ dữ liệu cho hệ khuyến nghị du lịch bao gồm các giai đoạn chính như: Thu thập dữ liệu (3.1) sẽ gồm các bước như cào dữ liệu từ website, lựa chọn và lọc lại dữ liệu. Kế tiếp là Xử lý dữ liệu (3.2) gồm các bước như loại bỏ các mẫu dữ liệu trùng lặp, xử lý chuỗi, khoảng trắng thừa và phân tích lỗi

dữ liệu để nhằm tạo ra những bộ dữ liệu đa dạng và đầy đủ thông tin. Chi tiết về quy trình thu thập dữ liệu được trình bày ở Hình 1. Mục tiêu là giúp hệ thống khuyến nghị có khả năng cung cấp các gợi ý chính xác và hấp dẫn. Thông tin chi tiết về các thuộc tính của bộ dữ liệu được thể hiện trong Bảng 3, 4, 5.

3.2 Xử Lý Dữ Liệu

Chất lượng của bộ dữ liệu đóng một vai trò quan trọng trong quá trình tạo ra các mô hình dự đoán có độ chính xác cao và khả năng tổng quát hóa. Để đáp ứng yêu cầu của từng nhiệm vụ cụ thể, chúng tôi tiếp cận với việc xử lý dữ liệu thông qua các phương pháp đa dạng với mẫu dữ liệu của các bộ dữ liệu khách sạn, nhà hàng và địa điểm thu hút sau khi được làm sạch được thể hiện ở bảng 1, 2, 3, 4, 5, 6. Tuy nhiên, các bộ dữ liệu như Hotels, Restaurants, Attractions Dataset của chúng tôi phải đổi mới với một thách thức lớn đó là sự thưa thớt của dữ liệu. Số lượng người dùng để lại đánh giá cho khách sạn, nhà hàng và điểm thu hút không nhiều, và rất nhiều người dùng chỉ đánh giá cho một vài khách sạn trong bộ dữ liệu. Điều này gây ra sự thưa thớt trong bộ dữ liệu của chúng tôi, làm tăng khó khăn trong việc đạt được kết quả cao khi triển khai hệ thống khuyến nghị.

Trong báo cáo này, chúng tôi đã tiến hành xây dựng hệ thống khuyến nghị bằng hai phương pháp chính: Lọc Cộng Tác (Collaborative Filtering) và Lọc Dựa Trên Nội Dung (Content-based Filtering). Trong Lọc Cộng Tác, chúng tôi đã triển khai hai chiến lược: Lọc Cộng Tác Dựa Trên Người Dùng và Lọc Cộng Tác Dựa Trên Sản Phẩm. Ngoài ra, chúng tôi đã đề xuất giải thuật Hybrid cho hệ thống khuyến nghị và sử dụng phương pháp Model-based để dự đoán đánh giá của người dùng đối với khách sạn, nhà hàng và điểm thu hút.

3.3 Bộ Dữ Liệu GroundTruth

Chúng tôi đã tạo bộ dữ liệu groundtruth cho khách sạn, nhà hàng và điểm thu hút từ Booking.com và TripAdvisor. Thông tin thu thập bao gồm tên, địa chỉ, mô tả, đánh giá và các thuộc tính khác của đối tượng, được mô tả ở hình 7, 8, 9. Sau đó, chúng tôi tiến hành tiền xử lý để loại bỏ các bản ghi trùng lặp và chuẩn hóa định dạng. Bộ dữ liệu groundtruth bây giờ không chỉ chứa các lượt ratings từ người dùng mà còn bao gồm thứ tự ưu tiên của họ đối với từng đối tượng. Đối với user-based, chúng tôi chọn những user có từ 15 lần rating trở lên, tổng cộng 7,759 user thỏa mãn và sau đó chúng tôi sẽ

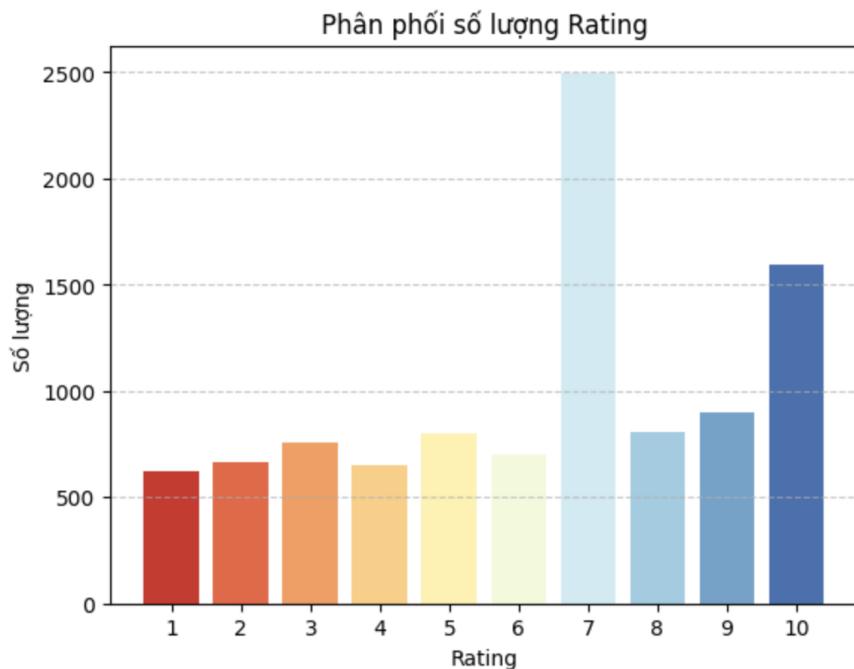


Figure 6: Phân phối số lượng Rating.

Index	Hotel Name	Recommendations
0	Khách sạn Thái Bình	[Khách San Minh Thắng', 'Changed Hotel 5', 'SixPearl Hotel', 'Khách sạn Trọng Tin', 'Changed Hotel 31', 'Golf Star Hotel', 'Changed Hotel 73', 'Đường Thành Bình 2', 'Hàng Ôn Côn Đảo Hotel', 'Đường Thành Bình']
1	Hotel De Condor	[O Songchi Hotel', 'An Phat Hotel', 'Changed Hotel 46', 'Changed Hotel 52', 'Havana Con Dao Hotel', 'Kim Ngan Hotel', 'Thien Tan Hotel', 'Con Son Blue Sea Hotel', 'Changed Hotel 36', 'SaConDor Hotel']
2	The Mystery Con Dao	[Changed Hotel 51', 'Con Son Blue Sea Hotel', 'Maya Hotel 3', 'Changed Hotel 14', 'SeaConDor Hotel', 'Maya Hotel 2', 'Hàng Ôn Côn Đảo Hotel', 'Changed Hotel 71', 'Tuần Ninh Hotel I', 'Thien Tan Hotel']
3	Uyen's House	[SixPearl Hotel', 'Homestay Hoàng Long', 'KHÁCH SẠN HOÀNG SƠN', 'Nhà nghỉ Hạnh Minh', 'Bà Đoàn 2 Hotel', 'Changed Hotel 66', 'Ngô Anh Hotel', 'Tuần Ninh Hotel I', 'Changed Hotel 4', 'Changed Hotel 39']
4	Kim Ngân II hotel	[Kim Ngân Hotel', 'KIM NGÂN HOME', 'Changed Hotel 73', 'THAO LIEN HOTEL', 'SixPearl Hotel', 'KHÁCH SẠN HOÀNG SƠN', 'KHANG HOTEL CON DAO', 'Havana Con Dao Hotel', 'Changed Hotel 21', 'Changed Hotel 19']

Figure 7: Mô tả cho Ground Truth Hotels Dataset.

Figure 8: Mô tả cho Ground Truth Restaurants Dataset.

Figure 9: Mô tả cho Ground Truth Attractions Dataset.

lấy đánh giá cũng như thứ tự ưu tiên của họ. Đối với item-based, chúng tôi chọn một số đối tượng có từ 15 lần rating và tương tự sẽ lấy đánh giá cũng như thứ tự ưu tiên từ người dùng về chúng. Đối với content-based, chúng tôi sử dụng mô tả và các

thuộc tính để tạo ra các đặc trưng cùng với thứ tự ưu tiên. Cuối cùng, chúng tôi chia bộ dữ liệu thành tập huấn luyện và tập kiểm thử để đào tạo và đánh giá mô hình khuyến nghị của mình, đảm bảo rằng cả đánh giá và thứ tự đều được xem xét trong quá trình đánh giá hiệu suất.

