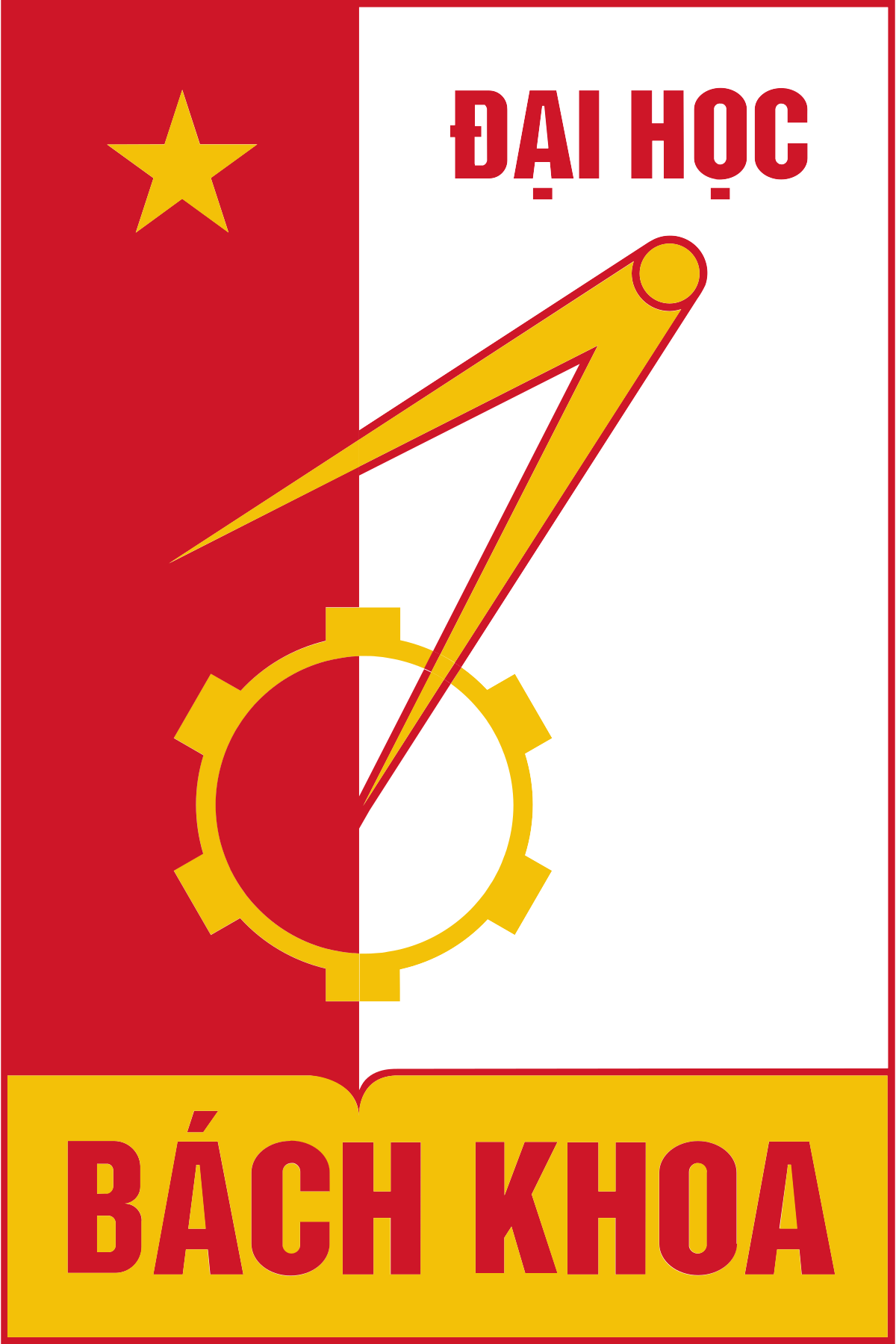
**TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI**

**Trường Công nghệ thông tin và Truyền thông**



**BÁO CÁO BÀI TẬP LỚN**

**HỆ THỐNG GỢI Ý PHIM**

Học phần: Nhập môn Trí tuệ Nhân tạo – IT3160

GVHD: TS. Trần Thế Hùng

**Nhóm 14**

|  |  |
| --- | --- |
| **Họ và tên** | **MSSV** |
| Phạm Việt Hải | 20215044 |
| Trần Bình Minh | 2021 |
| Nguyễn Lê Sơn | 2021 |
| Nguyễn Văn Dũng | 2021 |

# I. Lời mở đầu

## 1. Đề tài

Hệ thống gợi ý phim là một trong những ứng dụng quan trọng của trí tuệ nhân tạo và học máy, được sử dụng rộng rãi trên các nền tảng như Netflix, Amazon Prime, và nhiều dịch vụ xem phim trực tuyến khác. Mục tiêu chính của hệ thống này là đề xuất cho người dùng những bộ phim mà họ có thể thích dựa trên lịch sử xem phim của họ cũng như hành vi và sở thích của người dùng tương tự.

