Hệ thống gợi ý phim dựa trên lọc công tác và phân tích dự đoán đánh giá người dùng

1st Trương Nhật Quang
Trường Đại Học Công Nghệ Thông Tin
Thành Phố Hồ Chí Minh, Việt Nam
22521294@gm.uit.edu.vn

2nd Đỗ Tuấn Trực
Trường Đại Học Công Nghệ Thông Tin
Thành Phố Hồ Chí Minh, Việt Nam
22521548@gm.uit.edu.vn

3rd Nguyễn Khánh Minh
Trường Đại Học Công Nghệ Thông Tin
Thành Phố Hồ Chí Minh, Việt Nam
20520635@gm.uit.edu.vn

4th Trương Lưu Song Tâm
Trường Đại Học Công Nghệ Thông Tin
Thành Phố Hồ Chí Minh, Việt Nam
22521294@gm.uit.edu.vn

5th Nguyễn Đức Tài
Trường Đại Học Công Nghệ Thông Tin
Thành Phố Hồ Chí Minh, Việt Nam
22521277@gm.uit.edu.vn

GVHD: ThS.Nguyễn Văn Kiệt
Trường Đại Học Công Nghệ Thông Tin
Thành Phố Hồ Chí Minh, Việt Nam

***Tóm tắt nội dung*—Phim ảnh là một phần không thể thiếu trong cuộc sống hàng ngày, bởi vì nó không chỉ mang lại niềm vui mà còn tạo ra những giây phút thư giãn, giải trí cho mọi người. Nhận thức được tầm quan trọng này, chúng tôi muốn giới thiệu một hệ thống đề xuất phim tiên tiến, nhằm giúp người dùng dễ dàng tìm kiếm và lựa chọn bộ phim phù hợp với sở thích của mình. Hệ thống của chúng tôi sử dụng là Collaborative Filtering, một hệ thống phổ biến trong việc cung cấp các gợi ý cá nhân hóa dựa trên sở thích và hành vi của người dùng. Phục vụ cho cả người dùng mới và người dùng hiện tại. Được thiết kế để mang đến sự mới mẻ và cá nhân hóa, hệ thống này tạo ra các gợi ý phim phù hợp với sở thích riêng biệt của từng người. Hệ thống đề xuất phim của chúng tôi hướng tới việc mang lại sự tiện lợi và phù hợp với sở thích cá nhân của người dùng. Bằng cách khai thác thông tin từ cơ sở dữ liệu phim và hành vi của người dùng, hệ thống của chúng tôi tạo ra những gợi ý phim đa dạng và hấp dẫn, giúp người dùng khám phá ra những tác phẩm điện ảnh mới mẻ mà họ chưa từng biết đến. Chúng tôi tin rằng hệ thống này sẽ mang đến cho người dùng một trải nghiệm xem phim độc đáo và thú vị.**

***Từ khóa* — Hệ thống đề xuất, Phim, Lọc cộng tác, Học máy.**

I. GIỚI THIỆU

Kể từ khi được phát minh, Internet đã phát triển nhanh chóng và tiếp tục mở rộng mỗi ngày. Sự phong phú của thông tin trực tuyến khiến việc tìm kiếm thông tin phù hợp trở nên khó khăn hơn bao giờ hết [**1**]. Với sự bùng nổ của ngành công nghiệp điện ảnh và sự đa dạng ngày càng tăng của các thể loại phim, việc tìm kiếm những bộ phim phù hợp có thể trở thành một nhiệm vụ mất thời gian và phức tạp.

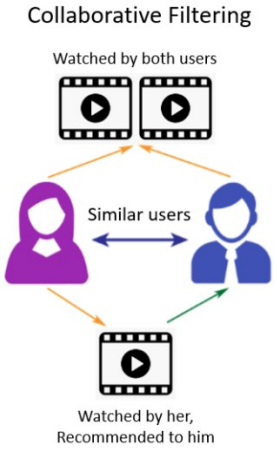
Điều này làm cho hệ thống gợi ý phim trở thành một công cụ quan trọng để giúp người dùng dễ dàng tìm kiếm và khám phá các bộ phim được chúng tôi đề cập trong bài nghiên cứu này bởi vì nó có thể giúp người dùng dễ dàng tìm kiếm những bộ phim một cách thuận tiện và hiệu quả.

Hệ thống gợi ý phim của chúng tôi sử dụng phương pháp collaborative filtering, dựa trên hành vi và sở thích của người dùng khác có cùng quan điểm. Hệ thống này sẽ đề xuất những bộ phim mà người dùng có thể quan tâm, dựa trên đánh giá và lựa chọn của những người dùng tương tự. Chỉ cần cung cấp cho chúng tôi thông tin về những bộ phim bạn đã thích, hệ thống sẽ tự động tạo ra danh sách gợi ý phim dựa trên sở thích chung của cộng đồng người dùng. Từ những bộ phim hành động gay cấn đến những bộ phim lãng mạn ngọt ngào, chúng tôi sẽ giúp bạn khám phá những tác phẩm điện ảnh tuyệt vời mà bạn có thể chưa biết đến.

