Hệ thống đề xuất phim dựa trên lọc nội dung và lọc cộng tác sử dụng Apache Spark

Trần Đức Anh
20021296@vnu.edu.vn

Phạm Huy Khôi
20021377@vnu.edu.vn

Phạm Đức Long
20021388@vnu.edu.vn

Hoàng Tú Tài
20021433@vnu.edu.vn

Vũ Thanh Tùng
20021473@vnu.edu.vn

Nguyễn Minh Trí
20020276@vnu.edu.vn

ĐẶT VẤN ĐỀ

Phim ảnh là một trong những hình thức giải trí, nhưng việc tìm ra một bộ phim phù hợp về mặt nội dung trong một số lượng rất lớn các bộ phim được sản xuất hằng năm lại là một công việc không hề dễ dàng. Vì vậy, một hệ khuyến nghị phim sẽ được xây dựng để có thể hỗ trợ người dùng trong trường hợp này. Mục tiêu của bài báo này là mô tả cách thức tiếp cận vấn đề, các công cụ, thuật toán sử dụng trong hệ thống đề xuất phim.

GIỚI THIỆU

Hệ thống đề xuất là các hệ thống dự đoán mà có khả năng gợi ý mạnh mẽ đối với người dùng hoặc gợi ý người dùng đến các mục tiêu, và đôi khi còn gợi ý người dùng đến nhau. Các công ty công nghệ lớn như YouTube, Amazon Prime và Netflix sử dụng các phương pháp tương tự để gợi ý nội dung video phù hợp với sở thích của người dùng. Với lượng dữ liệu khổng lồ trên Internet, việc tìm kiếm nội dung phù hợp có thể rất khó khăn và tốn thời gian, do đó hệ thống đề xuất đóng vai trò quan trọng trong việc tiết kiệm thời gian để tìm kiếm và sử dụng sản phẩm phù hợp cho người dùng. Các hệ thống này đang ngày càng phổ biến trong nhiều lĩnh vực như sách, video, âm nhạc, phim ảnh, ... Các hệ đề xuất sử dụng thông tin người dùng để cải thiện kết quả đề xuất và đưa ra lựa chọn phù hợp nhất. Sự hài lòng của người dùng/khách hàng là yếu tố quan trọng trong việc xây dựng các hệ đề xuất. Điều này có lợi cho cả khách hàng và các công ty, vì khách hàng càng hài lòng, khả năng họ muốn sử dụng hệ thống để tiện lợi càng cao, từ đó tạo ra doanh thu cho các công ty. Hệ thống đề xuất luôn cần được cải thiện, vì sở thích của người dùng có thể khác nhau so với những người dùng khác và nếu người dùng không hài lòng với kết quả, họ có thể không sử dụng lại hệ thống đó. Có hai phương pháp tiếp cận thường được sử dụng phổ biến cho các hệ đề xuất. Đầu tiên, lọc dựa trên nội dung (Content-Based Filtering) , phương pháp tiếp cận này dựa trên đặc điểm của các đối tượng để tìm ra sự tương đồng. Thứ hai, lọc cộng tác (Collaborative Filtering), phương pháp tiếp cận này tập trung vào sự tương đồng giữa các người dùng và đưa ra dự đoán sản phẩm phù hợp cho họ.

NHỮNG NGHIÊN CỨU LIÊN QUAN

Trong dự án này, hệ đề xuất phim của chúng tôi sẽ được tiếp cận bằng các kĩ thuật phổ biến thường được sử dụng đó là Lọc cộng tác (Collaborative Filtering) và Lọc dựa trên nội dung (Content-Based Filtering) sử dụng Framework SparkML và HDFS (Hadoop Distributed Files System).

A. *Lọc cộng tác (Collaborative Filtering)*

Ý tưởng của việc sử dụng lọc cộng tác trong hệ đề xuất phim là dựa trên nguyên tắc rằng những người có sở thích tương đồng trong quá khứ sẽ có xu hướng có các sở thích tương tự trong tương lai. Hệ thống sẽ sử dụng thông tin từ các người dùng khác để dự đoán và đề xuất các bộ phim mà một người dùng cụ thể có thể quan tâm. Phương pháp lọc cộng tác dựa trên ALS, một thuật toán dùng để phân tích ma trận. Ưu điểm của collaborative filtering là nó không đòi hỏi thông tin chi tiết về bộ phim và người dùng, chỉ cần dựa vào lịch sử đánh giá hoặc lịch sử xem phim để tìm ra sự tương đồng. Ngoài ra, hệ thống cũng có khả năng đề xuất các bộ phim mới hoặc ít được biết đến dựa trên sở thích của những người dùng tương tự.