3.4 Bộ Dữ Liệu Cho Lọc Dựa Trên Nội Dung

Để tạo ra bộ dữ liệu cho lọc dựa trên nội dung, đầu tiên chúng tôi tiến hành đếm số lần rating của từng user để chọn ra các user có từ 15 lần rating trở lên, kết quả thu được 6,759 user thỏa mãn. Tiếp theo, chúng tôi lọc ra các user đó từ bộ dữ liệu Hotels, Restaurants, Attractions Dataset để tạo ra tập dữ liệu mới. Từ tập dữ liệu này, với từng user khác nhau sẽ có từ 15 lượt rating trở lên, chúng tôi sẽ lấy 5 lượt rating để làm tập testing. Tập training là các dữ liệu còn lại trong bộ dữ liệu Hotels, Restaurants, Attractions Dataset đã được loại bỏ đi các dữ liệu nằm trong testing. Kết quả thu được hai tập testing và training với số lượng dữ liệu lần lượt là 2,875 dòng dữ liệu và 5,884 dòng dữ liệu. Đối với tập training chúng tôi tiếp tục xử lý dữ liệu như sau: Mỗi khách sạn, nhà hàng hoặc điểm thu hút sẽ gồm nhiều mô tả như về vị trí địa lý, cơ sở hạ tầng, nội thất, đồ ăn thức uống và một số dịch vụ khác. Các mô tả này trên trang website Booking.com và TripAdvisor cơ bản sẽ nằm ở những mục khác nhau, do đó chúng tôi sẽ tiến hành gộp tất cả mô tả đã thu

Table 1: Mẫu dữ liệu từ bộ dữ liệu khách sạn

Hotel ID	Hotel Name	Services
1	Khách sạn Thái Bình	Airport shuttle Non-smoking Free WiFi Family rooms Good breakfast
2	Hotel De Condor	Non-smoking rooms Free WiFi Family rooms Bar
3	The Mystery Con Dao	Outdoor swimming pool Non-smoking rooms Free parking Restaurant Free WiFi
4	Uyen's House	Free parking Free WiFi Bar
5	Kim Ngân II hotel	Non-smoking rooms Free parking Family rooms Bar

Table 2: Mẫu dữ liệu từ bộ dữ liệu đánh giá khách sạn

User ID	Hotel ID	Rating
1	72	9
1	393	5
1	652	10
1	4750	7
1	6680	7

thập được ở các mục khác nhau của các khách sạn, nhà hàng hoặc điểm thu hút mà mỗi user đã đánh giá thành một mô tả cho user, từ đó thu được 7,889 điểm dữ liệu (tương ứng với 7,889 users); tiếp theo, lấy từng descriptions của mỗi khách sạn, chúng tôi thu được tất cả 10,205 điểm dữ liệu (tương ứng với 10,205 khách sạn).

3.5 Bộ Dữ Liệu Cho Lọc Cộng Tác Dựa Trên User

Để tạo ra bộ dữ liệu cho lọc cộng tác dựa trên user, đầu tiên chúng tôi tiến hành đếm số lần rating của từng user để chọn ra các user có từ 15 lần rating trở lên, kết quả thu được 5,701 user thỏa mãn. Tiếp theo, chúng tôi lọc ra các user đó từ bộ dữ liệu Hotels, Restaurant, Attractions Dataset để tạo ra tập dữ liệu mới. Từ tập dữ liệu này, với từng khách sạn, nhà hàng và điểm thu hút khác nhau, chúng tôi sẽ chia ra với tỉ lệ 20% dữ liệu cho tập testing. Tập training là 80% dữ liệu còn lại trong bộ dữ liệu Hotels, Restaurants, Attractions Dataset. Kết quả thu được hai tập testing và training với số lượng dữ liệu lần lượt là 1,264 dòng dữ liệu và 5054 dòng dữ liệu.

Kết quả thu được hai tập testing và training với số lượng dữ liệu lần lượt là 1,701 dòng dữ liệu và 12,507 dòng dữ liệu.

3.6 Bộ Dữ Liệu Cho Lọc Cộng Tác Dựa Trên Item

Để tạo ra bộ dữ liệu cho lọc cộng tác dựa trên item, đầu tiên chúng tôi tiến hành đếm số lượng rating của từng khách sạn, nhà hàng và điểm thu hút để chọn ra các khách sạn có từ 20 user rating trở lên, kết quả thu được 4,916 khách sạn, 867 nhà hàng và 535 địa điểm thu hút thỏa mãn. Tiếp theo, chúng tôi lọc ra các khách sạn, nhà hàng và điểm thu hút đó từ bộ dữ liệu Hotels, Restaurants, Attractions Dataset để tạo ra tập dữ liệu mới. Từ tập dữ liệu này, với từng khách sạn, nhà hàng và điểm thu hút khác nhau, chúng tôi sẽ chia ra với tỉ lệ 20% dữ liệu cho tập testing. Tập training là 80% dữ liệu còn lại trong bộ dữ liệu Hotels, Restaurants, Attractions Dataset. Kết quả thu được hai tập testing và training với số lượng dữ liệu lần lượt là 1,264 dòng dữ liệu và 5054 dòng dữ liệu.

3.7 Bộ Dữ Liệu Cho Hybrid Filtering

Để xây dựng bộ dữ liệu cho phương pháp lọc hybrid, chúng tôi kết hợp thông tin từ cả lọc dựa trên nội dung và lọc cộng tác dựa trên user. Quá trình này bao gồm các bước chi tiết như sau:

Lọc Dựa Trên Nội Dung (Content-Based): Chúng tôi bắt đầu bằng việc chọn những người dùng có đủ lượng đánh giá, đạt từ 15 lượt đánh giá trở lên. Kết quả là 6,759 người dùng được chọn. Từ

Table 3: Mẫu dữ liệu từ bộ dữ liệu nhà hàng

Restaurant ID	Restaurant Name	Food
1	Hoang's Vietnamese Restaurant - Vegan Food	Vietnamese, Healthy, Vegetarian Friendly, Vegan Options, Gluten Free Options
2	Vy's Market Restaurant	Vegetarian Friendly, Vegan Options, Gluten Free Options, Lunch, Dinner, Brunch, Late Night
3	Ho Lo Quan	Asian, Vietnamese, Vegetarian Friendly, Vegan Options, Gluten Free Options
4	Red Bean Restaurant	Vietnamese, Vegetarian Friendly, Vegan Options, Gluten Free Options
5	Nhu Bau Restaurant (Family Kitchen)	Vegetarian Friendly, Vegan Options, Gluten Free Options, Lunch, Dinner, Brunch, Late Night

Table 4: Mẫu dữ liệu từ bộ dữ liệu đánh giá nhà hàng

User ID	Restaurant ID	Rating
1	347	8
1	892	3
1	121	9
1	573	5
1	1049	7

nhóm người dùng này, chúng tôi tạo tập dữ liệu dựa trên nội dung bằng cách lấy mô tả của khách sạn, nhà hàng và điểm thu hút mà họ đã đánh giá. Mỗi người dùng được chọn ngẫu nhiên 5 lượt đánh giá làm tập testing, và dữ liệu còn lại làm tập training. Kết quả là hai tập testing và training với số lượng dữ liệu lần lượt là 2,875 dòng và 7,889 dòng.

Lọc Cộng Tác Dựa Trên User (User-Based Collaborative Filtering): Tiếp tục sử dụng nhóm người dùng có từ 15 lượt đánh giá trở lên (6,759 người dùng). Dựa trên nhóm người dùng này, chúng tôi chọn ra 5 dòng dữ liệu cho mỗi người dùng để tạo tập testing. Tập training là dữ liệu còn lại không nằm trong tập testing. Kết quả là hai tập testing và training với tổng số dữ liệu là 1,701 dòng và 12,507 dòng.

Kết Hợp Dữ Liệu: Cuối cùng, chúng tôi kết hợp dữ liệu từ cả hai phương pháp trên để tạo ra bộ dữ liệu cho phương pháp lọc hybrid. Điều này giúp cung cấp một cơ sở dữ liệu đa dạng và phong phú, sẵn sàng cho việc đào tạo và đánh giá mô hình lọc hybrid của chúng tôi. Quá quy trình này, chúng tôi đã tạo ra một bộ dữ liệu đầy đủ và đa chiều, tích hợp thông tin từ cả hai phương pháp để đảm bảo mô hình lọc hybrid có khả năng đưa ra các khuyến nghị chính xác và đa dạng.

