FoodSim: Hệ thống đề xuất món ăn dựa trên đa thang đo kết hợp

Đặng Chí Thành, Đặng Thị Thúy Hồng, Trần Huỳnh Quốc An, 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
{20520761, 20520523, 20520955}gm.uit.edu.vn
tinhv@uit.edu.vn

Tóm tắt nội dung

Chúng tôi nhận thấy rằng, trong thời đại kỷ nguyên số lên ngôi, lối sống của người trẻ thường xuyên bận rộn, khiến thời gian sinh hoạt của họ bị đảo lộn và việc bỏ bữa là điều thường xuyên không hiếm gặp. Để góp phần giải quyết vấn đề này, nhằm giúp người dùng, đặc biệt là người dùng Việt Nam, dễ dàng hơn trong việc lựa chọn món ăn phù hợp với khẩu vị và nhu cầu dinh dưỡng, chúng tôi đã phát triển một hệ thống đề xuất món ăn Việt Nam. Trong nghiên cứu này, chúng tôi thực hiện xây dựng bộ dữ liệu Vietnamese Food Dataset và tiến hành xử lý dữ liệu để tạo ra bộ dữ liệu chất lượng phù hợp với các phương pháp Collaborative Filtering và Content-based Filtering. Chúng tôi sử dụng hai độ đo là Cosine và Pearson để tính độ tương đồng giữa các món ăn hoặc người dùng nhằm đưa ra các khuyến nghị tốt nhất. Sau đó, chúng tôi so sánh kết quả đạt được thông qua các độ đo: MSE, RMSE, MAE, NMAE, MRR và Precision@K. Đối với từng phương pháp, kết quả tốt nhất chúng tôi đạt được là 4.2581 MSE, 2.0635 RMSE với User-user Cosine; 1,6902 MAE, 0.338 NMAE với Itemt-item Cosine và 70,77% MRR, 48,5% Precision@K khi sử dụng Content-Based.

1 Giới thiệu

Hệ thống Khuyến nghị, một lĩnh vực của Machine Learning, đã phát triển đáng kể trong những năm gần đây, được thúc đẩy bởi sự mở rộng nhanh chóng của internet. Khác với các tác vụ phân loại hoặc hồi quy thông thường, Hệ thống Khuyến nghị tập trung vào việc dự đoán sở thích của người dùng đối với các sản phẩm khác nhau. Các thực thể chính trong Hệ thống Khuyến nghị bao gồm người dùng và sản phẩm. Người dùng đại diện cho cá nhân, trong khi sản phẩm có thể đại diện cho nhiều sự vật khác nhau như phim, bài hát, sách, video, hoặc thậm chí là người dùng khác trong mạng xã hội. Mục tiêu cơ bản của Hệ thống Khuyến nghị là dự đoán mức độ quan tâm mà người dùng có thể có đối với một sản phẩm cụ thể, từ đó tạo ra các đề xuất cá nhân.

Từ khi Internet xuất hiện và phát triển một cách bùng nổ đã khiến cho khối lượng công việc của nhiều người trẻ tăng lên và họ có thể bỏ qua nhiều bữa trong ngày vì nhiều lý do, ví dụ như không có thời gian để chọn lựa món ăn, không biết nên chọn món nào hay là những món ăn họ mong muốn tốn quá nhiều thời gian để chuẩn bị. Nếu họ chỉ bỏ bữa với tần suất ít thì sẽ không có nhiều vấn đề xảy ra, nhưng nếu điều này diễn ra thường xuyên sẽ làm ảnh hưởng đến sức khỏe thể chất và cả tinh thần. Cơ thể sẽ dần trở nên thiếu chất, thiếu năng lượng, trầm trọng hơn là cơ thể có thể suy nhược, suy giảm chức năng não bộ khiến họ không thể tập trung làm việc, rối loạn tiêu hóa và còn nhiều tác hại khôn lường khác nữa. Vì vậy, để giúp người dùng tiết kiệm thời gian và công sức trong việc tìm kiếm món ăn phù hợp với khẩu vị và sở thích cá nhân, chúng tôi xây dựng một hệ khuyến nghị món ăn Việt Nam FoodSim bằng cách sử dụng các phương pháp khuyến nghị khác nhau. Từ đó, giúp việc tìm kiếm của người dùng trở nên dễ dàng và tối ưu hơn.

Trong nghiên cứu này, trước tiên chúng tôi giới thiệu các công trình liên quan trong Phần 2. Tiếp theo, trong Phần 3, chúng tôi trình bày về quy trình thu thập và tạo ra bộ dữ liệu Vietnamese Food Dataset để sử dụng cho bài toán Food Recommendation System. 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. Tiếp đến, chúng tôi trực quan hệ thống trong Phần 6 Cuối cùng, chúng tôi rút ra kết luận ở Phần 7.

2 Các 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) (Schafer et al.): là một kỹ thuật phổ biến được sử

dụng trong hầu hết các hệ thống khuyến nghị. Kỹ thuật này tìm kiếm các mối tương quan dựa trên thông tin giữa các người dùng, và dựa trên đó để lọc ra các item mà người dùng có thể thích. Cơ sở của việc lọc này là xếp hạng hoặc phản ứng của những người dùng khác có sự tương đồng trên item đó. Lọc cộng tác có hai loại chính: Lọc cộng tác dựa trên người dùng (User-User-based similarity/Collaborative Filtering). Lọc cộng tác dựa trên item (Item-Item-based similarity/Collaborative Filtering).

- Lọc dựa trên nội dung (Content-based filtering) (Son and Kim, 2017): là một trong những kỹ thuật đề xuất thành công nhất, nó dựa trên mối tương quan giữa các nội dung. Kỹ thuật này sử dụng thông tin về item, được biểu diễn dưới dạng thuộc tính, để tính toán sự giống nhau giữa các item.