Phương pháp collaborative filtering không chỉ giúp bạn tìm thấy những bộ phim mới dựa trên sở thích cá nhân mà còn mang đến cơ hội khám phá các tác phẩm có phong cách hoặc chủ đề tương tự mà bạn yêu thích. Nhờ vào khả năng phân tích dữ liệu người dùng một cách chi tiết và chính xác, hệ thống của chúng tôi đảm bảo mang đến cho bạn những đề xuất phim phù hợp và hấp dẫn nhất.

Động lực đằng sau việc áp dụng phương pháp lọc cộng tác là các đề xuất được thực hiện dựa trên sự tương tác của người dùng tương tự. Nó giải quyết vấn đề quá tải nội dung bằng cách lọc qua thư viện phim khổng lồ để gợi ý cho người dùng những lựa chọn có khả năng thu hút sự quan tâm của họ, từ đó cải thiện trải nghiệm người dùng tổng thể.
Một cách tổng quát, nghiên cứu này có các đóng góp:

- **Nâng cao chất lượng dịch vụ:** Hệ thống gợi ý phim này đóng góp vào việc nâng cao chất lượng dịch vụ hơn nữa có thể giải quyết tình trạng mà người dùng không thể quyết định chọn xem gì do có quá nhiều lựa chọn.
- **Hỗ trợ nhà sản xuất và phân phối:** Ngoài ra, hệ thống cũng giúp các nhà sản xuất và phân phối phim hiểu rõ hơn về sở thích của khán giả, từ đó có chiến lược sản xuất và marketing hiệu quả hơn.



Hình 1: Mô tả quy trình Collaborative Filtering recommendation movie system của nhóm

- **Mở rộng ứng dụng:** Hơn thế nữa, không chỉ phát triển ở trên website phim mà có thể mở rộng ra những ý tưởng góp phần tối ưu hơn các trang web cũng sử dụng hệ thống gợi ý như Youtube, Tiktok, Facebook,...thay vì đề xuất những bộ phim phù hợp, thì có thể gợi ý cho người dùng những video, bài viết phù hợp cho người dùng. Thậm chí có thể loại bỏ những nội dung nhạy cảm lách luật gây hại cho người dùng trong tương lai.

Trong phần còn lại bài báo này, chúng tôi sẽ trình bày tiếp đến **Phần II Các Nghiên cứu liên quan** và các cơ sở lý thuyết được áp dụng, sau đó là **Phần III là Bộ dữ liệu** sử dụng trong nghiên cứu này, **Phần IV Phương pháp sử dụng cho nghiên cứu** gồm các ý tưởng được để thực hiện thiết kế hệ thống dựa trên collaborative filtering, **Phần V là Các thực nghiệm nghiên cứu và kết quả đánh giá**. Cuối cùng **Phần VI Kết luận và định hướng trong tương lai**.

II. Các Nghiên cứu liên quan

Dựa trên bài báo "Movie recommendation and sentiment analysis using machine learning" của tác giả N. Pavitha [**2**], chúng tôi nhận thấy rằng việc tích hợp hệ thống gợi ý phim với phân tích cảm xúc có thể cung cấp những gợi ý chính xác và cá nhân hóa hơn. Nghiên cứu này đã khám phá và áp dụng các thuật toán học máy tiên tiến để không chỉ gợi ý phim dựa trên sở thích của người dùng mà còn phân tích phản hồi cảm xúc từ các đánh giá của người dùng để cải thiện chất lượng gợi ý.

A. Hướng phát triển

Dựa theo bài báo tham khảo [**2**] và tham khảo thêm các notebook của kaggle [**3**][**4**][**5**], chúng tôi đã cải tiến một hệ thống đề xuất phim sử dụng phương pháp lọc cộng tác (Collaborative Filtering). Với việc đưa ra hai mô hình: "Dự đoán bộ phim tương đồng dựa trên sở thích cá nhân" và "Dự đoán đánh giá(rating) cho từng bộ phim đối với những người chưa xem"

B. Cơ sở lý thuyết

Bài báo đã cung cấp cho chúng tôi một cái nhìn sâu sắc về việc sử dụng Độ Tương Đồng Cosine để gợi ý phim dựa trên sở thích của người dùng. Chúng tôi sử dụng hai thuật toán học máy SVD (Singular Value Decomposition) và KNN (K-nearest neighbors) để giúp xây dựng ý tưởng hướng phát triển trên.

SVD (Singular Value Decomposition) SVD là một phép phân rã ma trận thành một phép xoay, theo sau là một phép co giãn và một phép xoay khác. Nó tổng quát hóa phân rã giá trị riêng của một ma trận vuông bình thường với một cơ sở riêng orthonormal cho bất kỳ ma trận nào. Cụ thể, phân rã giá trị đơn lẻ của một ma trận phức là một phân rã có dạng

$$A = U\Sigma V^* \tag{1}$$

trong đó U là một ma trận đơn vị phức, Σ là một ma trận chéo hình chữ nhật với các số thực không âm trên đường chéo, V là một ma trận đơn vị phức, và V^* là chuyển vị phức liên hợp của V. Các mục nhập chéo của Σ được xác định duy nhất bởi A và được biết đến là các giá trị đơn lẻ của A. Các cột của U và các cột của V được gọi là các vector đơn lẻ trái và phải của A, tương ứng.