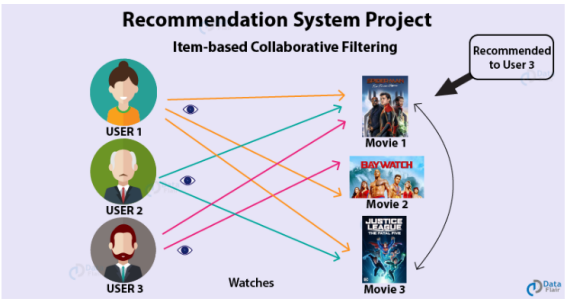

Fig. 1: Item-based Collaborative Filtering

*1) **Alternating Least Squares (ALS)**:* Thuật toán ALS được sử dụng trong trường hợp xử lí ma trận thưa trong việc xây dựng hệ đề xuất. Nguyên lý của thuật toán có thể được miêu tả như sau, dữ liệu đánh giá được biểu diện ở dạng ma trận M × N là R. Trong đó có n người dùng và m sản phẩm cần được đề xuất. Phần tử r_{ui} ở đây được định nghĩa là đánh giá của người dùng u cho sản phẩm thứ i. Ma trận R là một

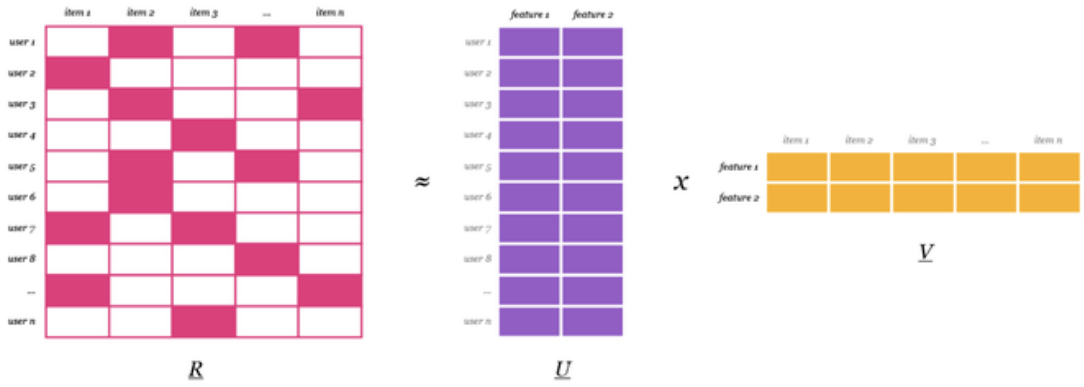

Fig. 2: Mô tả Alternating Least Squares (ALS)

ma trận thưa vì không phải tất cả các sản phẩm nhận được đánh giá của người dùng. Đó là lý do sẽ xuất hiện nhiều giá trị trống trong ma trận. Ta sử dụng phương pháp phân rã ma trận hay Matrix Factorization để có thể giải quyết vấn đề trên. Có 2 vector k chiều quan trọng trong việc phân ra chúng:

- x_u là một vector k chiều lưu lại thông tin đánh giá của người dùng u.
- y_i là một vector k chiều lưu lại thông tin đánh giá của mặt hàng i.

Ta có những biểu thức sau:

$$r_{ui} \approx x_u^T Y_i \tag{1}$$

$$x_u = x_1, x_2, x_3, \ldots, x_n \in \mathbb{R}^k \tag{2}$$

$$y_u = y_1, y_2, y_3, \ldots, y_n \in \mathbb{R}^k \tag{3}$$

Lúc này biểu thức (1) được đưa về dạng một bài toán tối ưu cần được tính toán

$$\text{argmin} \sum_{r_{ui}} \left(r_{ui} - x_u^T y_i\right)^2 + \lambda \left(\sum_u \|x_u\|^2 + \sum_i \|y_i\|^2\right) \tag{4}$$

