HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

BÁO CÁO BÀI TẬP LỚN

HỌC PHẦN: NHẬP MÔN KHOA HỌC DỮ LIỆU

ĐỀ TÀI: XÂY DỰNG HỆ THỐNG GỢI Ý PHIM

Nhóm Lớp: 06

Sinh viên thực hiện:

Trần Trọng Đại B22DCCN178

Giảng viên hướng dẫn: TS. Trần Tiến Công

HÀ NỘI - 2025

MỤC LỤC

1. THU THẬP DỮ LIỆU

1.1. Nguồn dữ liệu

Hệ thống sử dụng bộ dữ liệu MovieLens, một trong những bộ dữ liệu được sử dụng phổ biến nhất trong lĩnh vực nghiên cứu hệ thống gợi ý. Bộ dữ liệu được thu thập từ hai nguồn chính:

Kaggle: Bộ dữ liệu Movie Recommendation System

GroupLens: MovieLens Latest Small (ml-latest-small)

1.2. Cấu trúc dữ liệu

Dữ liệu thô bao gồm hai tập tin CSV chính:

movies.csv: Chứa thông tin về phim

- movieId: Định danh duy nhất của phim
- title: Tên phim và năm phát hành
- genres: Thể loại phim (phân cách bằng ký tự "|")

ratings.csv: Chứa thông tin đánh giá của người dùng

- userId: Định danh người dùng
- movieId: Định danh phim
- rating: Điểm đánh giá (thang điểm 0.5 - 5.0)
- timestamp: Thời điểm đánh giá

2. LÀM SẠCH VÀ CHUẨN BỊ DỮ LIỆU

2.1. Tiền Xử Lý Dữ Liệu Phim

Quá trình làm sạch dữ liệu phim được thực hiện qua các bước:

- Loại bỏ bản ghi trùng lặp: Xóa các phim có movieId trùng nhau
- Xử lý giá trị thiếu:
    - Loại bỏ các bản ghi thiếu movieId hoặc title
    - Điền giá trị "(no genres listed)" cho các phim thiếu thông tin thể loại

- Chuẩn hóa dữ liệu văn bản: Xóa khoảng trắng thừa trong tên phim
- Trích xuất thông tin bổ sung:
    - Phân tách thể loại thành danh sách
    - Trích xuất năm phát hành từ tên phim sử dụng biểu thức chính quy

2.2. Tiền Xử Lý Dữ Liệu Đánh Giá

Dữ liệu đánh giá được làm sạch theo quy trình:

Loại bỏ giá trị thiếu: Xóa các bản ghi có thuộc tính rỗng

Loại bỏ bản ghi trùng lặp: Mỗi cặp (userId, movieId) chỉ giữ lại một đánh giá

Lọc giá trị ngoại lai: Chỉ giữ lại các đánh giá trong khoảng 0.5 đến 5.0

Đảm bảo tính nhất quán: Loại bỏ đánh giá cho các phim không tồn tại trong tập dữ liệu phim

2.3. Lọc Tương Tác Tối Thiểu

Để đảm bảo chất lượng dữ liệu huấn luyện, hệ thống áp dụng ngưỡng lọc iterative:

Ngưỡng người dùng: Mỗi người dùng phải có tối thiểu 5 đánh giá

Ngưỡng phim: Mỗi phim phải có tối thiểu 10 đánh giá

Quá trình lọc được thực hiện lặp đi lặp lại cho đến khi đạt trạng thái hội tụ, đảm bảo không còn người dùng hoặc phim nào vi phạm ngưỡng.

2.4. Thống Kê Dữ Liệu

Sau khi làm sạch, hệ thống tính toán và báo cáo các thống kê quan trọng:

- Tổng số phim và đánh giá
- Số lượng người dùng duy nhất

- Phân bố điểm đánh giá (trung bình, trung vị, khoảng giá trị)
- Độ thưa của ma trận (Sparsity): Được tính theo công thức
- Sparsity = 1 - (số đánh giá / (số người dùng × số phim))

3. PHÂN TÍCH VÀ TRỰC QUAN HÓA DỮ LIỆU

3.1. Phân Tích Thể Loại

Hệ thống phân tích phân bố thể loại phim để hiểu đặc điểm của bộ dữ liệu. Thể loại được trích xuất và thống kê tần suất xuất hiện, giúp xác định các thể loại phổ biến nhất.