KNN (K-nearest neighbors) KNN là một thuật toán học có giám sát không tham số, được phát triển đầu tiên bởi Evelyn Fix và Joseph Hodges vào năm 1951, và sau đó được mở rộng bởi Thomas Cover. Nó được sử dụng cho cả phân loại và hồi quy. Trong cả hai trường hợp, đầu vào bao gồm các ví dụ đào tạo gần nhất k trong một tập dữ liệu. Đầu ra phụ thuộc vào việc KNN được sử dụng cho phân loại hay hồi quy:

- Trong phân loại KNN, đầu ra là thành viên của một lớp. Một đối tượng được phân loại bằng cách bỏ phiếu đa số của các hàng xóm của nó, với đối tượng được gán cho lớp phổ biến nhất trong số các hàng xóm gần nhất k của nó (k là một số nguyên dương, thường nhỏ).
- Trong hồi quy KNN, đầu ra là giá trị thuộc tính cho đối tượng. Giá trị này là trung bình của các giá trị của k hàng xóm gần nhất.

Cosine Similarity hay độ tương đồng cosin giữa hai đối tượng đo lường góc cosin giữa hai vector (đối tượng). Nó so sánh hai tài liệu trên một thang đo chuẩn hóa. Điều này có thể được thực hiện bằng cách tìm tích vô hướng giữa hai vector.

Công thức tính tương đồng cosin giữa hai vector $\mathbf{a}$ và $\mathbf{b}$ được cho bởi:

$$Cosinesimilarity(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|} \tag{2}$$

trong đó $\mathbf{a} \cdot \mathbf{b}$ là tích vô hướng của hai vector $\mathbf{a}$ và $\mathbf{b}$, và $\|\mathbf{a}\|$ và $\|\mathbf{b}\|$ là các norm Euclid của vector $\mathbf{a}$ và $\mathbf{b}$, tương ứng.

Cosine similarity có giá trị nằm trong khoảng từ -1 đến 1, trong đó 1 chỉ ra rằng hai văn bản giống nhau hoàn toàn, 0 có nghĩa là không có sự tương tự, và -1 cho thấy hai văn bản hoàn toàn khác nhau. Hình 1 trên cho thấy, góc giữa v1 và v2 là. Góc giữa hai vectơ càng nhỏ thì độ tương đồng càng lớn.

Điều đó có nghĩa là nếu góc giữa hai vectơ nhỏ thì chúng gần như giống nhau và nếu góc giữa hai vectơ lớn thì các vectơ rất khác nhau.

C. Ứng dụng thực tế

Chúng tôi đã áp dụng các kiến thức đã học từ bài báo vào việc xây dựng hệ thống đề xuất phim của mình. Hệ thống này không chỉ giúp người dùng tìm phim phù hợp mà còn phân loại đánh giá thành tích cực hoặc tiêu cực. Điều này góp phần nâng cao trải nghiệm người dùng khi lựa chọn phim. Nó không chỉ đóng vai trò là một thanh tìm kiếm mà còn có thể đưa ra những bộ phim gợi ý ngay sau khi họ vừa xem xong một bộ phim nào đó ở mục "các bộ phim liên quan" ngay dưới bộ phim đang xem.

III. Bộ dữ liệu

A. Thu thập bộ dữ liệu

Bộ dữ liệu được lấy từ Kaggle.com có tên là rating.csv và movie.csv được sử dụng cho nghiên cứu.

B. Mô tả bộ dữ liệu

Dữ liệu movies chứa thông tin về các bộ phim và được lưu trữ trong tập tin movies.csv(có kích thước là 9742 dòng). Bộ dữ liệu này gồm ba cột chính:

- movieId: Là một số nguyên duy nhất xác định từng bộ phim.(int64)
- title: Là tiêu đề của bộ phim.(object)
- genres: Là một chuỗi mô tả các thể loại của bộ phim.(object)

Dữ liệu ratings chứa thông tin về đánh giá của người dùng cho các bộ phim và được lưu trữ trong tập tin ratings.csv (có kích thước là 100836 dòng). Bộ dữ liệu này gồm ba cột chính:

- userId: Là một số nguyên duy nhất xác định từng người dùng.(int64)
- movieId: Là một số nguyên duy nhất xác định từng bộ phim, tương ứng với cột movieId trong tập dữ liệu movies.(int64)
- rating: Là đánh giá của người dùng cho bộ phim, thường nằm trong thang điểm từ 1 đến 5.(float64)

Việc có chung cột movieId trong cả hai bộ dữ liệu movies và ratings là rất quan trọng vì nó đóng vai trò như một khóa ngoại (foreign key) liên kết hai bảng này với nhau. Giúp tối ưu hóa hiệu suất truy vấn và cải thiện tốc độ xử lý dữ liệu. Điều này đặc biệt quan trọng khi làm việc với các bộ dữ liệu lớn và phức tạp trong các ứng dụng như hệ thống gợi ý phim.