Ở đây, λ là hệ số tinh chỉnh (regularization factor), được sử dụng để giải quyết vấn đề quá khớp (overfitting). Được gọi là "weighted-λ -regularization"(chế độ tinh chỉnh λ có trọng số). Giá trị của λ có thể được tinh chỉnh để giải quyết vấn đề quá khớp, trong khi giá trị mặc định là 1. Giả sử tập hợp biến x_u là hằng số, sau đó hàm mục tiêu của y_i là lõm và tập hợp biến y_i là hằng số, sau đó hàm mục tiêu của x_u là lõm. Do đó, giá trị tối ưu của x_u và y_i có thể được tìm bằng cách lặp lại phương pháp trên cho đến khi đạt đến sự hội tụ. Đây được gọi là phương pháp ALS, dưới đây là mã giả cho thuật toán ALS.

Mã giả của phương pháp ALS được thể hiện trong Algorithm 1.

Algorithm 1 Alternating Least Squares (ALS)

```
procedure ALS(x_u, y_i)
    Initialization x_u ← 0
    Initialization matrix y_i with random values
    repeat
        Fix y_i, solve x_u by minimizing the objective function (the sum of squared errors)
        Fix x_u, solve y_i by minimizing the objective function similarly
    until reaching the maximum iteration
    Return x_u, y_i
end procedure
```

B. Lọc dựa trên nội dung (Content-Based Filtering)

Content-Based Filtering là một phương pháp trong hệ thống đề xuất phim mà sử dụng thông tin về nội dung của phim để tìm kiếm và đề xuất những phim có sự tương đồng với các phim mà người dùng đã thích trước đó. Ý tưởng chính của Content-Based Filtering là sử dụng các thuộc tính của phim (thể loại, diễn viên, đạo diễn, từ khóa, đánh giá của người dùng, nhãn, ...) để xác định sự tương đồng giữa các phim và từ đó đề xuất những phim tương tự mà người dùng có thể quan tâm. Ưu điểm của Content-Based Filtering bao gồm:

- Độc lập với người dùng: Không cần thông tin về sở thích của người dùng khác để tạo ra các đề xuất phù hợp. Chỉ cần thông tin về nội dung của phim là đủ để tạo ra các đề xuất.
- Tính cá nhân hóa: Cung cấp các đề xuất phim dựa trên sở thích cá nhân của người dùng, không bị ảnh hưởng bởi sự phụ thuộc vào hành vi của người dùng khác.
- Giải quyết vấn đề "cold start": Content-based filtering có thể đề xuất phim cho người dùng ngay cả khi không có thông tin sẵn có về sở thích.

*1) **TF - Term Frequency**:* TF dùng để ước lượng tần xuất xuất hiện của từ trong văn bản hay nhãn của các bộ phim trong tập dữ liệu . Tuy nhiên với mỗi văn

bản thì có độ dài khác nhau, vì thế số lần xuất hiện của từ có thể nhiều hơn . Vì vậy số lần xuất hiện của từ sẽ được chia độ dài của văn bản (tổng số từ trong văn bản đó).

*2) **IDF - Inverse Document Frequency**:* IDF dùng để ước lượng mức độ quan trọng của một từ hay một nhãn nào đó. Khi tính tần số xuất hiện TF thì các từ đều được coi là quan trọng như nhau.

*3) **TF-IDF(Term Frequency - Inverse Document Frequency)**:* TF-IDF được tính bằng tích của TF và IDF, kết quả là một giá trị dương, giúp đánh giá mức độ quan trọng của một từ trong văn bản hay mức độ đại diện của nhãn đối với một bộ phim.

*4) **Độ tương tự Cosine (Cosine Similarity)**:* Cosine Similarity được sử dụng để tính độ tương tự hay độ giống nhau giữa 2 vector nhiều chiều dựa vào cos(θ) hay cosine của góc hợp bởi 2 vector.

$$\cos(\theta) = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| \cdot \|\vec{B}\|}$$

Giá trị của Cosine Similarity trong hệ đề xuất nằm trong khoảng từ -1 đến 1. Giá trị càng gần 1 hay góc θ càng gần bằng 0 °, tức là hai vector có hướng gần nhau, và giá trị càng gần -1 hay góc θ càng gần bằng 180° tức là hai vector có hướng ngược nhau. Nếu cosine similarity bằng 1, hai vector hoàn toàn giống nhau, trong khi nếu cosine similarity bằng -1, hai vector hoàn toàn trái ngược nhau. Khi cosine similarity bằng 0, hai vector không có sự tương quan tuyến tính. Trong hệ đề xuất phim của chúng tôi, độ tương tự Cosine được sử dụng để tìm ra các bộ phim có độ tương đồng với nhau.

