Hotels Recommendation System

Phạm Đức Thể[1,2*], Trần Thành Luân[1,2*], Nguyễn Văn Kiệt[1,2] and Huỳnh Văn Tín[1,2]

[1]Khoa Học & Kĩ Thuật Thông Tin, Trường Đại Học Công Nghệ Thông Tin, TP.HCM, Việt Nam.
[2]Đại Học Quốc Gia TP.HCM, Việt Nam.

*Corresponding author(s). E-mail(s): 19522253@gm.uit.edu.vn; 19521810@gm.uit.edu.vn;
Contributing authors: kietnv@uit.edu.vn; tinhv@uit.edu.vn;

Tóm tắt nội dung

Ngày nay, với sự phát triển bùng nổ của internet, việc tìm kiếm một thông tin hay một sản phẩm trên các trang web là việc không dễ dàng, nó làm cho người dùng mất nhiều thời gian để tìm kiếm các thông tin, sản phẩm phù hợp với nhu cầu của họ. Cụ thể trong lĩnh vực du lịch, việc đặt chỗ khách sạn trực tuyến thông qua các trang web là một điều tất yếu để cho du khách có được trải nghiệm du lịch trọn vẹn. Tuy nhiên, không phải lúc nào người dùng cũng tìm được khách sạn phù hợp với nhu cầu của họ, việc tìm kiếm này đôi khi lại mất rất nhiều thời gian và công sức. 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 các phương pháp khuyến nghị khác nhau. Từ đó, giúp cho việc tìm kiếm thông tin khách sạn trở nên dễ dàng và tối ưu hóa trải nghiệm người dùng trên các trang web đặt chỗ khách sạn du lịch trực tuyến. Trong báo cáo này, chúng tôi thực hiện xây dựng bộ dữ liệu Hotels Booking Dataset. Chúng tôi tiến hành xử lý dữ liệu để tạo ra các tập training và testing phù hợp với các phương pháp *collaborative filtering* và *content-based filtering*, dùng hai độ đo là *cosine* và *pearson* để tính độ tương đồ giữa *user-user* hoặc *item-item* để đứa ra các khuyến nghị. Sau đó, chúng tôi so sánh kết quả đạt được thông qua các độ đo: MSE, RMSE, MAE, NMAE, Accuracy. Kết quả tốt nhất theo mà chúng tôi đạt được là là 12.66 MSE, 3.56 RMSE, 2.60 MAE và 0.34 NMAE sử dụng *user-user cosine*.

Keywords: Hotels Recommendation System, Booking Hotels Dataset, Collaborative Filtering, Memory-based, Content-based Filtering.

1 Giới Thiệu

Hệ thống Khuyến nghị (Recommender System hoặc Recommendation System) là một mảng khá rộng của machine learning, và có tuổi đời ít hơn so với classification hay regression vì internet mới chỉ thực sự bùng nổ khoảng 10–15 năm gần đây. Có hai thực thể chính trong một recommendation system là *user* và *item*. *User* là *người dùng*; *item* là *sản phẩm*, ví dụ như các bộ phim, bài hát, cuốn sách, clip, hoặc cũng có thể là các người dùng khác trong bài toán gợi ý kết bạn. Mục đích chính của các recommender system là dự đoán mức độ quan tâm của một người dùng tới một sản phẩm nào đó, qua đó có chiến lược recommendation phù hợp.

Với sự phát triển mạnh mẽ của du lịch và internet, nhu cầu đặt phòng khách sạn trực tuyến càng ngày càng lớn. Khi World Wide Web tiếp tục phát triển theo cấp số nhân, quy mô và độ phức tạp của nhiều trang web cũng tăng theo. Đối với người dùng của các trang web này, cụ thể là trên các trang web đặt phòng khách sạn du lịch trực tuyến, việc tìm kiếm thông tin khách sạn ngày càng trở nên khó khăn và mất nhiều thời gian. 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 các phương pháp khuyến nghị khác nhau. Từ đó, giúp cho việc tìm kiếm thông tin khách sạn trở nên dễ dàng và tối ưu hóa trải nghiệm người dùng trên các trang web đặt chỗ khách sạn du lịch trực tuyến.