Trong bối cảnh ngày càng nhiều người sử dụng internet để tìm kiếm thông tin về thực phẩm và dinh dưỡng, việc phát triển các hệ thống đề xuất có thể cung cấp các gợi ý thực phẩm phù hợp với nhu cầu và sở thích cá nhân của người dùng đã trở nên cần thiết. Một số công trình nghiên cứu đã tập trung vào việc phát triển các hệ thống đề xuất thực phẩm dựa trên các thuật toán học máy như công trình nghiên cứu của Trang và cộng sự là 'An overview of recommender systems in the healthy food domain' (Trang Tran et al., 2018) đã phân tích các kỹ thuật khuyến nghị cho người dùng về các thực phẩm lành mạnh, tốt cho sức khỏe. Bài báo cũng đã thảo luận về các thách thức nghiên cứu liên quan đến việc phát triển các công nghệ đề xuất thực phẩm trong tương lai. Bên cạnh đó, cũng có một số công trình nghiên cứu khác về chủ đề này như "A Food Recommender System Considering Nutritional Information and User Preferences"(Toledo et al., 2019). Bài báo này trình bày một khung tổng quát cho các đề xuất về kế hoạch ăn uống hàng ngày, kết hợp với việc quản lý thông tin về dinh dưỡng và sở thích của người dùng. Điểm đặc biệt của công trình này là đề xuất việc quản lý đồng thời thông tin về dinh dưỡng và sở thích của người dùng, điều mà các công trình trước đây chưa từng có. Hay bài báo "Food Recommendation System Based on Data Clustering Techniques and User Nutrition Records"(Al-Chalabi and Jasim, 2022) đề xuất một hệ thống khuyến nghị thực phẩm dựa trên thông tin dinh dưỡng. Hệ thống này sử dụng các phương pháp khuyến nghị và phương pháp học máy để tạo ra các đề xuất cho các mặt hàng thực phẩm thiết yếu. Trong hệ thống này, phương pháp phân cụm K-means và phân loại Random Forest được sử dụng. Ngoài ra còn có bài báo "Food Recommender Systems Important Contributions, Challenges and Future"(Trattner and Elsweiler, 2017) thảo luận về các phương pháp dựa trên nội dung, lọc cộng tác và tiếp cận lai đã được áp dụng cho dữ liệu đánh giá nhà hàng để đề xuất thực đơn và dữ liệu mua sắm trực tuyến để khuyến nghị món ăn.

Tuy nhiên, việc phát triển các hệ thống đề xuất thực phẩm vẫn còn nhiều thách thức. Việc đảm bảo rằng các gợi ý thực phẩm không chỉ phù hợp với sở thích của người dùng mà còn phù hợp với các yêu cầu dinh dưỡng của họ cũng là một vấn đề đáng quan tâm. Đặc biệt, một trong những thách thức lớn nhất là việc xử lý và phân tích lượng lớn dữ liệu về thực phẩm và người dùng cũng rất quan trọng. Nhưng, hiện nay, chúng ta vẫn thiếu một bộ dữ liệu chất lượng về ẩm thực Việt Nam để phục vụ cho nghiên cứu. Do đó, trong khuôn khổ của đề tài này, chúng tôi đã tiến hành tự thu thập và xây dựng bộ dữ liệu Vietnamese Food Dataset. Chúng tôi hy vọng rằng với những đóng góp trong nghiên cứu này sẽ mở ra những hướng đi mới trong lĩnh vực đề xuất thực phẩm.

3 Bộ dữ liệu

Nhằm phục vụ nghiên cứu này, chúng tôi tổng hợp thông tin món ăn và đánh giá của người dùng từ nhiều nguồn khác nhau và lưu trữ ở định dạng CSV trong hai tập dữ liệu, bao gồm foods và ratings, tạo nên bộ dữ liệu Vietnamese Food Dataset. Trong đó, foods chứa thông tin về các món ăn, được sử dụng cho phương pháp lọc dựa trên nội dung (sẽ giới thiệu tại Phần 4.1) và ratings chứa đánh giá của người dùng được phục vụ phương pháp lọc cộng tác (sẽ giới thiệu tại Phần 4.2)

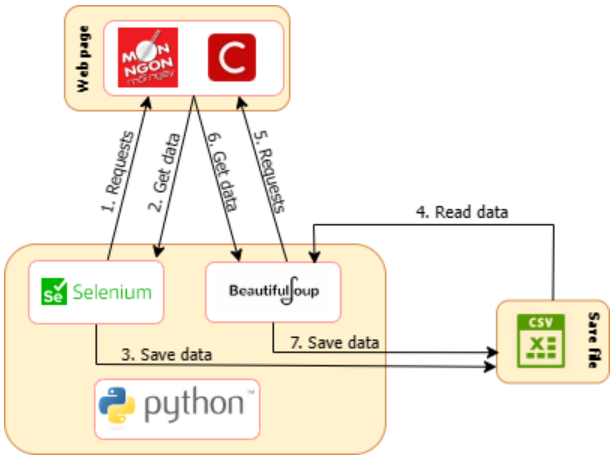
3.1 Thu thập dữ liệu

3.1.1 Bộ dữ liệu cho lọc dựa trên nội dung

Sử dụng hai thư viện thu thập dữ liệu trực tuyến mạnh mẽ là Selenium và BeautifulSoup, chúng tôi thu thập thông tin về các món ăn trên hai trang thông tin điện tử Việt Nam bao gồm Món Ngon Mỗi Ngày (monngonmoingay.com[1]) và Cooky.vn (cooky.vn[2]). Các thuộc tính được tổng hợp bao gồm tên món ăn, nguyên liệu, cách chế biến, giá trị dinh dưỡng, hình ảnh món ăn và một số thuộc tính phụ

[1]https://monngonmoingay.com/
[2]https://www.cooky.vn/cong-thuc

trợ khác. Trước khi đến với công đoạn tiền xử lý, dữ liệu thu thập từ 2 website trên được kết hợp và tạo thành foods. Chi tiết về quy trình thu thập dữ liệu được mô tả trong Hình 1:



Hình 1: Quy trình xây dựng bộ dữ liệu.