3.8 Bộ Dữ Liệu Cho Model-based Collaborative Filtering

Để xây dựng bộ dữ liệu cho phương pháp model-based, chúng ta cần thu thập thông tin về các thuộc tính quan trọng của user và item. Đầu tiên, lựa chọn các thuộc tính có thể ảnh hưởng đến sở thích hoặc đánh giá của người dùng đối với một item cụ thể, chẳng hạn như vị trí, tiện nghi, giá cả, hoặc đánh giá của người dùng trước đó. Dữ liệu này sau đó được sử dụng để xây dựng ma trận user-item, trong đó mỗi dòng đại diện cho một người dùng và mỗi cột đại diện cho một item.

Trước khi tiến hành xây dựng mô hình, cần xử lý dữ liệu thiếu, chúng tôi sử dụng các phương pháp như điền giá trị trung bình hoặc dự đoán giá trị thiếu bằng mô hình. Tiếp theo, chúng tôi sẽ tiến hành phân chia tỉ lệ dữ liệu với 80% cho tập training bao gồm một số các thuộc tính quan trọng của khách sạn, nhà hàng, điểm thu hút và 20% cho tập testing bao gồm các ratings của các user cho các item. Kế tiếp, chúng tôi sẽ sử dụng các kỹ thuật machine learning, chẳng hạn như học máy, mô hình tuyến tính, hoặc mô hình học sâu để huấn luyện mô hình dự đoán sở thích hoặc đánh giá của người dùng dựa trên các thuộc tính đã chọn.

Sau khi xây dựng mô hình, ta tiến hành đánh giá hiệu suất của mô hình. Việc kiểm tra chất lượng của mô hình thì chúng tôi sử dụng các thang đo đánh giá như Mean Squared Error (MSE), R2 (Coefficient of Determination) trên tập testing. Quá trình này đòi hỏi sự tìm hiểu sâu về các thuật toán machine learning và xử lý dữ liệu để tạo ra một model-based collaborative filtering hiệu quả.

Table 5: Mẫu dữ liệu từ bộ dữ liệu địa điểm thu hút

Attraction ID	Attraction Name	Services
1	Du thuyền đêm trên sông Hàn	Đón và trả khách Dịch vụ hướng dẫn Phương tiện đi lại khứ hồi Vé vào các điểm tham quan Lớp học nấu ăn
2	Tour Bà Nà Hills trọn ngày với bữa trưa buffet	Góc selfie Buffet trưa tại nhà hàng Vé vào các điểm tham quan
3	Vé vào cửa công viên suối khoáng nóng Núi Thần Tài	Lớp học nấu ăn
4	Lớp hướng dẫn làm lồng đèn	Góc selfie Đi cáp treo Đón và trả khách Vé vào các điểm tham quan Lớp học nấu ăn
5	Tour Cầu Vàng và Bà Nà Hills	Máy hát karaoke Góc selfie Dịch vụ hướng dẫn Lớp học nấu ăn

Table 6: Mẫu dữ liệu từ bộ dữ liệu đánh giá địa điểm thu hút

User ID	Attraction ID	Rating
1	473	3
1	126	1
1	589	5
1	324	5
1	213	2

4 Thí nghiệm

4.1 Hướng Tiếp Cận

Từ bộ dữ liệu Hotels, Restaurants, Attractions Dataset được chúng tôi thu thập ban đầu, sau quá trình xử lý để tạo ra các tập training và testing để xây dựng một hệ khuyến nghị sử dụng các phương pháp content-based filtering (CB – lọc dựa trên nội dung) [21], collaborative filtering (CF – lọc cộng tác) [13], hybrid filtering [26], model-based collaborative filtering [11] (được trình bày rõ hơn trong Phần 4.3 và 4.4). Để đánh giá hệ khuyến nghị sử dụng phương pháp lọc nội dung, lọc cộng tác, hybrid, graph-based [22], chúng tôi sử dụng các độ đo: Mean Squared Error, Coefficient of Determination; đối với hệ khuyến nghị sử dụng phương pháp model-based, chúng tôi đánh giá thông qua độ đo Precision@10 và NDCG@10 (chi tiết trong Phần 4.6).

4.2 Giới thiệu và triển khai kiến trúc hệ thống khuyến nghị du lịch

Hệ thống kiến trúc được xây dựng với ba phần chính bao gồm việc thu thập và lưu trữ dữ liệu, hệ thống khuyến nghị (RS), và giao diện người dùng. Như minh họa tổng quát trong Hình 10, ba đơn vị

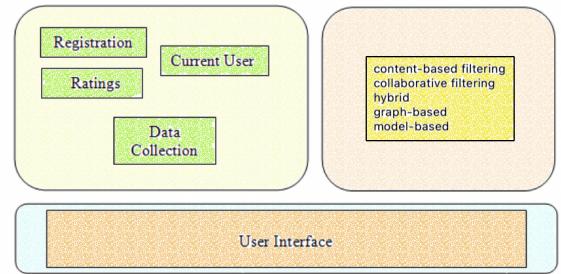


Figure 10: Tổng quan kiến trúc hệ thống khuyến nghị du lịch của chúng tôi.

thành phần này hoạt động một cách hợp tác.

Quá trình đăng ký (Registration) thu thập những thông tin mà người dùng lựa chọn hoặc nhập vào hệ thống, kèm theo đánh giá của họ về các khách sạn, nhà hàng, hoặc điểm thu hút. Các thông tin này được lưu trữ trong các bộ dữ liệu tương ứng, bao gồm Hotels, Restaurants, và Attractions Datasets. Trong kiến trúc của hệ thống khuyến nghị du lịch này, các bộ dữ liệu được lưu trữ ở phần bên trong.

Hệ thống khuyến nghị du lịch này sử dụng nhiều kỹ thuật như lọc dựa trên nội dung, lọc cộng tác, kết hợp cả hai (hybrid), lọc dựa trên cấu trúc đồ thị (graph-based), và mô hình dựa trên machine learning hoặc neural network. Cụ thể, hệ thống sử dụng các phép toán như cosine similarity, pearson similarity, và jaccard similarity để đo độ tương đồng giữa user và item trong hệ thống khuyến nghị du lịch.

Sở thích và đánh giá của người dùng về khách sạn, nhà hàng, hoặc điểm thu hút được chuyển vào phần bên trong để kết hợp với bộ dữ liệu lưu trữ và áp dụng các kỹ thuật lọc đã thiết lập sẵn trong hệ thống khuyến nghị du lịch. Kết quả sẽ đưa ra các gợi ý phù hợp với người dùng, và thông tin này

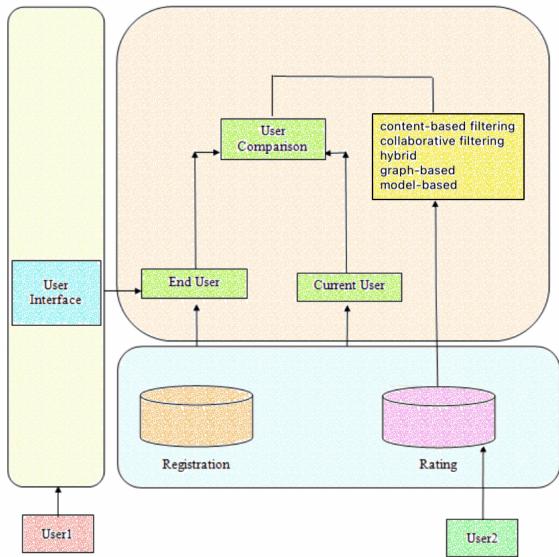


Figure 11: Triển khai kiến trúc lên hệ thống khuyến nghị du lịch.

được xuất ra thông qua giao diện người dùng. Chi tiết về kiến trúc hệ thống được mô tả ở hình 11.

4.3 Content-based Filtering

Trong hệ thống khuyến nghị dựa trên nội dung, chúng ta tận dụng các mô tả chi tiết của sản phẩm để tạo ra gợi ý [21]. Thuật ngữ "content" ở đây chỉ đơn giản là mô tả các đặc tính của sản phẩm. Trong phương pháp này, chúng ta kết hợp thông tin đánh giá và hành vi mua sắm của người dùng với nội dung chi tiết của sản phẩm [21].

Để minh họa, giả sử người dùng A đã đánh giá cao cho khách sạn X. Tuy nhiên, với hạn chế quyền truy cập vào đánh giá của người dùng khác, phương pháp lọc cộng tác truyền thống có thể không linh hoạt. Thay vào đó, chúng ta tập trung vào các mô tả chi tiết như vị trí, tiện ích, dịch vụ, diện tích, giá cả của khách sạn X.