Trong báo cáo này, trước tiên chúng tôi giới thiệu các công trình liên quan trong Phần 2. Tiếp theo, chúng tôi trình bày về quy trình thu thập và tạo ra bộ dữ liệu *Booking Hotel Dataset* để sử dụng cho bài toán *Hotel Recommendation System*. Đối với từng phương pháp recommendation khác nhau, chúng tôi tiến hành các phương pháp xử lý dữ liệu khác nhau để tạo ra các tập dữ liệu training và testing (Phần 3). Hướng tiếp cận bài toán được mô tả chi tiết trong Phần 4. Trong Phần 5, chúng tôi tiến hành thực nghiệm và phân tích kết quả của các phương pháp recommendation system trên các tập dữ liệu khác nhau. Cuối cùng, chúng tôi rút ra kết luận ở Phần 6.

2 Công Trình Liên Quan

Mục đích chính của các recommender system là dự đoán mức độ quan tâm của một người dùng tới một sản phẩm nào đó. Các recommendation system thường được chia thành hai nhóm lớn:

- Lọc cộng tác (Collaborative filtering)[1]: là quá trình lọc hoặc đánh giá các mục thông qua ý kiến của người khác. Lọc cộng tác được hầu hết các hệ thống đề xuất sử dụng để tìm các mẫu hoặc thông tin tương tự của người dùng, kỹ thuật này có thể lọc ra các mục mà người dùng thích trên cơ sở xếp hạng hoặc phản ứng của những người dùng tương tự. Có hai loại lọc cộng tác:
  + User-User-based similarity/Collaborative Filtering.

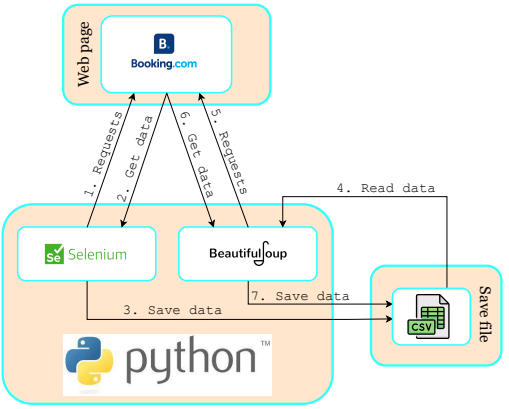
+ Item-Item-based similarity/Collaborative Filtering.

- Lọc dựa trên nội dung (Content-based filtering)[2]: là một trong những kỹ thuật đề xuất thành công nhất, dựa trên mối tương quan giữa các nội dung. Lọc dựa trên nội dung sử dụng thông tin item, được biểu diễn dưới dạng thuộc tính, để tính toán sự giống nhau giữa các item.

3 Bộ Dữ Liệu

3.1 Thu Thập Dữ Liệu

Bộ dữ liệu sử dụng trong báo cáo này có tên là *Booking Hotels Dataset* được chúng tôi thu thập từ trang web du lịch trực tuyến cho đặt chỗ booking.com. 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. Chi tiết về quy trình thu thập dữ liệu được trình bày ở Hình 1. Bộ dữ liệu được thu thập gồm 38,801 dòng dữ liệu và 9 thuộc tính với hơn 4,500 khách sạn thuộc 10 tỉnh/thành phố như: Đà Lạt, Hà Nội, TP. Hồ Chí Minh, . . . và gần 6,500 user khác nhau. Thông tin chi tiết về các thuộc tính của bộ dữ liệu được thể hiện trong Bảng 1.



Hình 1 Quy trình crawl data.