TẬP DỮ LIỆU

Trong nghiên cứu này nhằm đảm bảo tính đa dạng, trực quan về dữ liệu chúng tôi sử dụng bộ dữ liệu MovieLens. Bộ dữ liệu này cung cấp thông tin về 58,000 bộ phim và 27 triệu điểm đánh giá từ 280,000 người dùng. Dữ liệu bao gồm các thông tin về tên phim, thể loại, điểm đánh giá, gán nhãn của phim,...Việc sử dụng bộ dữ liệu này trong hệ thống đề xuất cung cấp sự phong phú về dữ liệu, dễ dàng hơn trong việc xây dựng các mô hình dự đoán. Điều này giúp chúng tôi có cái nhìn sâu hơn về sở thích của người dùng và khám phá các mối quan hệ tiềm năng giữa các yếu tố khác nhau trong bộ dữ liệu. Trong bộ dữ liệu MovieLens, các bảng dữ liệu được cung cấp bao gồm:

- movie: chứa thông tin về các bộ phim (id, tên phim, thể loại,...)
- rating: chứa thông tin về điểm đánh giá mà người dùng dành cho bộ phim đó
- tag: chứa các tags được người dùng gán nhãn cho các bộ phim
- links: chứa các liên kết đến các trang Imdb của từng bộ phim.

HỆ THỐNG ĐỀ XUẤT PHIM

A.Workflow

Workflow của hệ thống đề xuất phim đươc mô tả như trong Hình 5

B.Collaborative Filtering (Lọc cộng tác)

***1. Mô tả bài toán**:* Đối với phương pháp này chúng tôi tập trung vào thuật toán ALS, sử dụng trên bảng dữ liệu rating. Mục tiêu chính của phương pháp này là đề xuất những gợi ý và dự đoán đánh giá phim dựa trên sự tương tự giữa các người dùng và các phim. Bảng dữ liệu rating bao gồm các đánh giá người dùng đưa ra cho các bộ phim họ đã xem. Mục tiêu là dự đoán các đánh giá còn thiếu của người dùng cho các phim họ chưa xem để đưa ra gợi ý tốt nhất dựa trên điểm đánh giá cho từng người dùng.

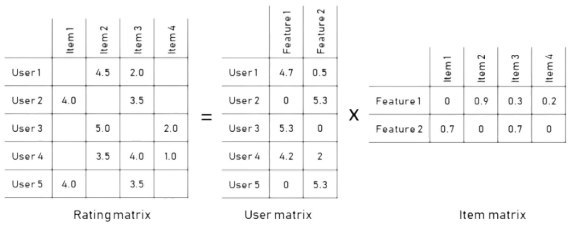

Fig. 3: Ma trận đánh giá người dùng - phim

***2. Xây dựng mô hình**:* Trước khi áp dụng ALS, bộ dữ liệu rating được chia thành hai tập:

- **Train** (80%): tập huấn luyện được sử dụng để huấn luyện mô hình ALS
- **Test** (20%): tập kiểm tra

Thuật toán ALS được áp dụng lên tập huấn luyện để học ma trận người dùng-phim tối ưu. Mô hình ALS sẽ tìm kiếm ma trận này bằng cách thay phiên nhau cập nhật ma trận người dùng khi giữ ma trận phim cố định và cập nhật ma trận phim khi giữ ma trận người dùng cố định

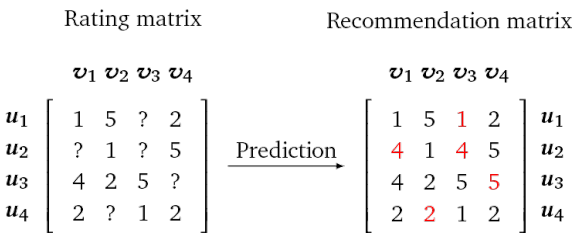

Fig. 4: Minh họa ma trận đánh giá sau khi dự đoán

***3. Đề xuất phim cho người dùng**:* Kết quả của mô hình ALS là ma trận người dùng-phim đã được học từ tập huấn luyện. Ma trận này chứa thông tin về mức độ quan tâm của mỗi người dùng đối với các phim từ đó dự đoán các đánh giá còn thiếu của người dùng.