3.2. Phân Tích Phân Bố Đánh Giá

Các phân tích được thực hiện bao gồm:

- Phân bố điểm đánh giá
- Phân bố năm phát hành phim
- Số lượng đánh giá theo người dùng và theo phim

3.3. Phân Tích Độ Thưa Ma Trận

Ma trận tương tác người dùng-phim thường có độ thưa rất cao (trên 95%), phản ánh thực tế rằng mỗi người dùng chỉ đánh giá một phần nhỏ trong tổng số phim có sẵn.

4. XÂY DỰNG MÔ HÌNH GỢI Ý

4.1. Mô Hình User-Based Collaborative Filtering

User-Based Collaborative Filtering dựa trên giả định: những người dùng có hành vi tương tự trong quá khứ sẽ tiếp tục có sở thích giống nhau trong tương lai. Mô hình gợi ý phim cho người dùng A bằng cách:

- Tìm những người dùng có đánh giá tương tự với A
- Tổng hợp đánh giá của những người dùng tương tự cho các phim A chưa xem

Độ tương tự giữa hai người dùng được tính bằng Cosine Similarity:

$$sim(u, v) = (R_u \cdot R_v) / (||R_u|| \times ||R_v||)$$

Trong đó R_u và R_v là vector đánh giá của người dùng u và v.

Do bộ dữ liệu có kích thước lớn, hệ thống áp dụng các kỹ thuật tối ưu:

- **Lấy mẫu người dùng (User Sampling)**: Khi số người dùng vượt quá ngưỡng (mặc định 10,000), hệ thống lấy mẫu ngẫu nhiên để tính ma trận độ tương tự, giảm độ phức tạp tính toán.
- **Lưu trữ Top-K láng giềng**: Thay vì lưu toàn bộ ma trận độ tương tự (N×N), chỉ lưu K láng giềng tương tự nhất cho mỗi người dùng (mặc định K=50). Điều này giảm đáng kể bộ nhớ sử dụng.
- **Tính toán theo lô (Batch Processing)**: Ma trận độ tương tự được tính theo từng lô với kích thước cố định, cho phép xử lý trên các hệ thống có bộ nhớ hạn chế.

Điểm dự đoán cho người dùng u và phim i được tính bằng trung bình có trọng số:

$$pred(u, i) = \Sigma(sim(u, v) \times rating(v, i)) / \Sigma|sim(u, v)|$$

Trong đó tổng được tính trên tập các người dùng v đã đánh giá phim i.

Đối với người dùng không nằm trong tập mẫu hoặc không có láng giềng tương tự, hệ thống sử dụng cơ chế dự phòng:

- Gợi ý dựa trên các phim được đánh giá cao bởi người dùng
- Gợi ý các phim phổ biến nhất (cold-start fallback)

4.2. Mô Hình Item-Based Collaborative Filtering

Item-Based Collaborative Filtering dựa trên nguyên tắc: người dùng có xu hướng thích những phim tương tự với những phim họ đã thích trước đó. Mô hình tính toán độ tương tự giữa các cặp phim dựa trên mẫu đánh giá của người dùng.

Ma trận tương tác được chuyển vị (items × users) để tính độ tương tự giữa các phim. Độ tương tự Cosine được áp dụng tương tự như User-Based CF:

$$sim(i, j) = (R_i \cdot R_j) / (||R_i|| \times ||R_j||)$$

Trong đó R_i là vector đánh giá mà phim i nhận được từ tất cả người dùng.

Điểm dự đoán cho người dùng u và phim i:

$$pred(u, i) = \Sigma(sim(i, j) \times rating(u, j)) / \Sigma|sim(i, j)|$$

Trong đó tổng được tính trên tập các phim j mà người dùng u đã đánh giá.