IV. Phương pháp sử dụng cho nghiên cứu

A. Dự đoán bộ phim tương đồng dự trên sở thích cá nhân

Trong bối cảnh công nghệ hiện đại, hệ thống gợi ý phim là một công cụ quan trọng giúp cá nhân hóa trải nghiệm người dùng bằng cách đề xuất những bộ phim phù hợp với sở thích của họ. Có nhiều phương pháp khác nhau để xây dựng hệ thống gợi ý, trong đó các phương pháp máy học như **SVD (Singular Value Decomposition)** và **KNN (K-nearest neighbors)** được sử dụng rộng rãi. Dưới đây là cách tiếp cận và triển khai một hệ thống gợi ý phim dựa trên các kỹ thuật này:

P**hương pháp Truncated SVD kết hợp độ tương đồng cosine:** Truncated SVD là một phương pháp giảm kích thước dữ liệu, hữu ích trong việc xử lý các ma trận lớn như ma trận người dùng-phim. Chúng ta sử dụng SVD để giảm số lượng đặc trưng và sau đó tính toán độ tương đồng cosine giữa các người dùng.

Sử dụng thư viện Surprise với SVD: Thư viện surprise cung cấp các công cụ mạnh mẽ để xây dựng và đánh giá các hệ thống gợi ý. Chúng ta sẽ sử dụng thuật toán SVD của thư viện này để dự đoán đánh giá của người dùng cho các phim chưa xem và gợi ý phim dựa trên các dự đoán này.

Phương pháp KNN: là một thuật toán đơn giản nhưng hiệu quả trong việc tìm kiếm các phần tử tương tự. Chúng ta sử dụng KNN để tìm các phim tương tự dựa trên các đánh giá của người dùng.

B. Dự đoán đánh giá cho từng bộ phim đối với những người chưa xem

Việc dự đoán chính xác đánh giá của người dùng cho từng bộ phim là một nhiệm vụ quan trọng nhằm nâng cao trải nghiệm cá nhân hóa. Các hệ thống gợi ý dựa trên những đánh giá này không chỉ giúp người dùng khám phá những bộ phim mới phù hợp với sở thích của họ mà còn hỗ trợ các nhà cung cấp dịch vụ tối ưu hóa nội dung đề xuất. Để đạt được điều này, chúng ta có thể áp dụng nhiều phương pháp máy học khác nhau. Trong bài viết này, chúng tôi sẽ trình bày hai phương pháp chính: SVD (Singular Value Decomposition) và KNNBasic (K-nearest neighbors) để dự đoán đánh giá của người dùng cho từng bộ phim kết hợp sửa dụng thư viện Surprise.

Phương pháp SVD: là một thuật toán phân rã ma trận, hữu ích trong việc giảm chiều dữ liệu và phát hiện các yếu tố tiềm ẩn ảnh hưởng đến đánh giá của người dùng. Chúng tôi sử dụng thư viện surprise để triển khai phương pháp này. Dữ liệu đánh giá phim được nạp vào dưới định dạng của thư viện surprise và sau đó được chia thành tập huấn luyện và tập kiểm tra.

Phương pháp KNNBasic: là một thuật toán đơn giản nhưng hiệu quả trong việc tìm kiếm các phần tử tương tự. Chúng tôi cũng sử dụng thư viện surprise để triển khai phương pháp này. Dữ liệu được chia thành tập huấn luyện và tập kiểm tra, sau đó mô hình KNNBasic được huấn luyện và sử dụng để dự đoán đánh giá.

V. Các thực nghiệm nghiên cứu và kết quả đánh giá

A. Cài đặt thực nghiệm

1/Dự đoán bộ phim tương đồng dự trên sở thích cá nhân

Phương pháp Truncated SVD kết hợp độ tương đồng cosine: Tạo ra ma trận người dùng-phim (usermatrix). Tiếp theo, chúng ta sử dụng phương pháp Truncated SVD để giảm chiều của ma trận người dùng-phim xuống 50 chiều. Kết quả là ma trận

matrixsvd. Chúng ta sau đó tính toán độ tương đồng cosine giữa các người dùng dựa trên ma trận matrixsvd để tạo ra ma trận độ tương đồng người dùng (usersimilarity). Cuối cùng, chúng ta có thể sử dụng hàm recommendmovies để gợi ý phim cho một người dùng.

Sử dụng thư viện Surprise với SVD: Sử dụng thư viện surprise để tải dữ liệu đánh giá từ dataframe. Sau đó sử dụng phương pháp SVD với số lượng yếu tố là 100 và hệ số điều chuẩn là 0.1 để huấn luyện mô hình trên tập huấn luyện. Hàm recommendmovies nhận vào một mô hình SVD đã được huấn luyện để gợi ý cho một người dùng.

Phương pháp KNN: tạo ra ma trận người dùng-phim (usermatrix) và chuyển đổi ma trận này thành dạng thưa (sparsematrix) để tiết kiệm bộ nhớ. Sau đó huấn luyện mô hình KNN trên ma trận người dùng-phim thưa. Cuối cùng, in ra danh sách các bộ phim được gợi ý cho từng người dùng.

Qua đó chúng ta có thể xây dựng một hệ thống gợi ý phim hiệu quả và linh hoạt. Mỗi phương pháp có những ưu điểm riêng, và việc kết hợp chúng giúp tăng cường khả năng gợi ý của hệ thống.