Tổng cộng, foods bao gồm 15 thuộc tính và 5509 món ăn, đại diện cho hầu hết các món ăn thông dụng tại Việt Nam. Bảng 1 mô tả các thông tin của các thuộc tính và giải thích ý nghĩa của các thuộc tính có trong tập dữ liệu foods.

Đáng lưu ý rằng foods không bao gồm thuộc tính định danh (ID), thay vào đó, chúng tôi sử dụng chỉ mục (index) của dataframe để định danh mỗi món ăn trong bộ dữ liệu.

3.1.2 Bộ dữ liệu cho lọc cộng tác

Nhằm xây dựng bộ dữ liệu đánh giá của người dùng đối với món ăn, chúng tôi xây dựng một file bao gồm danh sách các món ăn cũng như các thông tin về món ăn đó và đến 100 người dùng, với mỗi người dùng sẽ đánh giá khoảng 500 món ăn trong tổng số 4000 món ăn. Bộ dữ liệu ban đầu thu được khoảng 50000 dòng về đánh giá của mỗi người dùng cho món ăn. Bộ dữ liệu bao gồm 3 cột là id người dùng, chỉ mục món ăn và mức độ ưa thích món ăn của người dùng đối với món ăn đó (xếp hạng từ 0.0 là không thích đến 5.0 là cực kỳ thích, bước nhảy 0.5). Số món ăn mà người dùng đánh giá ít nhất là 436 món ăn và nhiều nhất là 566 món ăn. Món ăn có lượt đánh giá ít nhất là 2 lần và nhiều nhất là 26 lần.

3.2 Tiền xử lý dữ liệu

3.2.1 Bộ dữ liệu cho lọc dựa trên nội dung

Bộ dữ liệu foods sau khi được thu thập có rất nhiều vấn đề đòi hỏi sự chỉnh sửa cẩn thận. Do đó, các phương pháp tiền xử lý quan trọng đã được áp dụng, bao gồm việc loại bỏ các dòng chứa giá trị null và loại bỏ các dòng có toàn bộ 3 thuộc tính trùng nhau bao gồm tên món ăn, nguyên liệu và cách chế biến. Lý giải cho điều này, chúng tôi nhận thấy rằng nhiều món ăn mặc dù có tên giống nhau nhưng lại khác nguyên liệu và phương pháp chế biến, dẫn đến hương vị món ăn cũng thay đổi. Nói cách khác, 2 món ăn hoàn toàn khác nhau. Sau khi đã loại bỏ một cách tổng quát các giá trị gây nhiễu, chúng tôi tiến hành xử lý các giá trị văn bản, bao gồm chuẩn hóa unicode, loại bỏ các biểu tượng cảm xúc (emoji), xóa các khoảng trắng thừa và thay thế các từ viết tắt (ví dụ “1/2m” thành “1/2 muỗng”, “1/3m” thành “1/3 muỗng”).

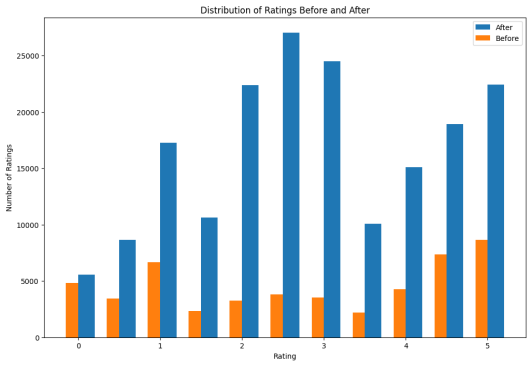
Cuối cùng, chúng tôi tạo một bản sao của foods với tên gọi foods_modeling, chứa một số ít các thuộc tính quan trọng, phục vụ cho phương pháp lọc dựa trên nội dung, bao gồm tên món ăn, nguyên liệu, giới thiệu món ăn, thẻ món ăn và hàm lượng dinh dưỡng. Chúng tôi cho rằng đây là những thuộc tính chủ yếu thể hiện đặc điểm của món ăn, đồng thời cho thấy sự tương quan với các món ăn khác trong tập dữ liệu. Bộ dữ liệu foods_modeling sẽ được vector hóa (vectorizer) theo phương pháp CountVectorizer và TF-IDF cho tất cả thuộc tính, ngoại trừ tên món ăn nhằm áp dụng các độ đo tương quan.

3.2.2 Bộ dữ liệu cho lọc cộng tác

Phân tích bộ dữ liệu ratings, chúng tôi phát hiện một số người dùng đánh giá lại cùng một món ăn. Để đảm bảo tính chính xác của dữ liệu, chúng tôi đã loại bỏ các đánh giá cũ, chỉ giữ lại đánh giá mới nhất của mỗi người dùng đối với mỗi món ăn. Ngoài ra, còn có một số lượng lớn điểm dữ liệu bị thiếu. Điều này có thể ảnh hưởng đến độ chính xác và hiệu suất của hệ thống khuyến nghị. Nhằm khắc phục vấn đề này, chúng tôi quyết định lấp đầy 40% các điểm dữ liệu còn trống bằng giá trị trung vị nhằm giữ nguyên tính thống kê của dữ liệu. Lý giải cho việc chúng tôi chỉ điền 40% dữ liệu bị khuyết chủ yếu đến từ giới hạn của phần cứng trong quá trình tính toán. Song, điều này cũng góp phần đảm bảo khả năng đại diện cho cảm nhận người dùng của dữ liệu. Sau khi đã điền các giá trị khuyết, chúng tôi thu được một bộ dữ liệu mới bao gồm khoảng 180,000 đánh giá, với số lượt người dùng đánh giá ít nhất là 1641 món ăn và nhiều nhất là 1989 món ăn. Bên cạnh đó, số món ăn có lượt đánh giá ít nhất là 27 lần và nhiều nhất là 68 lần. Hình 2 dưới đây thống kê số lượt đánh giá trước và sau khi thực hiện điền khuyết dữ liệu.