Bằng cách này, chúng ta có thể tìm ra những sản phẩm tương tự, như khách sạn Y hoặc Z, dựa trên sự tương đồng trong mô tả nội dung. Điều này giúp giải quyết thách thức khi không thể tiếp cận đánh giá của người dùng khác trong phương pháp lọc cộng tác truyền thống.

Hình 12 thể hiện chi tiết phương pháp khuyến nghị dựa trên nội dung của chúng tôi. Với một user đã rating n khách sạn khác nhau, chúng tôi sẽ có được descriptions cho user là tổng hợp n descriptions của các khách sạn đó. Sau đó, chúng tôi tính toán độ tương đồng giữa descriptions của user và descriptions của các khách sạn thuộc địa

CONTENT-BASED FILTERING

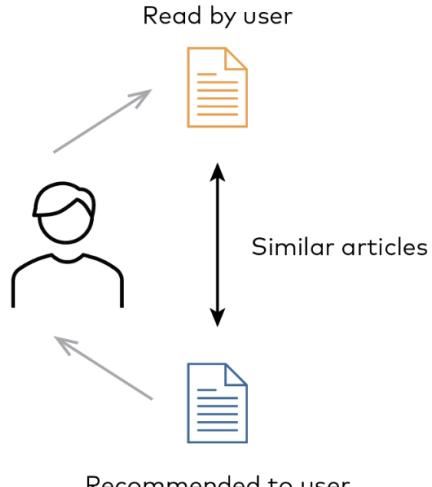


Figure 12: Content-based filtering.

điểm mà user đó muốn đi để khuyến nghị ra danh sách gồm m khách sạn có độ tương đồng cao nhất, tức là descriptions của m khách sạn gần giống nhất với descriptions của user. Làm tương tự cho nhà hàng và điểm thu hút. Chúng tôi sử dụng phương pháp TF-IDF (Term Frequency-Inverse Document Frequency) để tính toán trọng số độ tương đồng, đây là thước đo thường được sử dụng trong truy xuất thông tin và khai thác văn bản. TF-IDF có thể được tính toán như công thức 1:

$$TF-IDF(t, d) = tf(t, d) \times \log \frac{|D|}{|d : t \subseteq d|} \quad (1)$$

Trong đó: t là một từ, d là một văn bản, tf(t,d) là tần suất xuất hiện của từ t trong văn bản d, |D| là số lượng của tất cả các văn bản được quan sát.

4.4 Collaborative Filtering

Các mô hình lọc cộng tác (minh họa ở hình 13) tận dụng sức mạnh của cộng tác trong việc sử dụng đánh giá từ nhiều người dùng để đưa ra các đề xuất. Một trong những thách thức chính khi thiết kế các phương pháp lọc cộng tác là đối mặt với vấn đề sparse "thưa thớt" [4] của matrix rating "ma trận đánh giá" [28], nơi đa số các mục được đánh giá không đồng đều.

Ví dụ, xét một ứng dụng đặt phòng khách sạn, trong đó người dùng A thể hiện sở thích hoặc không thích đối với một số khách sạn cụ thể. Thường

COLLABORATIVE FILTERING

Algorithm 1 Lọc cộng tác với phép tính cosine similarity.

```

1: Initialization:
2: Parameters Initialization:
3:   CurrentUserID = 7
4: Data Collection:
5:   Hotels.csv
6: while ( $u < U_{\max}$ ) do
7:   for  $i = 1$  to  $MaxUserID$  do
8:     if  $iUserID = EndUserID$  then
9:       Copy HotelID, rating to
  endUser.csv
10:    end if
11:   end for
12:   Create endUserList
13:   for  $i = 1$  to  $MaxUserID$  do
14:     if  $iUserID = uUserID$  then
15:       Copy HotelID, rating to
  currentUser.csv
16:     end if
17:   end for
18:   Create currentUserList
19:   Compare endUserList and
  currentUserList create
  compareUserList
20: Compute Cosine Similarity:
21:   for  $w = 1$  to  $n$  do
22:     if numberCommonHotels > 0 then
23:
24:        $wi = \frac{\sum_i p_i \cdot q_i}{\sqrt{\sum_i p_i^2} \cdot \sqrt{\sum_i q_i^2}}$ 
25:
26:     end if
27:     Write uUserID and CosineSimilarity
  into GroundTruthHotels.csv
28:      $u = u + 1$ 
29:   end for
30: end while
31: Find SimiFeatures from Comparison.csv
32: Compute recommendedHotels and display

```

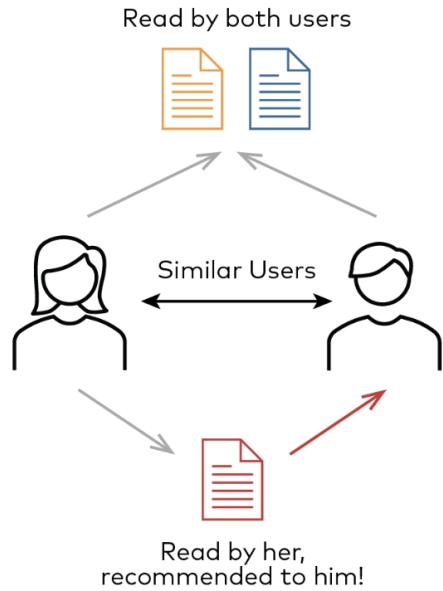


Figure 13: Collaborative filtering.

xuyên, người dùng chỉ đánh giá một phần nhỏ trong tổng số lượng lớn các khách sạn tại mỗi địa điểm. Kết quả là, đa số các đánh giá là không xác định, dẫn đến khả năng tính toán độ tương đồng không hiệu quả. Tình trạng tương tự cũng xảy ra với nhà hàng và điểm thu hút.

Để đánh giá độ tương đồng, chúng tôi sử dụng ba độ đo được thể hiện ở các công thức 2, 3, 4 là:

Độ đo tương đồng cosine (cosine similarity):

$$\text{sim}(U_u, U_v) = \cos(U_u, U_v) = \frac{\vec{U}_u \cdot \vec{U}_v}{\|\vec{U}_u\| \cdot \|\vec{U}_v\|} \quad (2)$$

Độ đo tương đồng pearson (pearson correlation):

$$\text{sim}(U_u, U_v) = \frac{\sum_i (r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_i (r_{u,i} - \bar{r}_u)^2} \sqrt{\sum_i (r_{v,i} - \bar{r}_v)^2}} \quad (3)$$

Độ đo tương đồng jaccard (jaccard similarity):

$$\text{sim}(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (4)$$

Trong đó: $r_{u,i}$ là giá trị rating của người dùng u đối với item i ; \bar{r}_u là giá trị rating trung bình của người dùng u .

Có hai loại phương pháp thường được sử dụng trong lọc cộng tác là user-based và item-based [2, 7]. Trong báo cáo này, chúng tôi sử dụng cả 2 phương pháp trên, ngoài ra chúng tôi còn triển khai thêm phương pháp model-based [11] và graph-based [22].

4.4.1 User-user collaborative filtering

Trong trường hợp này, rating được cung cấp bởi những người dùng có rating cho các item khác như người dùng mục tiêu A được sử dụng để đưa ra khuyến nghị cho A. Do đó, ý tưởng cơ bản là xác định những người dùng tương tự như người dùng mục tiêu A và đề xuất các item bằng cách tính toán độ tương đồng giữa người dùng A và các người dùng khác [2]. Ví dụ: nếu trước đây A và B đã rating các khách sạn theo cách tương tự; A đã rating cho khách sạn X và B chưa rating cho khách sạn X, thì người ta có thể sử dụng rating của A trên khách sạn X để dự đoán rating của B đối với khách sạn này. Nói chung, k người dùng tương tự nhất với B có thể được sử dụng để đưa ra dự đoán rating cho B. Để tính toán độ tương đồng giữa người dùng mục tiêu và những người dùng khác, ta có thể sử dụng cosine similarity (Công thức 2), pearson correlation (Công thức 3) hoặc jaccard similarity (Công thức 4). Minh họa về user-user collaborative filtering được thể hiện trong Hình 13.