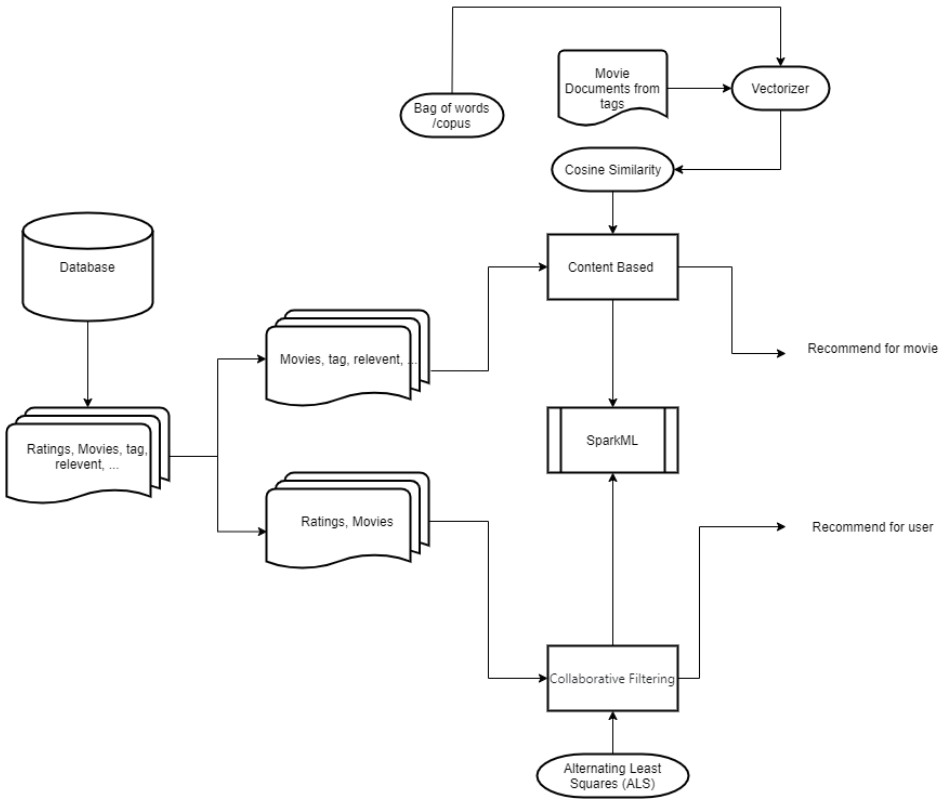

Fig. 5: Workflow hệ thống đề xuất phim

C. Content Based Filtering (Lọc dựa trên nội dung)

1. Tiền xử lý dữ liệu:

Trong hệ thống đề xuất phim, việc gắn các tag phù hợp cho các bộ phim là một yếu tố quan trọng để xác định tính tương đồng và khám phá nội dung của các phim. Trong bộ dữ liệu của chúng tôi số lượng phim được gán nhãn trong bảng tag chỉ là hơn 45000 bộ phim trong khi bảng dữ liệu movie có hơn 58000 bộ phim. Sự thiếu sót trong việc gán nhãn có thể làm giới hạn khả năng đề xuất chính xác.