STT	Thuộc tính	Ý nghĩa
1	Tên món	Tên của món ăn
2	Giới thiệu	Thông tin ngắn gọn mô tả chung về món ăn
3	Loại món	Món mặn hay món chay
4	Khẩu phần	Số lượng người có thể dùng món ăn
5	Thời gian nấu	Thời gian hoàn thành món ăn
6	Nguyên liệu	Nguyên liệu cần thiết để nấu món ăn
7	Cách làm	Hướng dẫn chi tiết về cách nấu món ăn
8	Thẻ món ăn	Những từ khóa hoặc nhãn liên quan đến món ăn giúp dễ dàng tìm kiếm, phân loại
9	Calories	Hàm lượng calo trong món ăn (calo)
10	Chất béo	Hàm lượng chất béo trong món ăn (gam)
11	Chất xơ	Hàm lượng chất xơ trong món ăn (gam)
12	Đường	Hàm lượng đường trong món ăn (gam)
13	Protein	Hàm lượng protein trong món ăn (gam)
14	Link ảnh	Liên kết dẫn đến hình ảnh của món ăn
15	Hàm lượng dinh dưỡng	Được tổng hợp từ các thuộc tính 9 đến 13

Bảng 1: Thông tin thuộc tính của tập dữ liệu foods



Hình 2: Thống kê số lượt đánh giá trước và sau khi thực hiện điền khuyết dữ liệu.

4 Hướng tiếp cận

Trong nghiên cứu này, chúng tôi áp dụng lọc dựa trên nội dung (Content-based Filtering) và lọc cộng tác (Collaborative Filtering), hai phương pháp khuyến nghị phổ biến và chính xác nhất. Đồng thời, chúng tôi cũng kiểm định mức độ hiệu quả của các hướng tiếp cận trên đối với bài toán khuyến nghị món ăn thông qua một số độ đo đánh giá và nhận định của thành viên trong nhóm.

4.1 Độ đo tương quan

4.1.1 Content-based Filtering

Nghiên cứu này sử dụng nhiều biện pháp tương quan để cung cấp mức độ bao quát lớn hơn về các kết quả được đề xuất và cải thiện độ chính xác.

Cụ thể, Cosine và Pearson được áp dụng đối với cả Content-based Filtering và Collaborative Filtering. Đây là 2 độ đo phổ biến nhất đối với bài toán khuyến nghị, với các công thức tính như sau:

$$\begin{aligned} \text{sim}_{\text{cosine}}(X,Y) = \cos(X,Y) &= \frac{X \cdot Y}{\|X\| \cdot \|Y\|} \\ &= \frac{\sum_{i=0}^{n-1} X_i \cdot Y_i}{\sqrt{\sum_{i=0}^{n-1} X_i^2} \cdot \sqrt{\sum_{i=0}^{n-1} Y_i^2}} \end{aligned} \tag{1}$$

$$\text{sim}_{\text{pearson}}(X,Y) = \frac{\sum_{i=0}^{n-1} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=0}^{n-1} \left(X_i - \bar{X}\right)^2} \sqrt{\sum_{i=0}^{n-1} \left(Y_i - \bar{Y}\right)^2}} \tag{2}$$

Trong đó:

- X_i và Y_i là giá trị của thuộc tính thứ i trong X và Y tương ứng.
- $\bar{X}$ và $\bar{Y}$ là giá trị trung bình của các giá trị trong X và Y tương ứng.

Bên cạnh đó, Content-based Filtering sẽ có thêm 4 thước đo tương quan khác là TfidfRecommender, Jaccard, BM25 và độ đo tổng hợp. Cụ thể, về định nghĩa, TfidfRecommender sử dụng phương pháp vectorize là TF-IDF và độ đo Cosine. Jaccard đánh giá sự tương đồng giữa hai tập hợp bằng cách tính tỉ lệ số phần tử chung giữa chúng và tổng số phần tử duy nhất trong cả hai tập hợp. Mặt khác, BM25 sử dụng trọng số IDF và tần suất xuất hiện của từ TF nhằm tính điểm tương đồng giữa một tài liệu và một truy vấn. Đối với độ đo tổng hợp, nó được tính dựa trên việc kết hợp các kết quả khuyến nghị của 5 độ đo trên và nhân điểm tương quan với trọng số là 0.2. Độ đo này nhằm trung hòa các đề xuất khác nhau, từ đó đưa ra cho người dùng đa dạng các đề xuất món ăn hơn.

4.1.2 Collaborative Filtering

Mặc dù cũng sử dụng chung độ đo Pearson và Cosine tuy nhiên về định nghĩa, hai công thức này có một số thay đổi nhằm phù hợp với bài toán lọc cộng tác. Cụ thể, công thức của các độ đo này như sau:

$$\text{sim}_{\text{pearson}}(U_u, U_v) = \frac{\sum_{i=0}^{n-1}(r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i=0}^{n-1}(r_{u,i} - \bar{r}_u)^2} \cdot \sqrt{\sum_{i=0}^{n-1}(r_{v,i} - \bar{r}_v)^2}} \quad (3)$$

$$\begin{aligned}\text{sim}_{\text{cosine}}(U_u, U_v) = \cos(U_u, U_v) &= \frac{U_u \cdot U_v}{\|U_u\| \cdot \|U_v\|} \\ &= \frac{\sum_{i=0}^{n-1} r_{u,i} \cdot r_{v,i}}{\sqrt{\sum_{i=0}^{n-1} r_{u,i}^2 \cdot \sum_{i=0}^{n-1} r_{v,i}^2}}\end{aligned} \quad (4)$$

Trong đó:

- $r_{u,i}$ và $r_{v,i}$ là giá trị mà người dùng u và v đánh giá cho món ăn i.
- $\bar{r}_u$ và $\bar{r}_v$ là giá trị trung bình đánh giá của người dùng u và v tương ứng.