4.4.2 Item-item collaborative filtering

Để đưa ra dự đoán rating cho item mục tiêu X bởi người dùng A, bước đầu tiên là xác định một tập S các item tương tự nhất với item mục tiêu X. Các rating trong tập item S, được chỉ định bởi A, được sử dụng để dự đoán liệu người dùng A có thích item X hay không. Do đó, rating của B trên các khách sạn tương tự như Y và Z có thể được sử dụng để dự đoán rating của B trên khách sạn T [7]. Để tính toán độ tương đồng giữa item mục tiêu và những item khác, ta có thể sử dụng cosine similarity (Công thức 2), pearson correlation (Công thức 3) hoặc jaccard similarity (Công thức 4).

4.4.3 Model-based collaborative filtering

Trong phương pháp lọc cộng tác dựa trên mô hình, chúng tôi sử dụng mô hình để dự đoán và đưa ra khuyến nghị cho người dùng. Các mô hình này thường được huấn luyện trên dữ liệu lịch sử rating để hiểu sâu hơn về mối quan hệ giữa người dùng và item [11].

Có nhiều phương pháp model-based, trong đó một số phổ biến bao gồm:

- **Neural Network (Deep Learning) [20]:** Sử dụng mạng neural để học các biểu diễn đặc trưng phức tạp của người dùng và item, có khả năng tự học và hiểu các mối quan hệ phức tạp.
- **Linear Regression [17]:** Sử dụng mô hình hồi quy tuyến tính để dự đoán các giá trị rating

dựa trên các biến đặc trưng.

- **Random Forest Regression [12]:** Sử dụng mô hình random forest để dự đoán rating, có khả năng xử lý nhiều và mô hình các mối quan hệ phi tuyến tính.

Phương pháp model-based [11] thường có khả năng xử lý hiệu suất cao và giải quyết vấn đề của ma trận thưa thoát. Tuy nhiên, chúng đòi hỏi lượng dữ liệu lớn và thời gian huấn luyện phức tạp.

4.4.4 Graph-based collaborative filtering

Trong phương pháp lọc cộng tác dựa trên đồ thị, chúng tôi chủ yếu tập trung vào mô hình User-Item Graph, một phương pháp mạnh mẽ để mô hình hóa mối quan hệ giữa user và item dựa trên cấu trúc đồ thị [22].

Đồ thị được xây dựng với các user và item là các đỉnh, và các cạnh biểu diễn mối quan hệ giữa chúng. Cụ thể, nếu một user đã rating hoặc tương tác với một item, một cạnh sẽ được thêm vào giữa đỉnh người dùng và đỉnh item, đại diện cho mối quan hệ giữa chúng.

Mô hình User-Item Graph có một số ưu điểm:

- **Mô hình Học Biểu diễn:** Đồ thị là một cách mạnh mẽ để học biểu diễn của user và item. Các thuật toán nhúng đồ thị có thể chuyển đổi user và item thành các vectơ đặc trưng trong không gian đa chiều, giúp capture mối quan hệ phức tạp và tiềm ẩn.
- **Tận Dụng Mối Quan Hệ:** Mô hình này tận dụng tất cả các mối quan hệ đã xảy ra giữa user và item. Việc này giúp cải thiện khả năng dự đoán và khuyến nghị.
- **Xử Lý Đánh Giá Thiếu:** Trong trường hợp có một số user chưa rating cho nhiều item hoặc có ít thông tin, đồ thị có thể giúp dự đoán mối quan hệ và tạo ra khuyến nghị chất lượng.

Phương pháp User-Item Graph cung cấp cơ hội mạnh mẽ để nắm bắt sâu sắc mối quan hệ người dùng và item, đặc biệt là khi dữ liệu thưa thoát và ít đánh giá.

4.5 Baseline Models and Settings

Để đánh giá khả năng của hệ thống khuyến nghị, chúng tôi sử dụng một loạt các mô hình từ cơ bản đến nâng cao, bao gồm cả mô hình dựa trên máy học và các mô hình học sâu. Ngoài các giải thuật đã đề cập trước đó (Content-based filtering, Collaborative filtering, hybrid, graph-based filtering),

chúng tôi mở rộng danh sách bằng cách giới thiệu thêm ba mô hình:

Mô Hình Hồi Quy Tuyến Tính (Linear Regression) [17]: Hồi quy tuyến tính là một mô hình truyền thống và đơn giản nhưng thường được sử dụng hiệu quả cho các nhiệm vụ hồi quy. Trong hệ thống khuyến nghị của chúng tôi, chúng tôi áp dụng hồi quy tuyến tính để dự đoán đánh giá của người dùng dựa trên các đặc điểm và thuộc tính liên quan đến các mục. Mô hình được huấn luyện bằng cách sử dụng phương pháp bình phương tối thiểu thông thường.

Mô Hình Random Forest Regression [12]: Random Forest là một phương pháp học tập tổ hợp sử dụng nhiều cây quyết định để đưa ra dự đoán. Trong ngữ cảnh của chúng tôi, chúng tôi áp dụng Random Forest Regression để nắm bắt các mối quan hệ phức tạp giữa tương tác người dùng-vật phẩm và các đặc điểm ngữ cảnh khác nhau. Mô hình này kết hợp các dự đoán từ nhiều cây quyết định để tăng cường độ chính xác và sức mạnh.

Mô Hình Mạng Nơ-ron (Neural Network) [20]: Mạng nơ-ron là một mô hình mạng thần kinh nhân tạo mạnh mẽ, có khả năng học và biểu diễn các mối quan hệ phức tạp trong dữ liệu. Trong hệ thống khuyến nghị của chúng tôi, chúng tôi triển khai một mô hình mạng nơ-ron để tự động học các biểu diễn ẩn của tương tác người dùng-vật phẩm và đặc điểm liên quan. Mô hình được xây dựng với các tầng ẩn và kích thước tầng được điều chỉnh để phù hợp với bài toán cụ thể của chúng tôi.

Tất cả các mô hình cơ bản, bao gồm Linear Regression [17], Random Forest Regression [12] và Neural Network [20], được triển khai và huấn luyện bằng sử dụng thư viện scikit-learn và TensorFlow. Các siêu tham số của mô hình được thiết lập như sau:

- Hồi Quy Tuyến Tính: learning_rate = 1e-5, max_iter = 100, Optimization Algorithm: Gradient Descent.
- Random Forest Regression: n_estimators = 100, random_state = 42, max_depth = 10, min_samples_split = 2, min_samples_leaf = 4, max_features = 'auto', bootstrap = True.
- Mạng Nơ-ron: learning_rate = 0.001, batch_size = 32, epochs = 20, hidden_layer_sizes = (100, 50), activation = 'relu', solver = 'adam'.

Các mô hình được huấn luyện trên các bộ dữ liệu phù hợp với các đặc điểm và thuộc tính liên quan.

Chúng tôi đánh giá hiệu suất của chúng bằng cách sử dụng các độ đo như Mean Squared Error (MSE) và Coefficient of Determination (R2) để đo lường độ chính xác của các đánh giá dự đoán so với các đánh giá thực tế do người dùng cung cấp. Các thử nghiệm được thực hiện trên môi trường tính toán và tài nguyên tương tự như đã mô tả ở trên.

4.6 Độ Đo Đánh Giá

Việc đánh giá hiệu suất của hệ thống khuyến nghị là một bước quan trọng để đảm bảo chất lượng và độ chính xác của các đề xuất. Trong phạm vi bài này, chúng tôi tập trung vào việc đánh giá hiệu suất thông qua các phương pháp không yêu cầu sự can thiệp trực tiếp của người dùng, tức là đánh giá không liên quan đến việc theo dõi thời gian thực của họ. Dưới đây là một số độ đo quan trọng được thể hiện qua bốn công thức 5, 6, 7 và 8 mà chúng tôi sử dụng để đánh giá chất lượng của hệ thống khuyến nghị:

- Mean Squared Error (MSE):

$$MSE = \frac{1}{n} \sum_{(u,i)} (r_{u,i} - \hat{r}_{u,i})^2 \quad (5)$$

- R-squared (R²):

$$R^2 = 1 - \frac{\sum_{(u,i)} (r_{u,i} - \hat{r}_{u,i})^2}{\sum_{(u,i)} (r_{u,i} - \bar{r}_u)^2} \quad (6)$$

- P@10:

$$P@10 = \frac{\text{Các dự đoán đúng trong top 10}}{10} \quad (7)$$

- NDCG@10 (Normalized Discounted Cumulative Gain):

$$NDCG@10 = \frac{1}{Z} \sum_{i=1}^{10} \frac{2^{rel_i} - 1}{\log_2(i+1)} \quad (8)$$

trong đó rel_i là độ quan trọng của item ở vị trí i trong danh sách đề xuất, và Z là hằng số chuẩn hóa để đảm bảo giá trị nằm trong khoảng [0, 1].