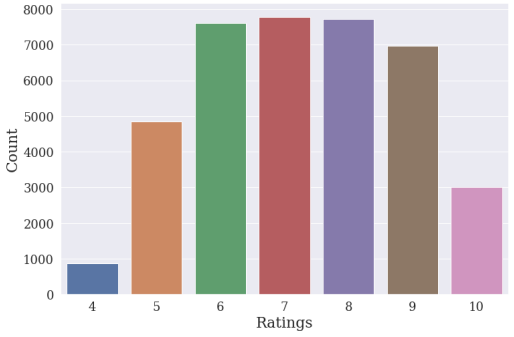
Bộ dữ liệu được thu thập từ tháng 11/2022 được sử dụng riêng cho mục đích học tập và nghiên cứu môn học Hệ Khuyến Nghị. Hình 2 thể hiện ví dụ về các điểm dữ liệu trong bộ dữ liệu. Các rating trong bộ dữ liệu phân bố từ 4 – 10 và không đồng đều (Hình 3), với rating thấp nhất là 4 và cũng là rating có số lượng ít nhất (gần 1,000 lượt rating) và rating cao nhất là 10 với gần 3,000 lượt rating. Khoảng rating từ 6 – 9 khá đồng đều với số lượng từ 7,000 – 8,000 lượt rating, trong đó cao nhất là 7 với gần 8,000 lượt rating.

Index	Thuộc tính	Ý nghĩa
0	URL Hotel	URL của khách sạn.
1	Location	Địa điểm của khách sạn (Tên tỉnh/TP).
2	HotelID	ID của khách sạn.
3	Name Hotel	Tên của khách sạn.
4	Descriptions	Mô tả, giới thiệu của khách sạn.
5	Address	Địa chỉ của khách sạn.
6	UserID	ID của user.
7	User	Tên của user (người đánh giá).
8	Rating	Rating của user.

Bảng 1 Thông tin chi tiết của các thuộc tính.

	URL Hotel	Location	HotelID	Name Hotel	Descriptions	Address	UserID	User	Rating
0	https://www.booking.com/hotel/vn/dalat-wind.vi...	Đà Lạt	4064	Dalat Wind Deluxe Hotel	Toạ lạc tại thành phố Đà Lạt, cách Hồ Xuân Hươ...	Lot R2 03-04, Golf Valley, Ward 2, Đà Lạt, Việ...	1187	Thảo	6
1	https://www.booking.com/hotel/vn/dalat-wind.vi...	Đà Lạt	4064	Dalat Wind Deluxe Hotel	Toạ lạc tại thành phố Đà Lạt, cách Hồ Xuân Hươ...	Lot R2 03-04, Golf Valley, Ward 2, Đà Lạt, Việ...	1284	Tran	5
2	https://www.booking.com/hotel/vn/dalat-wind.vi...	Đà Lạt	4064	Dalat Wind Deluxe Hotel	Toạ lạc tại thành phố Đà Lạt, cách Hồ Xuân Hươ...	Lot R2 03-04, Golf Valley, Ward 2, Đà Lạt, Việ...	5866	Tho	6
3	https://www.booking.com/hotel/vn/dalat-wind.vi...	Đà Lạt	4064	Dalat Wind Deluxe Hotel	Toạ lạc tại thành phố Đà Lạt, cách Hồ Xuân Hươ...	Lot R2 03-04, Golf Valley, Ward 2, Đà Lạt, Việ...	3033	Tuan	9
4	https://www.booking.com/hotel/vn/dalat-wind.vi...	Đà Lạt	4064	Dalat Wind Deluxe Hotel	Toạ lạc tại thành phố Đà Lạt, cách Hồ Xuân Hươ...	Lot R2 03-04, Golf Valley, Ward 2, Đà Lạt, Việ...	1406	Phan	7

Hình 2 5 điểm dữ liệu đầu tiên của bộ dữ liệu.



Hình 3 Phân bố rating của user.