2/Dự đoán đánh giá cho từng bộ phim đối với những người chưa xem

Phương pháp SVD: Tải dữ liệu đánh giá từ dataframe ratings vào định dạng của thư viện surprise. Sau đó chia dữ liệu thành tập huấn luyện và tập kiểm tra với tỷ lệ 80:20. Số lượng yếu tố là 100 và hệ số điều chuẩn là 0.1 để huấn luyện mô hình trên tập huấn luyện. đánh giá thực tế (actual rating), và đánh giá dự đoán (predicted rating). Sau đó chọn một bộ phim cụ thể. Nó sẽ hiển thị các dự đoán so với các giá trị thực tế của từng người dùng cho bộ phim đó

Phương pháp KNNBasic: tương tự với cách trên nhưng sẽ sử dụng phương pháp KNNBasic để huấn luyện mô hình trên tập huấn luyện. Kết quả trả về vẫn các dự đoán so với các đánh giá thực tế và dự đoán đánh giá của từng user.

B. Độ đo đánh giá

Root Mean Square Error (RMSE) là một chỉ số được sử dụng để đo lường độ chênh lệch giữa các giá trị được dự đoán bởi mô hình và các giá trị thực tế. RMSE tính toán căn bậc hai của giá trị trung bình của bình phương sai số giữa các giá trị dự đoán và giá trị thực tế.

Công thức tính RMSE:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

trong đó:

- n là số lượng mẫu.
- y_i là giá trị thực tế.
- $\hat{y}_i$ là giá trị dự đoán.

Mean Absolute Error (MAE) là một chỉ số khác để đo lường độ chênh lệch giữa các giá trị dự đoán và giá trị thực tế. MAE tính toán giá trị trung bình của các giá trị tuyệt đối của sai số giữa các giá trị dự đoán và giá trị thực tế.

Công thức tính MAE:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4}$$

trong đó:

- n là số lượng mẫu.
- y_i là giá trị thực tế.
- $\hat{y}_i$ là giá trị dự đoán.

RMSE và MAE đều là các chỉ số đánh giá hiệu suất của mô hình dự đoán, với mục tiêu là đạt được các giá trị càng thấp càng tốt để chỉ ra rằng mô hình có sai số dự đoán nhỏ và hiệu suất tốt. Trong đoạn mã trên, RMSE và MAE được sử dụng để đo lường và so sánh độ chính xác của mô hình
Khi đánh giá độ chính xác của một mô hình dự đoán rating phim với biên độ rating từ 0 đến 5, các giá trị của RMSE và MAE sẽ phản ánh sai số giữa dự đoán và thực tế. Nên biên độ giá trị của RMSE và MAE trong bài báo này sẽ là từ 0 đến 5. Cả MAE và RMSE đều có giá trị nhỏ nhất là 0, cho thấy mô hình dự đoán hoàn toàn chính xác. Cả MAE và RMSE đều có giá trị lớn nhất là 5, cho thấy sự khác biệt tối đa giữa dự đoán và giá trị thực tế. Đánh giá độ chính xác của mô hình dự đoán: giá trị càng nhỏ, mô hình càng chính xác.

C. Kết quả

1/Dự đoán bộ phim tương đồng dự trên sở thích cá nhân
Do kết quả đưa ra danh sách gợi ý phim cho từng cá nhân nên cần tính đa dạng nên theo ý kiến có nhân sẽ không tập trung vào các chỉ số chính xác mà nên tập trung vào sự phong phú của hệ thống đưa ra gợi ý cho người dùng. Có thể đối với người dùng này bộ phim được gợi ý họ không thích, nhưng với người dùng khác thì họ lại có sự tò mò về bộ phim được đề xuất mà họ chưa xem. Tuy nhiên bên cạnh đó, nhóm cũng có sử dụng SVD kết hợp với Surprise để train mô hình đưa ra hai chỉ số RMSE và MAE cho việc "Dự đoán bộ phim tương đồng dự trên sở thích cá nhân"

	movieId	title	genres
31	32	Twelve Monkeys (a.k.a. 12 Monkeys) (1995)	Mystery|Sci-Fi|Thriller
507	589	Terminator 2: Judgment Day (1991)	Action|Sci-Fi
793	1036	Die Hard (1988)	Action|Crime|Thriller
902	1200	Aliens (1986)	Action|Adventure|Horror|Sci-Fi
916	1215	Army of Darkness (1993)	Action|Adventure|Comedy|Fantasy|Horror
1067	1387	Jaws (1975)	Action|Horror
1211	1610	Hunt for Red October, The (1990)	Action|Adventure|Thriller
1404	1923	There's Something About Mary (1998)	Comedy|Romance
2078	2762	Sixth Sense, The (1999)	Drama|Horror|Mystery
2393	3175	Galaxy Quest (1999)	Adventure|Comedy|Sci-Fi