Trong các công thức trên:

- $r_{u,i}$ là giá trị rating thực tế của người dùng u đối với item i .
- $\hat{r}_{u,i}$ là giá trị rating dự đoán của hệ thống cho người dùng u đối với item i .

- \bar{r}_u là giá trị rating trung bình của người dùng u .
- n là số lượng điểm dữ liệu được sử dụng để đánh giá.

Đối với Precision@10 ở công thức 7, giả sử hệ thống khuyến nghị một danh sách S gồm k item cho người dùng u , và t là số lượng item trong S mà người dùng u đã tương tác, thì t được tính là số lượng dự đoán đúng. Công thức NDCG@10 ở công thức 8 đánh giá sự quan trọng của việc đưa ra các dự đoán chính xác ở những vị trí đầu trong danh sách đề xuất.

5 Phân tích kết quả

Sau quá trình xử lý để tạo ra các tập training và testing, chúng tôi xây dựng một hệ khuyến nghị sử dụng các phương pháp được trình bày trong Phần 4 và đánh giá kết quả thông qua các độ đo MSE, R2, Precision@10 và NDCG@10 để so sánh kết quả đạt được.

5.1 Collaborative Filtering

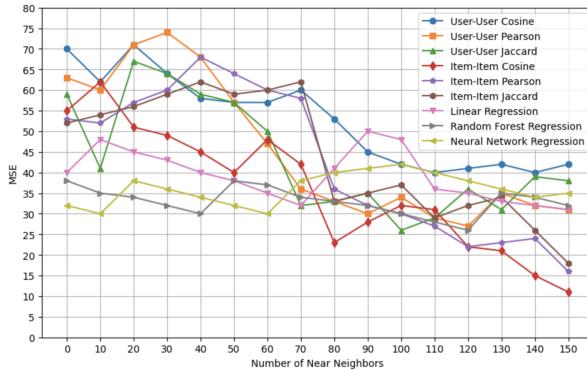


Figure 14: Thể hiện giá trị của độ đo MSE cho một số giải thuật collaborative filtering và model-based filtering.

Với hai phương pháp là user-user CF và item-item CF sử dụng ba độ tương đồng là cosine similarity, pearson similarity và jaccard similarity, qua đó chúng tôi có được sáu mô hình là user-user cosine, user-user pearson, user-user jaccard và item-item cosine, item-item pearson, item-item jaccard được đánh giá và so sánh thông qua các độ đo MSE, R2. Ngoài ra, chúng tôi còn tiến hành vẽ biểu đồ đường ở hình 14, 15, 17, 18 để so sánh giá trị về MSE, R2, Precision@10, NDCG@10 cho 3 mô hình là linear regression, random forest regression, neural network và các giải thuật lọc cộng tác như item-based, user-based. Bảng 7 trình bày toàn bộ

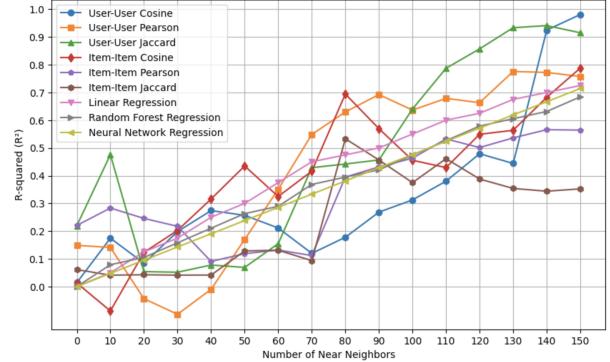


Figure 15: Thể hiện giá trị của độ đo R-2 cho một số giải thuật collaborative filtering và model-based filtering.

kết quả thực nghiệm của chúng tôi, ta có thể thấy rằng:

- **Về số lượng lân cận gần nhất:** Để giảm thời gian tính toán của các mô hình, chúng tôi lấy giới hạn các lân cận gần nhất là từ 10 đến 150 với bước nhảy là 10 cho các mô hình. Dựa vào kết quả của các độ đo, chúng ta có thể chọn ra số lượng lân cận gần nhất tốt nhất là 150, 50, 90, 10 tương ứng với các mô hình user-user cosine, user-user pearson, user-user jaccard và item-item cosine, item-item pearson, item-item jaccard.
- **Về mô hình và độ đo:** Với các lân cận gần nhất được chọn như trên, ta thu được các kết quả tốt nhất của các mô hình được thể hiện ở Bảng 7. Có thể thấy mô hình user-user cosine là mô hình cho kết quả tốt nhất trong số sáu mô hình được thực nghiệm ở tất cả các độ đo. Mô hình user-user cosine cho kết quả tốt nhất nhưng các thông số về độ đo vẫn còn khá cao, điều này có thể bị ảnh hưởng do rating matrix bị hiện tượng thưa thoát nhẹ. Ngược lại, mô hình item-item cho ra kết quả kém nhất. Nhìn chung, các mô hình khác cho ra các kết quả không đồng đều và bị biến động, nguyên nhân có thể là vì rating matrix bị hiện tượng thưa thoát nghiêm trọng và độ tương đồng pearson là không thực sự tốt trong trường hợp này.

5.2 Content-based Filtering

Để có thể đưa ra các khuyến nghị chính xác cho người dùng mà không tốn quá nhiều chi phí cũng như thời gian người dùng xem tất cả các đề xuất để lựa chọn khách sạn, nhà hàng hoặc điểm thu hút mà người dùng muốn đến trong lần tiếp theo thì chúng tôi giới hạn số lượng lân cận tương đồng từ

Table 7: Mô tả tổng quan kết quả thử nghiệm

Mô hình	Độ đo				Thời gian chạy (s)
	MSE	R2	Precision@10	NDCG@10	
Content-based cosine	0.15	0.81	0.80	0.88	1
Content-based pearson	0.18	0.85	0.72	0.8	2
Content-based jaccard	0.15	0.80	0.78	0.78	30
User-User cosine	0.09	0.93	0.88	0.92	1
User-User pearson	0.20	0.80	0.70	0.85	2
User-User jaccard	0.25	0.75	0.75	0.82	33
Item-Item cosine	0.38	0.82	0.79	0.86	1
Item-Item pearson	0.32	0.78	0.58	0.84	2
Item-Item jaccard	0.40	0.73	0.63	0.80	31
Hybrid	0.12	0.88	0.82	0.90	18
Linear Regression	0.28	0.87	0.60	0.70	4
Random Forest Regression	0.22	0.88	0.62	0.72	6
Neural Network	0.27	0.86	0.76	0.68	27
Graph-based	0.30	0.83	0.72	0.80	15