3.2 Xử Lý Dữ Liệu

Dữ liệu tốt đóng vai trò quan trọng trong việc tạo ra các mô hình dự báo có độ chính xác cao và tổng quát hóa. Để có được bộ dữ liệu tốt ứng với từng tác vụ thì ta có những cách xử lý dữ liệu khác nhau. Bộ dữ liệu *Booking Hotels Dataset* của chúng tôi gặp phải một vấn đề lớn là dữ liệu bị thưa, khi mà số lượng user để lại rating cho khách sạn được thu thập là không nhiều và có rất nhiều user chỉ rating cho một khách sạn trong bộ dữ liệu, đó cũng là lý do chính làm cho bộ dữ hiệu của chúng tôi bị thưa dẫn đến khó đạt được các kết quả cao khi đánh giá hệ khuyến nghị. Trong báo cáo này, chúng tôi tiến hành xây dựng hệ khuyến nghị dựa trên hai phương pháp chính là *Collaborative Filtering (Lọc cộng tác)* và *Content-based Filtering (Lọc dựa trên nội dung)*,

đối với phương pháp lọc cộng tác ta chia làm hai phần là lọc cộng tác dựa trên user và lọc cộng tác dựa trên item. Vì vậy để dữ liệu phù hợp với từng phương pháp khác nhau, chúng tôi tiến hành xử lý dữ liệu để tạo ra các bộ dữ liệu training và testing cho từng phương pháp.

3.2.1 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 30 lần rating trở lên, kết quả thu được 139 user thỏa mãn. Tiếp theo, chúng tôi lọc ra các user đó từ bộ dữ liệu *Booking Hotels Dataset* để tạo ra tập dữ liệu mới. Từ tập dữ liệu này, với từng user khác nhau, chúng tôi chọn ra 5 dòng dữ liệu, có nghĩa là với mỗi user sẽ có 5 khách sạn được user này rating để làm tập testing. Tập training là các dữ liệu còn lại trong bộ dữ liệu *Booking Hotels Dataset* đã được loại bỏ đi các dữ liệu nằm trong tập testing. Kết quả thu được hai tập testing và training với số lượng dữ liệu lần lượt là 695 dòng dữ liệu và 38,066 dòng dữ liệu.

3.2.2 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 để chọn ra các khách sạn có từ 10 user rating trở lên, kết quả thu được 1,916 khách sạn thỏa mãn. Tiếp theo, chúng tôi lọc ra các khách sạn đó từ bộ dữ liệu *Booking Hotels Dataset* để tạo ra tập dữ liệu mới. Từ tập dữ liệu này, với từng khách sạn khác nhau, chúng tôi chọn ra 3 dòng dữ liệu, có nghĩa là với mỗi khách sạn sẽ chọn ra 3 user đã rating cho khách sạn để là tập testing. Tập training là các dữ liệu còn lại trong bộ dữ liệu *Booking Hotels 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 5,748 dòng dữ liệu và 32,273 dòng dữ liệu. Vì số lượng dữ liệu của tập testing khá lớn nên chúng tôi chỉ sử dụng 500 trong tổng số 5,748 dòng dữ liệu để test đánh giá hệ khuyến nghị.

3.2.3 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 10 lần rating trở lên, kết quả thu được 739 user thỏa mãn. Tiếp theo, chúng tôi lọc ra các user đó từ bộ dữ liệu *Booking Hotels Dataset* để tạo ra tập dữ liệu mới. Từ tập dữ liệu này, với từng user khác nhau, chúng tôi chọn ra 5 dòng dữ liệu, có nghĩa là với mỗi user sẽ có 5 khách sạn được user này rating để làm tập testing. Tập training là các dữ liệu còn lại trong bộ dữ liệu *Booking Hotels 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 3,695 dòng dữ liệu và 34,912 dòng dữ liệu. Đối với tập training chúng tôi tiếp tục xử lý dữ liệu như sau: gộp tất cả descriptions của các khách sạn mà mỗi user đã đi thành một descriptions cho user, từ đó thu được 6,741 điểm dữ liệu (tương ứng với 6,471 users); tiếp theo,