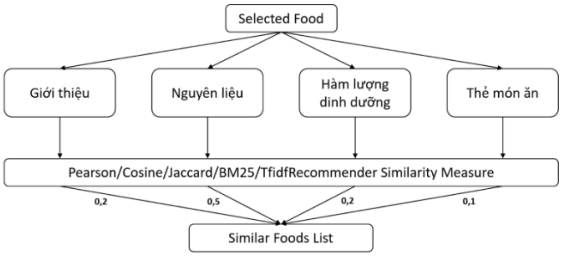
4.2 Content-based Filtering

Hệ thống khuyến nghị dựa trên nội dung được thiết kế nhằm đề xuất cho người dùng các sản phẩm họ có thể quan tâm, dựa trên đặc điểm và thuộc tính của những sản phẩm đã được họ thích trước đó. Áp dụng ý tưởng này vào nghiên cứu, chúng tôi sử dụng bộ dữ liệu foods_modeling, với các thuộc tính đã được vectorize, nhằm tính toán điểm tương tự một cách riêng biệt nhau thông qua 5 độ đo đã giới thiệu ở phần trên. Cụ thể, khi người dùng chọn một món ăn yêu thích và chỉ định một độ đo tương quan, hệ thống sẽ tính điểm tương tự giữa từng thuộc tính của món ăn đó với các món ăn khác trong bộ dữ liệu theo độ đo đã được chỉ định. Sau đó, chúng tôi tổng hợp các đề xuất dựa trên các thuộc tính và đưa ra danh sách khuyến nghị cho người dùng. Sự tổng hợp này sử dụng phương pháp nhân trọng số, trong đó danh sách trọng số được xác định sau quá trình thử nghiệm vét cạn. Cụ thể, chúng tôi tính toán các trọng số lần lượt từ 0.1 đến 0.9, với bước nhảy nhỏ nhất là 0.05. cho từng thuộc tính. Sau 80 lần thử nghiệm từng giá trị, kết quả tốt nhất chúng tôi thu được thể hiện trong Bảng 2 như sau:

Thuộc tính	Giới thiệu	Nguyên liệu	Hàm lượng dinh dưỡng	Thẻ món ăn
Trọng số	0.25	0.6	0.05	0.1

Bảng 2: Mô tả của bảng

Qua đó, chúng tôi nhận thấy thuộc tính nguyên liệu có khả năng thể hiện đặc trưng món ăn mạnh mẽ nhất và ngược lại là hàm lượng dinh dưỡng.Đồng thời, Hình 3 sau đây mô tả toàn bộ quy trình khuyến nghị món ăn theo phương pháp lọc dựa trên nội dung:



Hình 3: Quy trình khuyến nghị món ăn theo phương pháp Content-based Filtering.

Song, nếu người dùng chọn độ đo là tổng hợp (khác với 5 độ đo trong Hình 3 và đã được trình bày trong Phần 4.1.1) thì hệ thống sẽ thực hiện 5 quy trình như trên song song với nhau, tương ứng với từng độ đo.

4.3 Collaborative Filtering

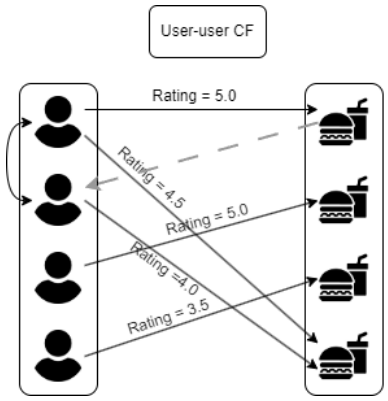
Lọc cộng tác là một trong những phương pháp phổ biến được sử dụng để thiết kế hệ khuyến nghị. Đây là phương pháp giúp cá nhân hóa trải nghiệm người dùng trên các nền tảng trực tuyến như các trang thông tin thương mại điện tử, các dịch vụ trực tuyến, hay các hệ thống đề xuất nội dung. Lọc cộng tác có thể triển khai mà không cần những dữ liệu mô tả về sản phẩm cũng như có độ tin cậy tương đối cao. Tuy nhiên, một số thách thức lớn trong quá trình xây dựng các phương pháp lọc cộng tác đó là vấn đề dữ liệu thưa thớt trong ma trận đánh giá của người dùng đối với sản phẩm. Bên cạnh đó, phương pháp này cũng phải đối mặt với vấn đề "khởi động lạnh"(cold start), khi không có đủ dữ liệu để tạo ra dự đoán cho người dùng mới hoặc sản phẩm mới.

Để tính toán độ tương đồng giữa các người dùng hoặc giữa các sản phẩm, chúng tôi sử lần lượt sử dụng hai độ đo, như đã trình bày trong Phần 4.1.2, bao gồm Cosine và Pearson. Đồng thời, các hướng tiếp cận thường được sử dụng trong phương pháp lọc cộng tác là memory-based và model-based. Trong báo cáo này, chúng tôi chỉ sử dụng phương pháp memory-based gồm:

4.3.1 User-user Collaborative Filtering

User-user Collaborative Filtering sẽ tập trung vào sự tương đồng giữa những người dùng với nhau, từ đó có thể đưa ra đánh giá sản phẩm gợi ý cho người dùng dựa trên những người dùng có độ tương đồng cao nhất. Ý tưởng cơ bản ở đây là xác định những người dùng tương tự người dùng mục tiêu A và đề xuất sản phẩm bằng cách tính toán độ tương đồng giữa dùng A và các người dùng khác. Ví dụ người dùng A và người dùng B đều rating danh sách các món ăn, B đã rating cho món ăn X, còn A chưa rating cho món ăn X, thì người ta có thể dựa vào rating của người dùng B trên món ăn X để dự đoán đánh giá của người dùng A đối với món ăn này.