Để giải quyết vấn đề này chúng tôi sử dụng kỹ thuật crawl data từ các nguồn bên ngoài như Imdb, nơi cung cấp thông tin phong phú về các bộ phim, bao gồm cả các tags mô tả nội dung của chúng. Qua việc crawl data tags từ IMDb, ta có thể bổ sung thông tin chi tiết về các tags cho các bộ phim chưa có đầy đủ tag, từ đó nâng cao chất lượng và hiệu quả của hệ thống đề xuất phim. Quy trình crawl data tags từ IMDb:

- Trích xuất danh sách các bộ phim đã có tag từ tệp tags.csv, đồng thời đối chiếu với tệp movies.csv để xác định các bộ phim chưa có đầy đủ tag
- Kết hợp danh sách các bộ phim chưa có tag với imdbId từ tệp links.csv để xác định danh sách các bộ phim cần crawl thông tin tags.
- Sử dụng kỹ thuật crawl dữ liệu từ trang IMDb theo định dạng URL "https://www.imdb.com/title/imdbId/keywords", trong đó imdbId là mã id của bộ phim.
- Sau khi crawl thông tin tags từ trang IMDb và lưu vào tệp tags2.csv với hai cột là imdbId và tag, chúng ta chuyển đổi cột imdbId sang movieId và kết hợp với tệp tags.csv ban đầu để cập nhật thông tin tags cho các bộ phim chưa có tag.

2. Trích xuất đặc trưng:

- **Tạo documents cho các tags** : Mỗi tag sẽ được coi là một document riêng, và các tags liên quan đến cùng một phim sẽ được kết hợp thành một document.
- **Bag of words**: Sau đó, chúng ta xây dựng một bag of words (BOW) cho mỗi document. BOW là một vector đại diện cho mỗi document, trong đó mỗi thành phần của vector biểu diễn số lần xuất hiện của từ trong document đó.
- **Áp dụng TF-IDF**: Tiếp theo, chúng ta áp dụng phương pháp TF-IDF (Term Frequency-Inverse Document Frequency) để đánh giá tầm quan trọng của từ trong mỗi document. TF-IDF giúp chúng ta xác định độ quan trọng của từ dựa trên tần suất xuất hiện của từ trong document và tần suất xuất hiện của từ đó trong toàn bộ bộ dữ liệu
- **Tính độ cosine giữa các phim**: Dựa trên TF-IDF, chúng ta tính độ cosine giữa các vector biểu diễn TF-IDF của các phim để đo độ tương đồng giữa chúng

3. Đề xuất phim tương tự:

Sau khi có được độ tương đồng cosine chúng ta sẽ sử dụng nó để đề xuất các phim tương ứng cho một bộ phim nhất định sắp xếp theo độ tương đồng cosine giảm dần

D. Cold Start - Giải quyết vấn đề người dùng mới

Để giải quyết vấn đề Cold Start cho người dùng mới, chúng tôi đã triển khai một phương pháp đặc biệt nhằm mang đến trải nghiệm cá nhân hóa và đáp ứng nhu cầu giải trí từ ngay lần đầu tiên người dùng truy cập vào hệ thống. Phương pháp này cho phép người dùng mới chọn một thể loại mà họ quan tâm hoặc muốn khám phá, và dựa trên thể loại này, chúng tôi sẽ đề xuất cho họ danh sách các bộ phim liên quan đến thể loại đó.
Ý tưởng của phương pháp này là tạo ra một khởi đầu tương tác tích cực và khám phá cho người dùng mới. Thay vì yêu cầu người dùng điền vào một loạt thông tin cá nhân chi tiết hoặc đánh giá phim trước khi có thể nhận được đề xuất, chúng tôi cho phép họ chọn một thể loại mà họ có quan tâm. Điều này không chỉ giúp giảm bớt rào cản cho người dùng mới mà còn tạo ra một trải nghiệm đơn giản và thuận tiện. Sau khi người dùng nhập vào genres chúng tôi sẽ lọc ra các phim có ratings cao nhất theo thể loại đó. Sau đó chọn ngẫu nhiên một bộ phim trong danh sách các phim đó và áp dụng content-based như trên.

ĐÁNH GIÁ MÔ HÌNH

A. Lọc cộng tác

Tiêu chí đánh giá mô hình collaborative filtering được sử dụng trong nghiên cứu của chúng tôi là Root Mean Square Error (RMSE). RMSE là một phép đo đánh giá độ chính xác của mô hình dựa trên sự khác biệt giữa giá trị dự đoán và giá trị thực tế.

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

RMSE được sử dụng để đo lường sự khác biệt trung bình giữa các giá trị rating dự đoán với giá trị thực tế trong tập kiểm tra. RMSE càng nhỏ, thì mô hình càng chính xác trong việc dự đoán rating và đề xuất phim cho người dùng
Kết quả của các dự đoán hay độ lớn của RMSE tùy thuộc vào các tham số trong thuật toán ALS . Bảng I sẽ cho biết sư thay đổi của RMSE dựa theo các thông số đầu vào.