Hình 2: SVD với cosine similarity

	movieId	title	genres
277	318	Shawshank Redemption, The (1994)	Crime|Drama
602	750	Dr. Strangelove or: How I Learned to Stop Worr...	Comedy|War
686	904	Rear Window (1954)	Mystery|Thriller
694	912	Casablanca (1942)	Drama|Romance
841	1104	Streetcar Named Desire, A (1951)	Drama
906	1204	Lawrence of Arabia (1962)	Adventure|Drama|War
949	1250	Bridge on the River Kwai, The (1957)	Adventure|Drama|War
2462	3275	Boondock Saints, The (2000)	Action|Crime|Drama|Thriller
4909	7361	Eternal Sunshine of the Spotless Mind (2004)	Drama|Romance|Sci-Fi
6648	56782	There Will Be Blood (2007)	Drama|Western

Hình 3: SVD với thư viện Surprise

	movieId	title	genres
18	19	Ace Ventura: When Nature Calls (1995)	Comedy
38	42	Dead Presidents (1995)	Action|Crime|Drama
44	48	Pocahontas (1995)	Animation|Children|Drama|Musical|Romance
56	63	Don't Be a Menace to South Central While Drink...	Comedy|Crime
90	102	Mr. Wrong (1996)	Comedy
265	305	Ready to Wear (Pret-A-Porter) (1994)	Comedy
287	329	Star Trek: Generations (1994)	Adventure|Drama|Sci-Fi
312	354	Cobb (1994)	Drama
451	516	Renaissance Man (1994)	Comedy|Drama
468	535	Short Cuts (1993)	Drama

Hình 4: KNN với cosine similarity

Qua đó ta thấy được sự đa dạng của từng phương pháp đề xuất phim, những gợi ý phong phú mang lại cho người dùng. Do phương pháp SVD với cosine similarity, phương pháp KNN tập trung vào việc tìm phim tương tự để đề xuất, không phải là dự đoán chính xác điểm số. Mục tiêu là tìm các phim mà người dùng có thể thích dựa trên các phim tương tự mà họ đã xem. Còn SVD với thư viện Surprise do bao gồm việc huấn luyện mô hình SVD, dự đoán điểm số và đánh giá độ chính xác của mô hình dự đoán. RMSE và MAE là các chỉ số quan trọng để đánh giá hiệu suất của mô hình dự đoán.
Những chỉ số này cho thấy mô hình đang hoạt động khá hiệu quả trong bài toán dự đoán đánh giá phim. Điều này có nghĩa là sự chênh lệch giữa dự đoán và đánh giá thực tế của người dùng là tương đối tốt. Việc sử dụng phương pháp collaborative filtering giúp mô hình đưa ra các gợi ý phim dựa trên sở thích và đánh giá của người dùng có cùng gu phim. Nhờ đó, những bộ phim được gợi ý có xu hướng tương đồng với thể loại phim mà người dùng mong muốn và có đánh giá gần với các phim mà họ đã xem trước đó. Điều này làm tăng khả năng người dùng sẽ yêu thích và đồng ý với các gợi ý phim của mô hình, tạo ra trải nghiệm cá nhân hóa và đáng tin cậy hơn.

2/Dự đoán đánh giá cho từng bộ phim đối với những người chưa xem

Sau khi cài đặt và cho chạy thực nghiệm, chúng tôi thu được kết quả sau.

RMSE: 0.8857
MAE: 0.6825

	userId	movieId	actual_rating	predicted_rating	title	genres
0	39	1196	5.0	4.305388	Star Wars: Episode V - The Empire Strikes Back...	Action|Adventure|Sci-Fi
1	215	1196	4.5	4.188010	Star Wars: Episode V - The Empire Strikes Back...	Action|Adventure|Sci-Fi
2	186	1196	5.0	4.589631	Star Wars: Episode V - The Empire Strikes Back...	Action|Adventure|Sci-Fi
3	28	1196	4.0	3.609270	Star Wars: Episode V - The Empire Strikes Back...	Action|Adventure|Sci-Fi
4	82	1196	4.0	3.986403	Star Wars: Episode V - The Empire Strikes Back...	Action|Adventure|Sci-Fi
5	256	1196	4.0	4.447170	Star Wars: Episode V - The Empire Strikes Back...	Action|Adventure|Sci-Fi
6	32	1196	4.0	4.213460	Star Wars: Episode V - The Empire Strikes Back...	Action|Adventure|Sci-Fi
7	112	1196	5.0	3.979018	Star Wars: Episode V - The Empire Strikes Back...	Action|Adventure|Sci-Fi
8	64	1196	3.5	4.228626	Star Wars: Episode V - The Empire Strikes Back...	Action|Adventure|Sci-Fi
9	166	1196	4.5	4.397221	Star Wars: Episode V - The Empire Strikes Back...	Action|Adventure|Sci-Fi
10	128	1196	4.0	4.602564	Star Wars: Episode V - The Empire Strikes Back...	Action|Adventure|Sci-Fi