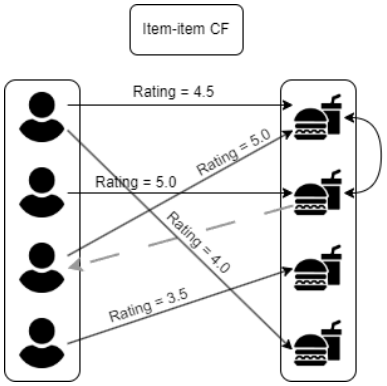
Sự tương đồng giữa người dùng với nhau được tính toán bằng độ đo Cosine (Công thức (1)) và Pearson (Công thức (2)). Hình 4 minh họa về User-user Collaborative Filtering.



Hình 4: User-user Collaborative Filtering.

4.3.2 Item-item Collaborative Filtering

Thay vì dựa vào thông tin người dùng, Item-item Collaborative Filtering dựa vào độ tương đồng giữa các sản phẩm để dự đoán cho người dùng dựa trên đánh giá của người dùng đó với các sản phẩm tương tự. Ví dụ để dự đoán đánh giá của người dùng A đối với món ăn X, quá trình sẽ bắt đầu bằng việc xác định tập S gồm các món ăn tương đồng với món ăn X. Các đánh giá trong danh sách món ăn S mà được người dùng A đánh giá sẽ được sử dụng để dự đoán xem người dùng A có thích món ăn X hay không. Tương tự như vậy, đánh giá của người dùng A trên các món ăn tương tự như Y và Z cũng có thể sử dụng để dự đoán đánh giá của người dùng B trên món ăn T. Độ tương đồng được sử dụng ở đây cũng tương tự như phương pháp User-user Collaborative Filtering, là độ đo cosine similarity (Công thức (1)) hoặc pearson similarity (Công thức (2)). Hình 5 sau đây minh họa về Item-item Collaborative Filtering.



Hình 5: Item-item Collaborative Filtering

5 Thực nghiệm và phân tích kết quả

5.1 Độ đo đánh giá

5.1.1 Content-based Filtering

Mặc dù mang lại những ưu điểm trong việc đề xuất dựa trên đặc tính nội dung của món ăn, tuy nhiên phương pháp Content-based Filtering thường đối mặt với thách thức đánh giá chính xác do hầu hết các bộ dữ liệu không đi kèm với nền tảng sự thật (ground truth) cụ thể. Để giải quyết vấn đề này, chúng tôi quyết định thực hiện gán nhãn trên khoảng 200 món ăn, chiếm 5% tổng số dữ liệu. Quá trình này được thực hiện bởi các thành viên trong nhóm, những người sẽ đánh giá sự tương đồng giữa món ăn hiện tại và món ăn được đề xuất dựa trên mối liên hệ giữa nguyên liệu của các món ăn và các thông tin phụ trợ đi kèm. Sau đó chúng tôi sẽ chọn top 5 món ăn đề xuất liên quan nhất, đánh nhãn "recommend" và sắp xếp theo thứ tự để so sánh với kết quả khuyến nghị của hệ thống nhằm đánh giá mức độ hiệu quả, sử dụng các độ đo Precision@K và Mean Reciprocal Rank (MRR): Cụ thể. Precision@K là tỷ lệ món ăn được đánh nhãn đề xuất trong top N món ăn khuyến nghị, công thức tính Precision@K như sau:

$$Precision@K = \frac{\text{Number of items labeled "recommend"in the top k}}{k} \tag{5}$$

Khác với Precision@K khi độ đo này không quan tâm đến thứ tự món ăn được khuyến nghị, MRR tìm vị trí của món ăn đầu tiên của tập thực tế (ground truth) trong tập dự đoán và tính trung bình các thứ hạng này. Công thức tính của MRR có thể được diễn giải như sau:

$$MRR = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\text{rank}_i} \tag{6}$$

Trong đó, n là số lượng món ăn được đánh giá và rank_i là vị trí của món ăn đầu tiên của tập thực tế trong tập dự đoán. Như vậy có thể thấy, mục tiêu của hệ thống khuyến nghị theo phương pháp lọc dựa trên nội dung là đạt được giá trị tối đa của 2 thang đo đánh giá MRR và Precision@K.

5.1.2 Collaborative Filtering

Để tiến hành đánh giá phương pháp lọc cộng tác, chúng tôi tiến hành so sánh kết quả dự đoán điểm rating của hệ thống và điểm rating thực tế đối với 200 món ăn. Cụ thể mỗi món ăn sẽ có 1 điểm rating và mỗi điểm thể hiện đánh giá của một người dùng, được tách ra từ tập dữ liệu ratings gốc với tỉ lệ xấp xỉ 1:900, tạo thành tập kiểm thử. Trong tập dữ liệu ratings gốc, món ăn được đánh giá ít nhất là 27 lần và người dùng đánh giá ít nhất là 1641 món ăn. Do đó, xác suất một người dùng hoặc món ăn chỉ xuất hiện trong tập kiểm thử mà không có trong tập ratings là rất thấp. Ngoài ra, sau khi tiến hành trích xuất, chúng tôi đã trực quan hóa và kiểm tra nhằm đảm bảo không có người dùng hoặc món ăn nào chỉ xuất hiện trong tập kiểm thử mà không xuất hiện trong tập ratings.