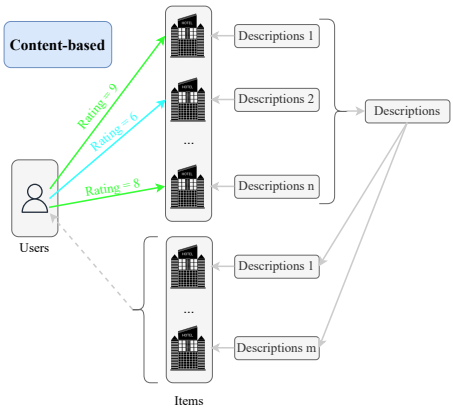
lấy từng descriptions của mỗi khách sạn, chúng tôi thu được tất cả 4,506 điểm dữ liệu (tương ứng với 4,506 khách sạn).

4 Hướng Tiếp Cận

Từ bộ dữ liệu *Booking Hotel 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)* và *collaborative filtering (CF – lọc công tác)* (được trình bày rõ hơn trong Phần 3.1 và 3.2). Để đánh giá hệ khuyến nghị sử dụng phương pháp lọc cộng tác, chúng tôi sử dụng các độ đo: Mean Squared Error, Root Mean Squared Error, Mean Absolute Error, Normalized Mean Absolute Error; đối với hệ khuyến nghị sử dụng phương pháp lọc dựa trên nội dung, chúng tôi đánh gia thông qua độ đo Accuracy (chi tiết trong Phần **??**).

4.1 Content-based Filtering

Trong các hệ thống khuyến nghị dựa trên nội dung, các thuộc tính mô tả của các *item* được sử dụng để đưa ra các khuyến nghị. Thuật ngữ *"content"* đề cập đến những mô tả này. Trong các phương pháp dựa trên nội dung, rating và hành vi mua của người dùng được kết hợp với thông tin nội dung có sẵn trong các *item* [4]. Ví dụ: hãy xem xét một tình huống trong đó người dùng A đã rating cao cho khách sạn X, nhưng chúng tôi không có quyền truy cập vào rating của những người dùng khác. Do đó, các phương pháp lọc cộng tác bị loại trừ. Tuy nhiên, phần mô tả của khách sạn X chứa các từ khóa về vị trí địa lý, tiện ích, dịch vụ, diện tích, giá tiền, . . . tương tự như các khách sạn khác, chẳng hạn như khách sạn Y, Z, ... Trong những trường hợp như vậy, những khách sạn này có thể được khuyến nghị cho người dùng A.



Hình 4 Content-based filtering

Hình 4 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 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. 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:

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

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.

4.2 Collaborative Filtering

Các mô hình lọc cộng tác sử dụng sức mạnh cộng tác của rating do nhiều người dùng cung cấp để đưa ra khuyến nghị. Thách thức chính trong việc thiết kế các phương pháp lọc cộng tác là vấn đề *sparse (thưa thớt)* của các *rating matix (ma trận đánh giá)*. Hãy xem xét một ví dụ về ứng dụng đặt phòng khách sạn trong đó người dùng A rating cho biết họ thích hoặc không thích các khách sạn cụ thể. Hầu hết người dùng sẽ chỉ đặt một phần nhỏ trong vô số các khách sạn ở từng địa điểm. Kết quả là đa số các rating là không xác định. Do đó, dẫn đến việt tính toán độ tương đồng không đạt được hiệu quả cao.