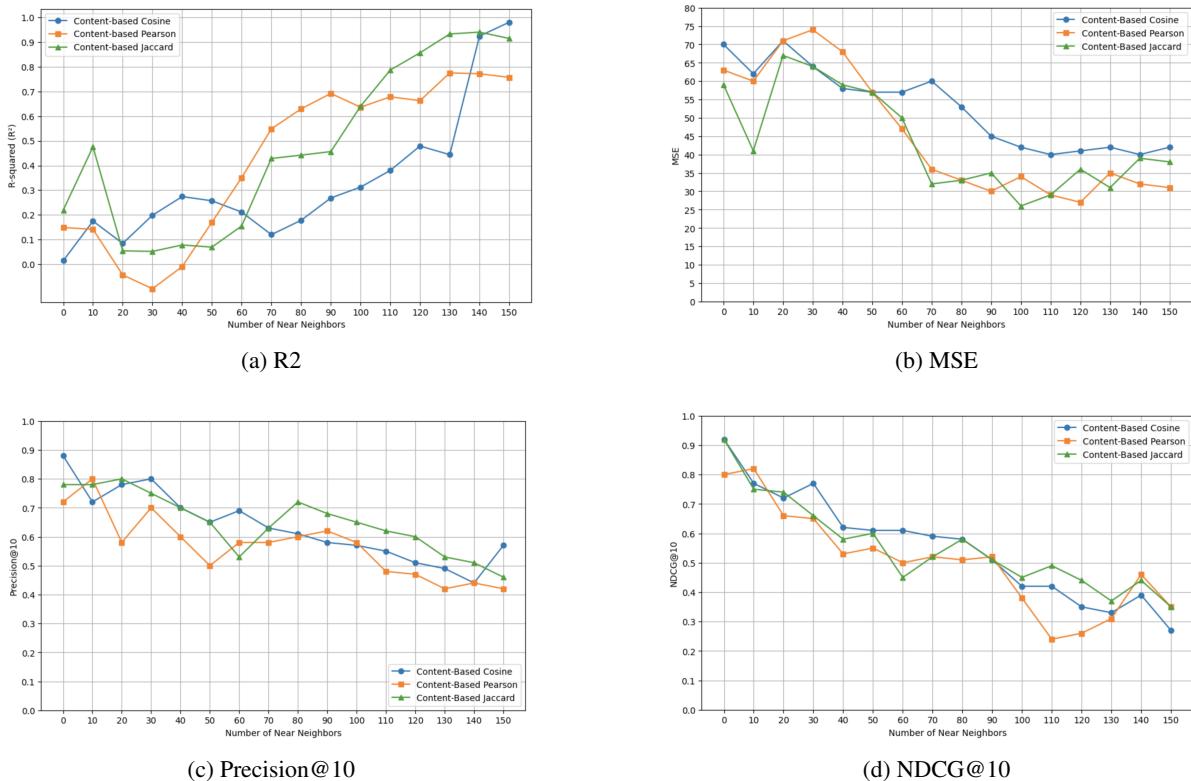


Figure 16: Kết quả của 4 độ đo gồm MSE, R2, Precision@10 và NDCG@10 cho giải thuật Content-based.

5-50 với các bước nhảy là 5 để tìm ra lân cận tốt nhất.

Hình 16 trình bày kết quả R2, MSE, Precision@10 và NDCG@10 theo số lượng khách sạn được khuyến nghị của content-based filtering, với số lượng 15 lân cận thì kết quả ở các độ đo Precision@10, NDCG@10 bắt đầu giảm. Vì vậy, chúng tôi chọn 15 là số lân cận tốt nhất dựa vào các tiêu chí đã đặt ra (giảm chi phí tính toán và giảm thời gian người dùng tìm kiếm khách sạn thích hợp trong danh sách các khách sạn được khuyến nghị) với MSE là 9%, R2 là 93%, Precision@10 là 88%, và NDCG@10 là 92% với thời gian tính toán khuyến nghị là 1s.

5.3 Hybrid Filtering

Hybrid Filtering là một phương pháp tích hợp cả Collaborative Filtering (CF) và Content-based Filtering (CBF) để tận dụng ưu điểm của cả hai mô hình. Trong nghiên cứu của chúng tôi, chúng tôi triển khai một mô hình Hybrid có thể kết hợp đồng thời độ tương đồng người dùng và nội dung của sản phẩm để cung cấp khuyến nghị chính xác và đa dạng.

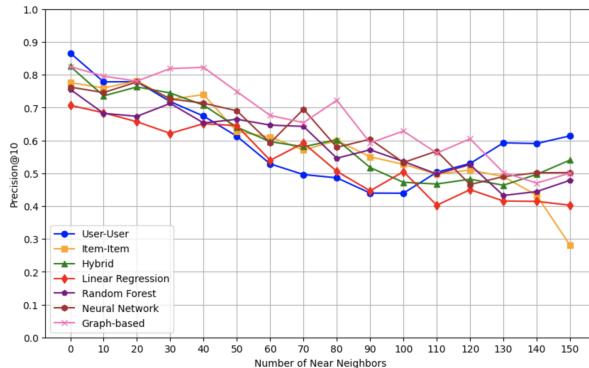


Figure 17: Thể hiện giá trị của độ đo Precision@10 cho một số giải thuật collaborative filtering và model-based filtering.

Đối với mô hình Hybrid Filtering [26], chúng tôi tính toán độ tương đồng người dùng bằng các phương pháp như cosine similarity, pearson similarity và jaccard similarity. Tương tự, đối với độ tương đồng nội dung, chúng tôi sử dụng các phương pháp đánh giá sự tương đồng giữa các sản phẩm dựa trên các đặc trưng của chúng.

Mô hình Hybrid Filtering [26] của chúng tôi tổng hợp thông tin từ cả hai phương pháp để đưa ra các đề xuất tối ưu cho người dùng. Qua đánh giá sử dụng các độ đo như MSE, R2, Precision@10 và NDCG@10, chúng tôi so sánh hiệu suất của

mô hình Hybrid với các mô hình Collaborative Filtering và Content-based Filtering riêng lẻ.

Hình 17 thể hiện sự so sánh giữa kết quả Precision@10 của mô hình Hybrid với các mô hình khác. Có thể thấy rằng mô hình Hybrid không chỉ giữ được những đặc tính tích cực của cả hai phương pháp mà còn cải thiện khả năng đề xuất so với việc sử dụng chúng độc lập.

Điều này làm nổi bật sức mạnh của việc kết hợp các phương pháp lọc cộng tác và lọc dựa trên nội dung để tối ưu hóa khả năng đề xuất trong hệ thống khuyến nghị của chúng tôi.

5.4 Graph-based Filtering

Trong nghiên cứu của chúng tôi, chúng tôi cũng đánh giá hiệu suất của mô hình Graph-based Filtering [22], một phương pháp mới sử dụng đồ thị để tối ưu hóa việc đề xuất. Đối với Graph-based Filtering [22], chúng tôi xây dựng đồ thị từ dữ liệu người dùng và sản phẩm, trong đó mỗi đỉnh đại diện cho một người dùng hoặc một sản phẩm và các cạnh biểu thị mức độ tương tác giữa chúng.

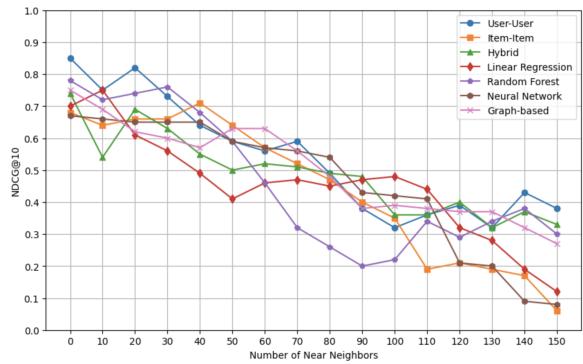


Figure 18: Thể hiện giá trị của độ đo NDCG@10 cho một số giải thuật collaborative filtering và model-based filtering.

Mô hình Graph-based Filtering [22] của chúng tôi sử dụng các thuật toán đồ thị như PageRank để đánh giá tầm quan trọng của mỗi đỉnh trong đồ thị. Dựa trên đánh giá này, chúng tôi tạo ra danh sách các sản phẩm được ưu tiên để đề xuất cho mỗi người dùng.

Qua việc so sánh kết quả sử dụng các độ đo như Precision@10 và NDCG@10, chúng tôi đánh giá khả năng đề xuất của mô hình Graph-based Filtering [22] so với các phương pháp khác.

Hình 18 thể hiện so sánh hiệu suất của mô hình Graph-based Filtering và các mô hình Collaborative Filtering, Content-based Filtering và Hybrid Filtering. Kết quả này giúp chúng tôi đánh giá

tính ổn định và hiệu quả của Graph-based Filtering trong bối cảnh của hệ thống khuyến nghị của chúng tôi.

Mô hình Graph-based Filtering là một bước tiến quan trọng để mở rộng khả năng đề xuất và đồng thời cung cấp sự linh hoạt trong việc tích hợp các yếu tố tương tác và mối quan hệ giữa người dùng và sản phẩm vào quá trình đề xuất.