Do thiếu sự cân bằng giữa số lượng đánh giá được đề xuất và số lượng đánh giá thực tế người dùng thích dẫn đến việc sử dụng độ đo chính xác (Accuracy) có thể đạt hiệu quả không cao và không phản ánh không đúng hiệu suất của phương pháp. Vì vậy, chúng tôi quyết định sử dụng 4 độ đo là Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) và Normalized Mean Absolute Error (NMAE) nhằm đánh giá hệ khuyến nghị áp dụng phương pháp Collaborative Filtering. Các độ đo được định nghĩa như sau:

$$\text{MSE} = \frac{1}{n}\sum_{(u,i)}(\hat{r}_{u,i} - r_{u,i})^2 \tag{7}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{(u,i)}(\hat{r}_{u,i} - r_{u,i})^2} \tag{8}$$

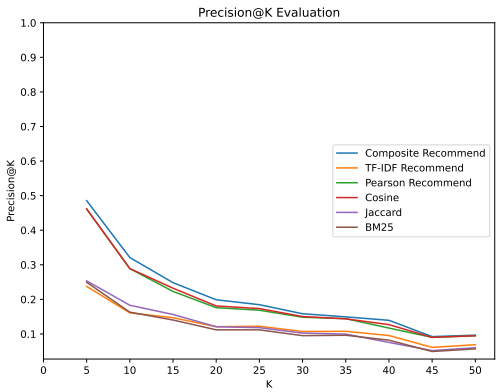
$$\text{MAE} = \frac{1}{n}\sum_{(u,i)}|\hat{r}_{u,i} - r_{u,i}| \tag{9}$$

$$\text{NMAE} = \frac{\text{MAE}}{r_{\max} - r_{\min}} \tag{10}$$

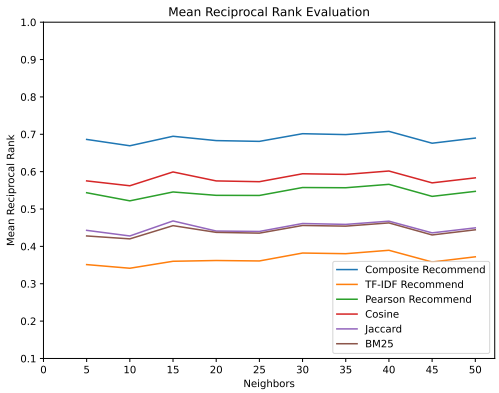
Trong đó:

- $\hat{r}_{u,i}$ là giá trị đánh giá dự đoán của người dùng u cho món ăn i.
- $r_{u,i}$ là giá trị đánh giá thực tế của người dùng u cho món ăn i.
- n là số lượng điểm dữ liệu được sử dụng để đánh giá.
- $r_{\max}$ và $r_{\min}$ lần lượt là giá trị đánh giá cao nhất và thấp nhất trong tập dữ liệu được sử dụng để đánh giá.

5.2 Content-based Filtering



Hình 6: Đánh giá các phương pháp với Precison@K.



Hình 7: Đánh giá các phương pháp với Mean Reciprocal Rank.

Để đưa ra các khuyến nghị chính xác cho người dùng mà không làm tăng chi phí và thời gian tính

toán, chúng tôi đã giới hạn số lượng lân cận từ 5 đến 50, với bước nhảy là 5 trong quá trình thực nghiệm. Các kết quả đánh giá được thể hiện thông qua Hình 6 và Hình 7.

Nhìn chung, hệ thống khuyến nghị dựa trên nội dung vẫn chưa đạt được hiệu suất cao, với giá trị cao nhất của Precision@K chỉ ở mức trung bình và giá trị MRR luôn nằm trong khoảng từ 0,35 đến 0,7. Tuy nhiên, không thể phủ nhận rằng độ đo tổng hợp đã thể hiện rất tốt, khi đứng đầu cả kết quả đánh giá MRR và Precision@K. Ngược lại, Cosine kết hợp TF-IDF là độ đo kém hiệu quả nhất, khi thấp nhất trong các kết quả đánh giá. Ngoài ra, tất cả các độ đo khác đều không có sự nổi bật đặc biệt và kết quả nằm trong ngưỡng chấp nhận được.

Bên cạnh đó, dựa trên biểu đồ, có thể thấy số lượng lân cận lý tưởng nhất cho hệ thống khuyến nghị của chúng tôi là 15. Con số này cũng được chúng tôi sử dụng trong quá trình trực quan hệ thống, triển khai ứng dụng. Đồng thời, kết quả tốt nhất của các độ đo tương quan được trình bày chi tiết trong Bảng 3:

Bàn luận về kết quả này, chúng tôi cho rằng có 2 nguyên nhân chính dẫn đến việc thể hiện chưa tốt của các phương pháp như sau:

- Chúng tôi dự đoán rằng các thuộc tính trong bộ dữ liệu vẫn còn nhiều giá trị nhiễu, chưa thể đại diện cho các món ăn riêng biệt, dẫn đến không thể trích xuất đặc trưng hiệu quả các thuộc tính.

- Kết quả đánh giá có thể phần nào phụ thuộc vào quá trình gán nhãn ground truth. Một lần nữa, chúng tôi nghĩ rằng, việc gán nhãn dựa trên nhận định của con người, hay nói cách khác là yếu tố chủ quan, đã ảnh hưởng đến kết quả đánh giá các phương pháp.

5.3 Collaborative Filtering

Trong quá trình thực nghiệm cho khuyến nghị lọc cộng tác, chúng tôi kết hợp hai phương pháp là User-user Collaborative Filtering và Item-item Collaborative Filtering với hai độ tương đồng là Cosine và Pearson, tạo thành 4 mô hình. Sau khi tiến hành thực nghiệm với số lượng lân cận là 10 và so sánh với 200 điểm dữ liệu trong tập đánh giá, chúng tôi thu được kết quả ở Bảng 4.

Từ Bảng 4, chúng tôi nhận thấy rằng phương pháp User-user Cosine đạt kết quả tốt nhất với MSE là 4.2581 và RMSE là 2.0635. Ngược lại, Item-item Cosine cho kết quả tốt nhất ở MAE là 1.6902 và NMAE là 0.338. Trong khi đó, phương pháp Item-item Pearson đều cho kết quả tệ nhất ở cả bốn độ đo với MSE là 6.5245, RMSE là 2.5543, MAE là 2.1250 và NMAE là 0.4250.