Để tính toán độ tương đồng, chúng tôi sử dụng hai độ đo là:

- *Độ đo tương đồng cosine (cosine similarity)*:

$$sim(U_u, U_v) = cos(U_u, U_v) = \frac{U_u \cdot U_v}{U_u \cdot U_v} = \frac{\sum_i r_{u,i} r_{v,i}}{\sqrt{\sum_i r_{u,i}^2} \sqrt{\sum_i r_{v,i}^2}} \tag{2}$$

- *Độ đo tương đồng pearson (pearson similarity)*:

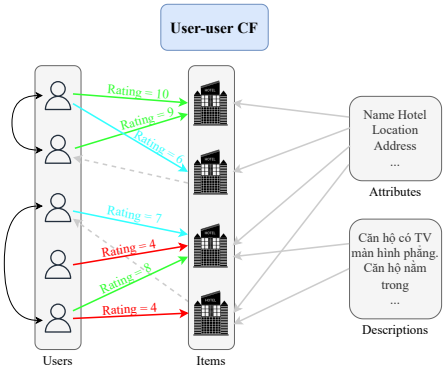
$$sim(U_u, U_v) = \frac{\sum_i (r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_i (r_{u,i} - \bar{r}_u)^2} \sqrt{\sum_i (r_{v,i} - \bar{r}_v)^2}} \tag{3}$$

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 memory-based và model-based [5, 6]. Trong báo cáo này, chúng tôi chỉ sử dụng phương pháp memory-based gồm:

4.2.1 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. 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) hoặc pearson similarity (Công thức 3). Minh họa về user-user collaborative filtering được thể hiện trong Hình 5.

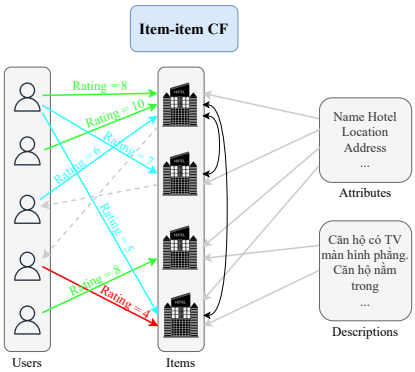


Hình 5 User-user collaborative filtering

4.2.2 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. Để tính toán độ tương đồng giữa item mục tiêu và những item khác, ta có thể sử dụng cosine similarity (Công thức 2) hoặc pearson similarity (Công thức 3). Minh họa về item-item collaborative filtering được thể hiện trong Hình 6.

Ưu điểm của các kỹ thuật memory-based là chúng dễ thực hiện và các khuyến nghị thu được thường dễ giải thích. Mặt khác, các thuật toán memory-based không hoạt động tốt với các rating matrix thưa thớt. Ví dụ: có thể khó tìm được những người dùng tương tự như B, người đã rating cho X. Trong



Hình 6 Item-item collaborative filtering

những trường hợp như vậy, rất khó để dự đoán chính xác rating của B cho X. Nói cách khác, những phương pháp như vậy có thể không bao phủ đầy đủ các dự đoán rating.

4.3 Độ đo đánh giá

Có nhiều phương pháp để đánh giá hệ khuyến nghị như: *đánh giá offline không quan tâm thứ tự sắp xếp, đánh giá offline quan tâm thứ tự sắp xếp, đánh giá online và A/B testing.* Trong bài này, chúng tôi dùng phương pháp *đánh giá offline không quan tâm thứ tự sắp xếp* sử dụng *Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE)* và *Normalized Mean Absolute Error (NMAE)* để đánh giá hệ khuyến nghị áp dụng kĩ thuật *collaborative filtering*; *Accuracy (A)* để đánh giá hệ khuyến nghị áp dụng kĩ thuật *content-based filtering.* Các độ đo được định nghĩa như sau [4]:

$$MSE = \frac{\sum_{(u,i)} (\hat{r}_{u,i} - r_{u,i})^2}{n} \tag{4}$$

$$RMSE = \sqrt{\frac{\sum_{(u,i)} (\hat{r}_{u,i} - r_{u,i})^2}{n}} \tag{5}$$

$$MAE = \frac{\sum_{(u,i)} |\hat{r}_{u,i} - r_{u,i}|}{n} \tag{6}$$

$$NMAE = \frac{MAE}{r_{max} - r_{min}} \tag{7}$$

$$A = \frac{t}{n} \tag{8}$$

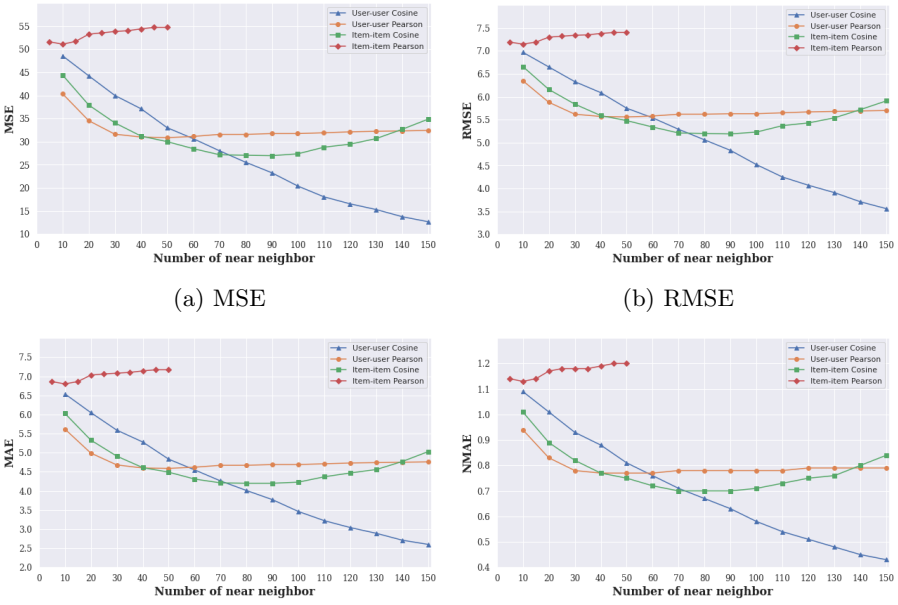
trong đó: $\hat{r}_{u,i}$ là giá trị rating dự đoán của user u cho item i; $r_{u,i}$ là giá trị rating thực tế của user u cho item i; n là số lượng điểm dữ liệu được sử dụng để đánh giá; r_{max} và r_{min} lần lượt là giá trị rating cao nhất và thấp nhất

trong tập dữ liệu được sử dụng để đánh giá; t là số lượng điểm dữ liệu được dự đoán đúng trong tập đánh giá, giả sử hệ khuyến nghị sẽ khuyến nghị một danh sách S gồm k khách sạn cho user u, nếu user u đã đi khách sạn i và khách sạn $i \in S$ thì sẽ được tính là một dự đoán đúng.

5 Thực Nghiệm và 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 3 và đánh giá kết quả thông qua các độ đo $MSE, RMSE, MAE, NMAE$ và Accuracy để so sánh kết quả đạt được.

5.1 Collaborative Filtering



(a) MSE

(b) RMSE

(c) MAE

(d) NMAE

Hình 7 Kết quả của collaborative filtering

Với hai phương pháp là *user-user CF* và *item-item CF* sử dụng hai độ tương đồng là *cosine similarity* và *pearson similarity*, qua đó chúng tôi có được bốn mô hình là *user-user cosine, user-user pearson, item-item cosine* và *item-item pearson* được đánh giá và so sánh thông qua các độ đo $MSE, RMSE, MAE, NMAE$. Hình 7 trình bày toàn bộ 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; đặc biệt, đối với mô hình *item-item pearson* chúng tôi chỉ lấy từ 5 đến 50 với bước nhảy là 5 để chọn ra số lượng lân cận gần nhất tốt nhất 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, item-item cosine* và *item-item pearson.*
- **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 2. Có thể thấy mô hình *user-user cosine* và *item-item pearson* lần lượt là hai mô hình cho kết quả tốt nhất và tệ nhất trong số bốn mô hình được thực nghiệm ở tất 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 thớt nhẹ. Ngược lại, mô hình *item-item pearson* cho kết quả tệ nhất, nguyên nhân có thể là vì rating matrix bị hiện tượng thưa thớt nghiêm trọng và độ tương đồng *pearson* là không thực thực tốt trong trường hợp này.