- **Rank** : Tham số xác định số lượng các yếu tố ẩn (latent factors)
- **Num of iters**: Số lần lặp để huấn luyện mô hình
- **regParam**: Tham số regularization, được sử dụng để kiểm soát sự phức tạp của mô hình và tránh overfitting

Rank	Nums of iters	reg_param	RMSE
5	5	0.01	0.8391
5	10	0.01	0.8306
5	25	0.01	0.8284
5	10	0.05	0.8183
10	5	0.01	0.8421
10	10	0.01	0.8337
10	25	0.05	0.8009
25	5	0.01	0.8520
25	10	0.05	0.7987
25	25	0.05	0.7874
50	50	0.05	0.7802
100	**50**	**0.05**	**0.7784**

TABLE I: Bảng kết quả với các tham số khác nhau và RMSE tương ứng

B. Lọc nội dung

Đối với mô hình Content-based chúng tôi sử dụng phương pháp đánh giá Hit-Rate. Nó đo lường tỷ lệ phần trăm các dự đoán đúng (hits) trên tổng số lượng các dự đoán. Trong trường hợp này, hit rate được sử dụng để đánh giá hiệu suất của mô hình đề xuất phim dựa trên genres. Bằng cách so sánh genres của các bộ phim được đề xuất với genres của một bộ phim cụ thể, hit rate đo lường khả năng mô hình đề xuất các bộ phim có genres tương tự với bộ phim đã cho.

$$HR = \frac{M}{M + A} \tag{2}$$

Tính tỷ lệ hit rate bằng cách chia số lượng genres trùng khớp cho tổng số lượng genres của các phim đề xuất. Điều này cho ta biết tỷ lệ phần trăm của các genres được đề xuất giống với genres của phim gốc

Lấy ví dụ một bộ phim có genres tương ứng là: ['Comedy', 'Romance'].
Các bộ phim được đề xuất có danh sách các genres: ['Comedy', 'Drama', 'Romance', 'Comedy', 'Romance', 'Drama', 'Comedy', 'Action', 'Comedy', 'Drama', 'Action', 'Adventure', 'Comedy', 'Comedy', 'Drama', 'Romance', 'Comedy', 'Drama', 'Comedy', 'Drama', 'Romance', 'Crime', 'Drama', 'Mystery', 'Romance', 'Thriller', 'Comedy', 'Horror', 'Comedy', 'Comedy', 'Comedy', 'Comedy', 'Comedy', 'Comedy', 'Musical', 'Comedy', 'Action', 'Comedy', 'Crime', 'Comedy']
Lúc này số lương genres trùng khớp là 23 trên tổng số 40 genres vậy nên HR = 23/40 = 0.575

ĐỀ XUẤT CẢI TIẾN TRONG TƯƠNG LAI

Cả hai mô hình đề xuất đã đưa ra đều có ưu và nhược điểm riêng. Lọc cộng tác tỏ ra hạn chế khi chỉ dựa trên ratings để đánh giá sự tương đồng và đề xuất, dẫn đến những khuyết điểm như cold-start problem (khó đề xuất cho người dùng mới) và sparsity problem (hiếm khi có đủ đánh giá). Trong khi lọc nội dung có thể gặp khó khăn trong việc tìm ra các thuộc tính phù hợp để đưa ra khuyến nghị, đặc biệt là đối với những

dữ liệu đa dạng việc tính toán độ tương đồng có thể rất mất thời gian.