Hình 5: Dự đoán ratting với SVD

RMSE: 0.9337
MAE: 0.7130

	userId	movieId	actual_rating	predicted_rating	title	genres
0	368	1196	3.0	4.279831	Star Wars: Episode V - The Empire Strikes Back...	Action|Adventure|Sci-Fi
1	593	1196	5.0	4.034725	Star Wars: Episode V - The Empire Strikes Back...	Action|Adventure|Sci-Fi
2	312	1196	5.0	4.316298	Star Wars: Episode V - The Empire Strikes Back...	Action|Adventure|Sci-Fi
3	122	1196	5.0	4.461797	Star Wars: Episode V - The Empire Strikes Back...	Action|Adventure|Sci-Fi
4	334	1196	4.0	4.074530	Star Wars: Episode V - The Empire Strikes Back...	Action|Adventure|Sci-Fi
5	305	1196	5.0	4.428147	Star Wars: Episode V - The Empire Strikes Back...	Action|Adventure|Sci-Fi
6	290	1196	5.0	4.378999	Star Wars: Episode V - The Empire Strikes Back...	Action|Adventure|Sci-Fi
7	370	1196	2.5	3.718273	Star Wars: Episode V - The Empire Strikes Back...	Action|Adventure|Sci-Fi
8	140	1196	3.0	4.137600	Star Wars: Episode V - The Empire Strikes Back...	Action|Adventure|Sci-Fi
9	580	1196	4.0	4.249955	Star Wars: Episode V - The Empire Strikes Back...	Action|Adventure|Sci-Fi
10	69	1196	5.0	4.680904	Star Wars: Episode V - The Empire Strikes Back...	Action|Adventure|Sci-Fi

Hình 6: Dự đoán ratting với KNN

Dự đoán đánh giá rating có giá trị từ 0 đến 5, thì giá trị của MAE (Mean Absolute Error) và RMSE (Root Mean Squared Error) sẽ nằm trong một phạm vi tương ứng với phạm vi của các rating này.
Chúng ta thấy sự chênh lệch giữa 'actual ratting' và 'prediction ratting', cũng cho thấy được hệ thống dự đoán khá sát với thực tế. Thêm vào đó cũng thấy được RMSE và MAE của từng mô hình là khác nhau:

- Mô Hình SVD: RMSE: 0.8857, MAE: 0.6825
- Mô Hình KNNBasic:RMSE: 0.9337, MAE: 0.7130

So Sánh

- RMSE của mô hình SVD (0.8857) nhỏ hơn RMSE của mô hình KNNBasic (0.9337). Điều này cho thấy mô hình SVD có khả năng dự đoán gần với giá trị thực tế hơn so với mô hình KNNBasic, khi xét theo tiêu chí RMSE.
- MAE của mô hình SVD (0.6825) cũng nhỏ hơn MAE của mô hình KNNBasic (0.7130). Điều này cho thấy trung bình các dự đoán của mô hình SVD sai lệch ít hơn so với mô hình KNNBasic, khi xét theo tiêu chí MAE.

Nó đánh giá hiệu suất của các mô hình, trích xuất các dự đoán cụ thể cho một bộ phim nhất định, và kết hợp các dự đoán này với các tiêu đề phim tương ứng để dễ dàng diễn giải.
So sánh này cho phép chúng ta thấy sự khác biệt về hiệu suất và chất lượng của các đề xuất giữa hai thuật toán.
Hiệu Suất Tốt Hơn: Mô hình SVD có hiệu suất tốt hơn so với mô hình KNNBasic dựa trên cả hai chỉ số RMSE và MAE. Điều này có nghĩa là SVD đưa ra các dự đoán chính xác hơn và ít sai lệch hơn so với KNNBasic.
Ưu Điểm Của SVD: SVD là một kỹ thuật phân tích giá trị đơn, thường hiệu quả hơn trong việc xử lý các dữ liệu có tính chất phức tạp và không có cấu trúc rõ ràng như dữ liệu đánh giá phim. Nó có khả năng mô hình hóa các tương tác phức tạp giữa người dùng và phim thông qua các nhân tố tiềm ẩn.
KNNBasic: KNNBasic, mặc dù đơn giản và dễ hiểu, có thể không xử lý tốt các dữ liệu phức tạp như trong trường hợp này. Nó hoạt động dựa trên việc tìm kiếm các láng giềng gần nhất, nhưng có thể gặp khó khăn khi dữ liệu có nhiều chiều và sự đa dạng trong sở thích của người dùng.
Vì vậy, dựa trên các kết quả trên, chúng ta có thể kết luận rằng mô hình SVD là lựa chọn tốt hơn cho bài toán dự đoán đánh giá phim trong trường hợp này.

D. Thảo luận

1/Dự đoán bộ phim tương đồng dự trên sở thích cá nhân
Hình ảnh minh họa cho thấy sự đa dạng trong các phương pháp đề xuất phim, đặc biệt là khi so sánh SVD với cosine similarity, KNN với cosine similarity, và SVD với thư viện Surprise. SVD với thư viện Surprise không chỉ bao gồm việc huấn luyện mô hình và dự đoán điểm số, mà còn đánh giá độ chính xác của mô hình dự đoán thông qua các chỉ số RMSE và MAE. Giá trị RMSE là 0.8748 và MAE là 0.6736 cho thấy mô hình SVD với Surprise library hoạt động khá tốt trong việc dự đoán đánh giá phim.
2/Dự đoán đánh giá cho từng bộ phim đối với những người chưa xem
Những kết quả này cho thấy mô hình SVD có khả năng dự đoán gần với giá trị thực tế hơn và ít sai lệch hơn so với KNNBasic. Điều này chứng tỏ rằng SVD là một kỹ thuật hiệu quả hơn trong việc xử lý dữ liệu đánh giá phim phức tạp và không có cấu trúc rõ ràng, nhờ khả năng mô hình hóa các tương tác phức tạp giữa người dùng và phim thông qua các nhân tố tiềm ẩn.
Sau tất cả xin nhấn mạnh đây chỉ là công cụ hỗ trợ cho và đưa ra gợi ý và chỉ mang tính chất tham khảo, nó vẫn phải phụ thuộc tình hình khác quan và quyết định của con người