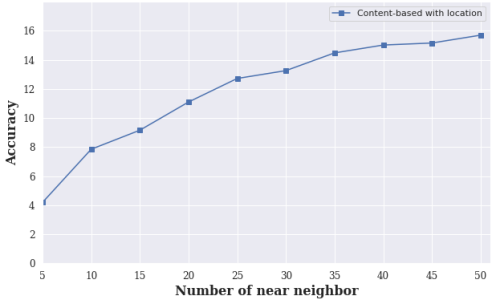
Mô hình	Độ đo				Thời gian khuyến nghị trung bình (s)
	MSE	RMSE	MAE	NMAE	
User-user cosine	**12.66**	**3.56**	**2.60**	**0.43**	55
User-User pearson	30.87	5.56	4.59	0.77	49
Item-item cosine	26.97	5.19	4.20	0.70	27
Item-item pearson	51.09	7.15	6.80	1.13	26

Bảng 2 Kết quả tốt nhất của các mô hình theo các độ đo.

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 mà người dùng muốn đến trong lần tiếp theo thì chúng tôi giới hạn số lượng lân cận tương đồng từ 5-50 với các bước nhảy là 5 để tìm ra lân cận tốt nhất.

Hình 8 trình bày kết quả accuracy theo số lượng khách sạn được khuyến nghị của content-based filtering, với số lượng 25 lân cận thì độ chính xác bắt đầu tăng chậm. Vì vậy, chúng tôi chọn 25 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 độ chính xác là 12.72% và thời gian tính toán khuyến nghị là 60s.



Hình 8 Kết quả của content-based filtering

6 Kết Luận

Trong báo cáo này, chúng tôi đã thu thập, xây dựng và trình bày bộ dữ liệu *Booking Hotels Dataset*, một bộ dữ mới cho bài toán khuyến nghị đặt phòng khách sạn du lịch trực tuyến. Bộ dữ liệu gồm 38,801 dòng dữ liệu và 9 thuộc tính với hơn 4,500 khách sạn thuộc 10 tỉnh/thành phố như: Đà Lạt, Hà Nội, TP. Hồ Chí Minh, ... và gần 6,500 user khác nhau. 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 bốn mô hình *memory-based* gồm: *user-user cosine, user-user pearson, item-item cosin* và *item-item pearson*; với phương pháp *content-based filtering* chúng tôi cũng đã cài đặt thành công mô hình *content-based*. Kết quả tốt nhất mà chúng tôi đạt được là 12.66 MSE, 3.56 RMSE, 2.60 MAE và 0.34 NMAE. Kết quả của chúng tôi đạt được không cao, nó 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 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 [4]: *Collaborative Filtering* dùng *model-based*, *Knowledge-Based Recommender Systems*, *Demographic Recommender Systems*, *Hybrid and Ensemble-Based Recommender Systems* ... để cải thiện kết quả dự đoán tốt hơn nữa.

Tài liệu

[1] Schafer, J.B., Frankowski, D., Herlocker, J., Sen, S. (2007). Collaborative Filtering Recommender Systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds) The Adaptive Web. Lecture Notes in Computer Science, vol 4321. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-72079-9_9

[2] Son, Jieun, and Seoung Bum Kim. "Content-based filtering for recommendation systems using multiattribute networks." Expert Systems with Applications 89 (2017): 404-412.

[3] Vũ Hữu, Tiệp. “Machine Learning cơ bản.” (2022).

[4] Aggarwal, Charu C. Recommender systems. Vol. 1. Cham: Springer International Publishing, 2016.

[5] J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. Collaborative filtering recommender systems. In The adaptive web, pages 291–324. Springer, 2007.

[6] Michael D Ekstrand, John T Riedl, Joseph A Konstan, et al. Collaborative filtering recommender systems. Foundations and Trends® in Human–Computer Interaction, 4(2):81–173, 2011.