6 Kết luận

Trong báo cáo này, chúng tôi đã thu thập, xây dựng và trình bày 3 bộ dữ liệu chính là Hotels, Restaurants và Attractions Dataset. Gồm 47 bộ dữ liệu cho khách sạn với tổng cộng 8,475 mẫu phục vụ cho 13 tỉnh/thành, mỗi bộ dữ liệu có 72 thuộc tính. Đối với nhà hàng, chúng tôi tạo ra 7 bộ dữ liệu với 1,251 mẫu cho 11 tỉnh/thành, mỗi bộ có 25 thuộc tính. Cuối cùng, với điểm thu hút, chúng tôi đã tạo 32 bộ dữ liệu với 671 mẫu cho 6 tỉnh/thành, mỗi bộ có 38 thuộc tính. Tổng cộng, có 86 bộ dữ liệu được tạo ra với hơn 7,500 user khác nhau và 10,397 lượt đánh giá, đảm bảo một quy trình nghiêm túc để đảm bảo chất lượng và hiệu suất của hệ thống. Bộ dữ liệu được xử lý để tạo ra các tập training và testing phù hợp với từng phương pháp khuyến nghị. Hiện tại, với phương pháp collaborative filtering chúng tôi đã cài đặt thành công sáu mô hình memory-based gồm: user-user cosine, user-user pearson, user-user jaccard và item-item cosine, item-item pearson, item-item jaccard; với phương pháp content-based filtering chúng tôi cũng đã cài đặt thành công mô hình content-based. Ngoài ra chúng tôi còn tiến hành xây dựng cho giải thuật hybrid-filtering, graph-based-filtering để đề xuất một danh sách các khách sạn, nhà hàng hoặc điểm thu hút cho hệ thống khuyến nghị. Chúng tôi còn tiến hành cài đặt thêm 3 mô hình gồm 2 mô hình machine learning là linear regression và random forest regression, 1 mô hình neural network để tiến hành dự đoán rating của 1 user cho 1 khách sạn, nhà hàng hoặc điểm thu hút bất kỳ. Kết quả tốt nhất mà chúng tôi đạt được là 9% MSE, 93% R2, 88% Precision@10 và 92% NDCG@10 với User-User cosine. Kết quả của chúng tôi đạt được khá cao từ đó cho thấy được hệ thống khuyến nghị du lịch của chúng tôi hoạt động khá tốt, có thể phát triển và ứng dụng vào thực tế. Bên cạnh đó, nó cũng đặt ra một thách thức cho các nhóm nghiên cứu sau về việc cải thiện kết quả cho bài toán.

Hướng phát triển trong tương lai:

- Bộ dữ liệu: Thu thập thêm dữ liệu từ các trang web đặt phòng khách sạn, nhà hàng hoặc điểm thu hút du lịch trực tuyến, cùng với đó là thu thập thêm các thuộc tính mới như: bình luận của người đánh giá, rating cho từng khía cạnh, giá,... để cho ra bộ dữ liệu đầy đủ thông tin hơn. Ngoài ra, xử lý hiện tượng thưa thớt trong các rating matrix cũng là một vấn đề rất quan trọng.
- Mô hình: Áp dụng các phương pháp, kỹ thuật khuyến nghị khác như: Collaborative Filtering dùng Knowledge-Based Recommender Systems [6], Demographic Recommender Systems [23], Ensemble-Based Recommender Systems [18],... để cải thiện kết quả dự đoán tốt hơn nữa.
- Ứng dụng sâu hơn vào bài toán thực tế: Phát triển cao hơn để giải quyết các bài toán như dùng mô hình để phân tích cảm xúc từ bình luận của user, sau đó dựa vào những đặc tính và cảm xúc của người dùng để đề xuất thứ họ mong muốn. Ngoài ra còn có thể giải quyết bài toán như dựa vào giọng nói để phân tích sâu hơn về tâm lý của người dùng, từ đó dự đoán và đề xuất ra sản phẩm,...

References

- [1] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering*, 17(6):734–749, 2005.
- [2] Shahriar Badsha, Xun Yi, Ibrahim Khalil, and Elisa Bertino. Privacy preserving user-based recommender system. In *2017 IEEE 37th international conference on Distributed Computing Systems (ICDCS)*, pages 1074–1083. IEEE, 2017.
- [3] Joan Borràs, Antonio Moreno, and Aida Valls. Intelligent tourism recommender systems: A survey. *Expert systems with applications*, 41(16):7370–7389, 2014.
- [4] Chenwei Cai, Ruining He, and Julian McAuley. Spmc: Socially-aware personalized markov chains for sparse sequential recommendation. *arXiv preprint arXiv:1708.04497*, 2017.
- [5] Lei Chen, Lu Zhang, Shanshan Cao, Zhiang Wu, and Jie Cao. Personalized itinerary recommendation: Deep and collaborative learning with textual information. *Expert Systems with Applications*, 144:113070, 2020.

- [6] Phung Do, Kha Nguyen, Thanh Nguyen Vu, Tran Nam Dung, and Tuan Dinh Le. Integrating knowledge-based reasoning algorithms and collaborative filtering into e-learning material recommendation system. In *Future Data and Security Engineering: 4th International Conference, FDSE 2017, Ho Chi Minh City, Vietnam, November 29–December 1, 2017, Proceedings 4*, pages 419–432. Springer, 2017.
- [7] Taehyun Ha and Sangwon Lee. Item-network-based collaborative filtering: A personalized recommendation method based on a user’s item network. *Information Processing & Management*, 53(5):1171–1184, 2017.
- [8] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pages 173–182, 2017.
- [9] Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. *Recommender systems: an introduction*. Cambridge University Press, 2010.
- [10] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.
- [11] Marcello Meldi and Alexandre Poux. A reduced order model based on kalman filtering for sequential data assimilation of turbulent flows. *Journal of Computational Physics*, 347:207–234, 2017.
- [12] Mustafa Misir and Michèle Sebag. Alors: An algorithm recommender system. *Artificial Intelligence*, 244:291–314, 2017.
- [13] Najdt Mustafa, Ashraf Osman Ibrahim, Ali Ahmed, and Afnizanfaizal Abdullah. Collaborative filtering: Techniques and applications. In *2017 International Conference on Communication, Control, Computing and Electronics Engineering (ICCCCEE)*, pages 1–6. IEEE, 2017.
- [14] Paromita Nitu, Joseph Coelho, and Praveen Madiraju. Improvising personalized travel recommendation system with recency effects. *Big Data Mining and Analytics*, 4(3):139–154, 2021.
- [15] Paromita Nitu, Joseph Coelho, and Praveen Madiraju. Improvising personalized travel recommendation system with recency effects. *Big Data Mining and Analytics*, 4(3):139–154, 2021.
- [16] Bansari Patel, Palak Desai, and Urvi Panchal. Methods of recommender system: A review. In *2017 international conference on innovations in information, embedded and communication systems (ICI-IECS)*, pages 1–4. IEEE, 2017.
- [17] S Kanaga Suba Raja, R Rishi, E Sundaresan, and V Srijit. Demand based crop recommender system for farmers. In *2017 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR)*, pages 194–199. IEEE, 2017.
- [18] Santosh Singh Rathore and Sandeep Kumar. Towards an ensemble based system for predicting the number of software faults. *Expert Systems with Applications*, 82:357–382, 2017.
- [19] Francesco Ricci, Lior Rokach, and Bracha Shapira. Introduction to recommender systems handbook. In *Recommender systems handbook*, pages 1–35. Springer, 2010.
- [20] Ayush Singhal, Pradeep Sinha, and Rakesh Pant. Use of deep learning in modern recommendation system: A summary of recent works. *arXiv preprint arXiv:1712.07525*, 2017.
- [21] Jieun Son and Seoung Bum Kim. Content-based filtering for recommendation systems using multiattribute networks. *Expert Systems with Applications*, 89:404–412, 2017.
- [22] Pengfei Song, Joshua D Trzasko, Armando Manduca, Runqing Huang, Ramanathan Kadirvel, David F Kallmes, and Shigao Chen. Improved super-resolution ultrasound microvessel imaging with spatiotemporal nonlocal means filtering and bipartite graph-based microbubble tracking. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, 65(2):149–167, 2017.
- [23] Nayana Vaidya and AR Khachane. Recommender systems-the need of the ecommerce era. In *2017 International Conference on Computing Methodologies and Communication (ICCMC)*, pages 100–104. IEEE, 2017.
- [24] Steven Van Canneyt, Olivier Van Laere, Steven Schockaert, and Bart Dhoedt. Using social media to find places of interest: a case study. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*, pages 2–8, 2012.
- [25] Lin Wan, Yuming Hong, Zhou Huang, Xia Peng, and Ran Li. A hybrid ensemble learning method for tourist route recommendations based on geo-tagged social networks. *International Journal of Geographical Information Science*, 32(11):2225–2246, 2018.
- [26] Yong Wang, Jiangzhou Deng, Jerry Gao, and Pu Zhang. A hybrid user similarity model for collaborative filtering. *Information Sciences*, 418:102–118, 2017.
- [27] Jian Wei, Jianhua He, Kai Chen, Yi Zhou, and Zuoyin Tang. Collaborative filtering and deep learning based recommendation system for cold start items. *Expert Systems with Applications*, 69:29–39, 2017.
- [28] Hong-Jian Xue, Xinyu Dai, Jianbing Zhang, Shujian Huang, and Jiajun Chen. Deep matrix factorization models for recommender systems. In *IJCAI*, volume 17, pages 3203–3209. Melbourne, Australia, 2017.