VI. Kết luận và định hướng tương lai

A. Kết luận

Hệ thống gợi ý phim dựa trên lọc cộng tác (Collaborative Filtering) đã chứng minh được tính hiệu quả và tiềm năng trong việc cung cấp các đề xuất phim phù hợp với sở thích cá nhân của người dùng. Bằng cách khai thác dữ liệu đánh giá của người dùng, hệ thống này có khả năng phân tích và hiểu rõ hơn về sở thích của từng cá nhân.
Collaborative Filtering giúp mở rộng tầm nhìn của người dùng tới những tác phẩm mới mà họ có thể chưa từng biết đến nhưng có sự tương đồng trong cách đánh giá với những bộ phim họ đã xem và yêu thích. Một lợi thế quan trọng của hệ thống lọc cộng tác là khả năng hoạt động hiệu quả ngay cả khi thông tin về nội dung phim hạn chế, vì nó chủ yếu dựa vào dữ liệu đánh giá và hành vi của người dùng. Tuy nhiên, hệ thống Collaborative Filtering cũng có một số hạn chế. Một trong số đó là vấn đề "cold start", khi có ít thông tin về người dùng mới hoặc phim mới. Điều này có thể dẫn đến việc hệ thống không thể đưa ra các đề xuất chính xác.

Để tối ưu hóa hệ thống này, cần liên tục cập nhật và mở rộng cơ sở dữ liệu đánh giá phim, cũng như kết hợp với các phương pháp gợi ý khác như Content-Based Filtering để đạt được độ chính xác và tính đa dạng cao hơn trong các đề xuất. Bằng việc liên tục cải tiến và kết hợp các phương pháp khác nhau, hệ thống gợi ý phim có thể mang lại trải nghiệm giải trí tốt nhất cho người dùng, từ đó tăng cường sự hài lòng và gắn kết của họ với dịch vụ.

B. Định hướng tương lai

Phát triển hệ thống gợi ý lai (Hybrid Recommendation System): Hệ thống gợi ý lai có thể kết hợp nhiều phương pháp khác nhau, bao gồm Content-Based, Collaborative Filtering, và Knowledge-Based Systems. Có thể giúp tận dụng ưu điểm của các phương pháp. Điều này sẽ giúp nâng cao độ chính xác và tính đa dạng của các gợi ý, đồng thời khắc phục các hạn chế của từng phương pháp riêng lẻ. Việc kết hợp này sẽ giúp hệ thống đề xuất các phim không chỉ dựa trên nội dung mà còn dựa trên hành vi và sở thích của những người dùng có thị hiếu tương tự.
Tuy nhiên, hệ thống không phải lúc nào cũng hoàn hảo. Có thể xảy ra trường hợp hệ thống đưa ra đề xuất không chính xác hoặc không phù hợp. Do đó, việc kiểm tra và xác nhận từ đề xuất vẫn cần sự can thiệp của người dùng.

Acknowledgment

Cảm ơn Thầy ThS.Nguyễn Văn Kiệt đã giúp nhóm hoàn thiện bài nghiên cứu.

References

[1] M. Fu, H. Qu, Z. Yi, L. Lu, and Y. Liu, “A novel deep multi-criteria collaborative filtering model for recommendation systems,” Knowledge-Based Systems, vol. 187, Article 104811, 2020.

[2] N. Pavitha, V. Pungliya, A. Raut, R. Bhonsle, A. Purohit, A. Patel, and R. Shashidhar, "Movie recommendation and sentiment analysis using machine learning," Name of the Journal, vol. Volume Number, no. Issue Number, pp. Page Range, 2024.

[3] R. Banik, “Movie Recommender Systems,” https://www.kaggle.com/code/rounakbanik/movie-recommender-systems/notebook.

[4] A. Kamr, “Recommendation System from Zero to Hero,” https://www.kaggle.com/code/abdokamr/recommendation-system-form-zero-to-hero/notebook#Recommendation-system-Collaborative-Filtering.

[5] Krish Cross, ”Netflix’s Movie Recommendation System,” https://www.kaggle.com/code/krishcross/netflix-s-movie-recommendation-system/notebook.

Tên(MSSV)	Đóng góp	Phần trăm
Trương Nhật Quang(22521207)	Viết bài nghiên cứu+ làm slide thuyết trình+ làm code	25%
Nguyễn Khánh Minh(20520635)	Làm code + góp ý bài nghiên cứu + thuyết trình	25%
Đỗ Tuấn Trực(22521548)	Góp ý code+ sửa lỗi bài nghiên cứu +thuyết trình	20%
Nguyễn Đức Tài(22521277)	Thuyết trình	15%
Trương Lưu Song Tâm(22521294)	Thuyết trình	15%

Hình 7: Đóng góp