Trong thực tế, cả hai phương pháp đều được sử dụng tùy thuộc vào mục đích sử dụng và tính chất của sản phẩm hay nội dung. Ví dụ, phương pháp collaborative filtering thường được sử dụng trong các hệ thống khuyến nghị sản phẩm như Amazon, Netflix, hay Spotify, trong khi phương pháp content-based filtering thường được sử dụng trong các hệ thống khuyến nghị nội dung như Google News hay Youtube. Đặc biệt đối với lĩnh vực đề xuất phim trên thực tế Netflix hay các trang web xem phim nổi tiếng cũng đã và đang sử dụng kết hợp hai phương pháp

Mô hình Hybrid Recommend thay vì chỉ sử dụng rating như lọc công tác hay tag như lọc nội dung thì sẽ kết hợp những thông tin này. Cụ thể quá trình xây dựng mô hình hybrid bao gồm các bước:

- Xây dựng mô hình Lọc Cộng Tác
- Xây dựng mô hình Lọc nội dung
- Kết hợp hai mô hình: Lúc này thay vì đưa ra đề xuất cho người dùng dựa trên mỗi rating thì chúng tôi sẽ đưa ra đề xuất dựa trên các bộ phim tương tự với bộ phim người dùng đã đưa ra đánh giá và sắp xếp theo rating dự đoán cao nhất

Kết hợp Collaborative Filtering và Content-Based Recommendation thành mô hình Hybrid có thể sẽ đem lại độ chính xác và tin cậy cao hơn trong việc đề xuất các bộ phim cho người dùng. Vì chung quy lại tất cả các sản phẩm mà các doanh nghiệp làm ra đều phải hướng đến người dùng nhất là trong lĩnh vực giải trí như phim ảnh, âm nhạc,...

KẾT LUẬN

Bài báo đã đề xuất mô hình gồm hai phương pháp là lọc cộng tác dựa trên thuật toán Alternating Least Squares (ALS) và lọc dựa trên nội dung dựa trên kỹ thuật TF-IDF, cosine similarity của Framework SparkML. Trong phương pháp lọc cộng tác, việc lựa chọn các tham số của thuật toán ALS có thể ảnh hưởng đến hiệu suất của hệ thống. Với lọc theo nội dung, việc gán nhãn của người dùng cho các bộ phim là rất quan trọng, với nhiều nhãn cho mỗi bộ phim từ đó có thể xây dựng mô hình đề xuất tốt hơn.

Trong tương lai việc kết hợp cả hai phương pháp, một mô hình hybrid recommend đáng tin cậy mạnh mẽ sẽ được tạo ra Phương pháp hybrid cho phép hệ thống đề xuất phim dựa trên sự kết hợp thông tin về mức độ quan tâm của người dùng đối với các phim từ mô hình ALS và sự tương đồng về nội dung của các bộ phim từ phương pháp lọc theo nội dung. Kết quả là một hệ thống đề xuất phim có khả năng cung cấp những đề xuất phù hợp và cá nhân hóa cho từng người dùng

Tóm lại, hệ thống đề xuất phim mang lại nhiều lợi ích quan trọng cho cả người dùng và các dịch vụ giải trí trực tuyến. Nó là hệ thống không thể thiếu ở các sản phẩm giải trí trực tuyến như Netflix, Spotify, Youtube,... Không chỉ giúp cá nhân tiết kiệm thời gian, đa dạng hóa lựa chọn mà còn giúp tối ưu hóa dịch vụ, tăng doanh thu cho các dịch vụ giải trí.

TÀI LIỆU THAM KHẢO

[1] Recommendation System for E-commerce using Alternating Least Squares (ALS) on Apache Spark; https://www.diva-portal.org/smash/get/diva2:1504620/FULLTEXT02.pdf

[2] Feature Extraction and Transformation - RDD-based API; https://spark.apache.org/docs/latest/mllib-feature-extraction.html

[3] Collaborative Filtering SparkML; https://spark.apache.org/docs/latest/ml-collaborative-filtering.html

[4] Movie Recommender System; https://www.codeheroku.com/static/workshop/hw/movierecommendation/MovieRecommenderSystem

[5] A Recommendation Engine for Predicting Movie Ratings Using a Big Data Approach;Mazhar Javed Awan, Rafia Asad Khan, Haitham Nobanee, Awais Yasin, Syed Muhammad Anwar, Usman Naseem, and Vishwa Pratap Singh https://www.mdpi.com/2079-9292/10/10/1215

[6] Movie Recommender System; Yijie Zhuang(yz226), Boyang Xu(bx15), Hao Wu(hw135), Shaoyi Han(sh335), Hong Jin(hj68); https://www.scribd.com/document/510788403/Project-Report-Movie-Recommender-System

[7] Internet Movie Database (IMDB); https://www.imdb.comhome