6 Trực quan hệ thống

Chúng tôi quyết định sử dụng Streamlit, một framework khá thông dụng, nhận được sự quan tâm lớn từ cộng đồng do khả năng mạnh mẽ và sự tiện lợi của nó, nhằm tối ưu hóa quá trình phát triển hệ thống và dễ dàng chuyển đổi thành ứng dụng web. Đồng thời, vì phương pháp khuyến nghị dựa trên nội dung và phương pháp lọc cộng tác dựa trên sản phẩm có đầu vào khá giống nhau và cùng khác với phương pháp User-user Collaborative Filtering, chúng tôi chia ứng dụng thành 2 trang riêng biệt. Chi tiết về giao diện hệ thống có thể truy cập thông qua Youtube([3]).

6.1 Content-based Filtering và Item-item Collaborative Filtering

Đối với phương pháp khuyến nghị dựa trên nội dung và lọc cộng tác dựa trên sản phẩm, người dùng cần cung cấp thông tin về một món ăn mà họ yêu thích. Từ đó, thông tin về các thuộc tính liên quan được hệ thống thu thập và tính toán sự tương đồng với các món ăn khác, sử dụng một trong sáu độ đo khác nhau. Sau cùng, khi tính toán xong, người dùng có thể lọc lại các món ăn được đề xuất thông qua 4 tiêu chí, bao gồm khẩu phần, thời gian nấu, hàm lượng calories và loại món ăn. Những tiêu chí này được tạo nên từ các thuộc tính trong bộ dữ liệu, giúp người dùng có đa dạng sự lựa chọn và kết quả đề xuất được cá nhân hóa.

6.2 User-user Collaborative Filtering

Phương pháp này yêu cầu người dùng chọn số lượng không giới hạn các món ăn mà họ thích. Hệ thống sẽ tính điểm đánh giá cho các món ăn này và so sánh sự tương đồng với các người dùng khác để đưa ra các đề xuất phù hợp. Hiện tại, điểm đánh giá được xác định ngẫu nhiên từ 4 đến 5 điểm để tối ưu hóa tính trực quan trong ứng dụng, mặc dù có hạn chế nhất định, nhưng cách tính điểm này có khả năng giúp giao diện thân thiện và gọn gàng.

7 Kết luận

Trong báo cáo này, chúng tôi đã thu thập, xây dựng và trình bày bộ dữ liệu Vietnamese Food Dataset, một bộ dữ mới cho bài toán khuyến nghị món ăn

[3] https://www.youtube.com/watch?v=nM16OfhCrrA

Kết quả / Độ đo	Kết hợp	TF-IDF + Cosine	CountVectorizer + Cosine	CountVectorizer + Pearson	CountVectorizer + Jaccard	CountVectorizer + BM25
MRR	70.77%	38.9%	56.03%	60.1%	46.5%	45.8%
Precision@K	48.5%	23.6%	46.1%	45.9%	25.3%	24.4%

Bảng 3: Kết quả các mô hình theo từng độ đo

	MSE	RMSE	MAE	NMAE
User-user Cosine	**4.2581**	**2.0635**	1.7228	0.3445
User-user Pearson	5.4402	2.3324	1.9130	0.3826
Item-item Cosine	4.6168	2.1486	**1.6902**	**0.3380**
Item-item Pearson	6.5245	2.5543	2.1250	0.4250

Bảng 4: Kết quả các mô hình theo từng độ đo

Việt Nam. Bộ dữ liệu gồm tập foods với hơn 4000 dòng dữ liệu và 15 thuộc tính và tập ratings với hơn 180000 đánh giá.

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 Cosine 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 4,2581 MSE, 2,0635 RMSE, 1,6902 MAE và 0,3380 NMAE đối với phương pháp lọc cộng tác và 48.5% Precision@k và 70.77% MRR đối với lọc trên nội dung.

Hướng phát triển trong tương lai:

- Bộ dữ liệu: Thu thập thêm dữ liệu là thông tin món ăn từ các trang web Món Ngon Mỗi Ngày (monngonmoingay.com) và Cooky.vn (cooky.vn), cùng với đó là thu thập thêm các thuộc tính mới như: bình luận của người dùng, rating trên từng khía cạnh, giá cả, lịch sử tìm kiếm . . . để cho ra bộ dữ liệu đa dạng thông tin hơn.

- Mô hình: Áp dụng các phương pháp, kỹ thuật khuyến nghị khác như (Aggarwal and Aggarwal, 2016): 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.

References

Charu C Aggarwal and Charu C Aggarwal. 2016. An introduction to recommender systems. *Recommender systems: The textbook*, pages 1–28.

Hayder Hussein Al-Chalabi and Mahdi Nsaif Jasim. 2022. Food recommendation system based on data clustering techniques and user nutrition records. In *International Conference on New Trends in Information and Communications Technology Applications*, pages 139–161. Springer.

J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. Collaborative filtering recommender systems. In *The adaptive web: methods and strategies of web personalization*, pages 291–324. Springer.

Jieun Son and Seoung Bum Kim. 2017. Content-based filtering for recommendation systems using multiattribute networks. *Expert Systems with Applications*, 89:404–412.

Raciel Yera Toledo, Ahmad A Alzahrani, and Luis Martinez. 2019. A food recommender system considering nutritional information and user preferences. *IEEE Access*, 7:96695–96711.

Thi Ngoc Trang Tran, Müslüm Atas, Alexander Felfernig, and Martin Stettinger. 2018. An overview of recommender systems in the healthy food domain. *Journal of Intelligent Information Systems*, 50:501–526.

Christoph Trattner and David Elsweiler. 2017. Food recommender systems: important contributions, challenges and future research directions. *arXiv preprint arXiv:1711.02760*.