Mô hình Item-Based CF cung cấp khả năng tìm phim tương tự (get_similar_movies), cho phép:

- Gợi ý phim dựa trên một phim cụ thể
- Hỗ trợ tính năng "Có thể bạn cũng thích" trên giao diện người dùng

4.3. Mô Hình Neural Collaborative Filtering

Neural Collaborative Filtering (NCF) sử dụng mạng nơ-ron để học các mối quan hệ phi tuyến giữa người dùng và phim. Thay vì tính toán độ tương tự tường minh, mô hình học hàm ánh xạ từ cặp (user, item) sang điểm đánh giá thông qua quá trình huấn luyện.

Mô hình sử dụng **Multi-Layer Perceptron (MLP)** với sklearn's MLPRegressor:

- **Đầu vào**: Chỉ số người dùng và phim được chuẩn hóa về khoảng [0, 1]
- **Các lớp ẩn**: Kiến trúc (64, 32, 16) neurons
- **Đầu ra**: Điểm đánh giá dự đoán

Chỉ số người dùng và phim được chuẩn hóa trước khi đưa vào mô hình:

- user_idx_normalized = user_idx / n_users
- movie_idx_normalized = movie_idx / n_movies

Mô hình được huấn luyện với các tham số:

- **Tốc độ học (Learning Rate)**: 0.001
- **Số vòng lặp tối đa**: 20 epochs
- **Early Stopping**: Dừng sớm khi không cải thiện trên tập validation
- **Tỷ lệ validation**: 10% của tập huấn luyện

Để xử lý bộ dữ liệu lớn, hệ thống lấy mẫu tối đa 500,000 đánh giá cho quá trình huấn luyện.

Ưu Điểm So Với CF Truyền Thống:

- Học tương tác phi tuyến: Có thể nắm bắt các mẫu phức tạp mà phương pháp tuyến tính không thể
- Không cần tính ma trận độ tương tự: Tiết kiệm bộ nhớ cho dữ liệu lớn
- Dự đoán theo lô hiệu quả: Có thể dự đoán nhiều cặp (user, item) cùng lúc

4.4. Mô Hình Hybrid

Mô hình Hybrid kết hợp điểm mạnh của nhiều mô hình đơn lẻ để cải thiện chất lượng gợi ý. Hệ thống triển khai phương pháp Weighted Ensemble, trong đó dự đoán cuối cùng là trung bình có trọng số của các mô hình thành phần.

Mô hình Hybrid tích hợp ba mô hình:

- User-Based Collaborative Filtering
- Item-Based Collaborative Filtering
- Neural Collaborative Filtering

Trọng số mặc định được cấu hình như sau:

- User-Based CF: 30%
- Item-Based CF: 30%
- Neural CF: 40%

Trọng số được chuẩn hóa để tổng bằng 1, cho phép người dùng tùy chỉnh theo nhu cầu cụ thể.

Khi sinh gợi ý, mô hình Hybrid thực hiện:

- Lấy K×3 ứng viên từ mỗi mô hình thành phần
- Hợp nhất tập ứng viên (loại bỏ trùng lặp)
- Tính điểm Hybrid cho từng ứng viên
- Sắp xếp và trả về top-K gợi ý

5. ĐÁNH GIÁ MÔ HÌNH

5.1. Các phương pháp đánh giá

Do kích thước dữ liệu lớn, việc đánh giá được thực hiện trên mẫu:

- **RMSE/MAE**: Đánh giá trên 10,000 mẫu ngẫu nhiên từ tập kiểm thử
- **Precision@K/Recall@K**: Đánh giá trên 100 người dùng được chọn ngẫu nhiên

5.2. Kết quả đánh giá

	RMSE	MAE	Precision@K	Recall@K
User-Based	1.043	0.823	0.001	0.0007
Item-Based	0.861	0.638	0.005	0.006
Neural CF	1.039	0.821	0.01	0.006
Hybrid	0.940	0.738	0.014	0.007