## 2. Hệ thống gợi ý

### 2.1. Khái niệm

Hệ gợi ý (Recommender System) là một hệ thống sử dụng các kỹ thuật và phương pháp của trí tuệ nhân tạo, học máy và xử lý dữ liệu để dự đoán và đề xuất các mục (items) mà người dùng có thể quan tâm. Các mục này có thể là sản phẩm, dịch vụ, nội dung số như phim, âm nhạc, sách, bài viết, và nhiều loại hình khác. Hệ gợi ý được sử dụng rộng rãi trong các ngành công nghiệp như thương mại điện tử, truyền thông, giải trí và mạng xã hội.

### 2.2. Mục tiêu

Mục tiêu chính của hệ gợi ý là cải thiện trải nghiệm người dùng bằng cách cá nhân hóa các đề xuất dựa trên sở thích và hành vi của họ. Bằng cách này, hệ gợi ý giúp:

**- Tăng cường sự hài lòng của người dùng**: Đưa ra các gợi ý phù hợp với sở thích của người dùng, giúp họ tìm thấy những mục thú vị một cách dễ dàng.

**- Tăng doanh thu và tương tác**: Khuyến khích người dùng mua sắm hoặc tương tác nhiều hơn bằng cách đề xuất các sản phẩm hoặc nội dung liên quan.

**- Khám phá các mục mới**: Giúp người dùng khám phá những mục mà họ chưa biết đến nhưng có thể quan tâm.

### 2.3. Các phương pháp của hệ gợi ý

Có hai phương pháp chính được sử dụng trong hệ gợi ý:

**- Lọc cộng tác (Collaborative Filtering)**:

**+ Dựa trên người dùng (User-Based)**: Đề xuất các mục dựa trên sở thích của những người dùng có hành vi tương tự.

**+ Dựa trên mục (Item-Based)**: Đề xuất các mục dựa trên sự tương đồng giữa các mục với nhau.

**- Lọc nội dung (Content-Based Filtering)**: Đề xuất các mục dựa trên các thuộc tính và đặc điểm của chính các mục đó. Ví dụ: đề xuất các phim có cùng thể loại hoặc cùng đạo diễn.

# II. Cơ sở lý thuyết

## 1. Học máy

### 1.1. Khái niệm

**Học máy (machine learning)** là một lĩnh vực của trí tuệ nhân tạo (AI) tập trung vào việc xây dựng các mô hình và thuật toán cho phép máy tính học hỏi từ dữ liệu và cải thiện hiệu suất của chúng theo thời gian mà không cần phải lập trình một cách rõ ràng. Thay vì được lập trình cụ thể để thực hiện một nhiệm vụ, máy tính sử dụng các thuật toán học máy để tự động học hỏi và rút ra các mẫu hoặc xu hướng từ dữ liệu.

### 1.2. Tập dữ liệu tập luyện và kiểm tra

- Tập Dữ Liệu Tập Luyện (Train Set)

Khái niệm: Tập dữ liệu tập luyện là tập hợp các ví dụ được sử dụng để huấn luyện mô hình học máy. Dữ liệu trong tập luyện bao gồm cả đầu vào (input) và đầu ra (output) mong muốn.

Vai trò: Mô hình sử dụng dữ liệu này để tìm hiểu mối quan hệ giữa các biến đầu vào và đầu ra, điều chỉnh các tham số của mình sao cho có thể đưa ra dự đoán chính xác.

- Tập Dữ Liệu Kiểm Tra (Test Set)

Khái niệm: Tập dữ liệu kiểm tra là tập hợp các ví dụ được sử dụng để đánh giá hiệu suất của mô hình sau khi nó đã được huấn luyện. Dữ liệu trong tập kiểm tra thường không được sử dụng trong quá trình huấn luyện.

Vai trò: Mô hình sử dụng dữ liệu này để kiểm tra xem nó có thể tổng quát hóa tốt như thế nào với dữ liệu mới, tức là có thể dự đoán chính xác các đầu ra mà nó chưa từng thấy trước đó.

### 1.3. Xác thực chéo (Cross-Validation)

- Khái niệm: **Cross-validation (xác thực chéo)** là một kỹ thuật được sử dụng trong học máy để đánh giá độ chính xác của mô hình và đảm bảo rằng mô hình không bị overfitting (quá khớp) hoặc underfitting (không đủ khớp). Cross-validation giúp tận dụng tối đa dữ liệu có sẵn để huấn luyện và kiểm tra mô hình, đảm bảo rằng mô hình có thể tổng quát hóa tốt trên các dữ liệu chưa từng thấy.

- Phương pháp:

+ K-Fold: Chia tập dữ liệu thành k phần (fold) có kích thước gần bằng nhau. Lặp lại k lần: trong mỗi lần lặp, sử dụng k-1 phần để huấn luyện mô hình và 1 phần để kiểm tra mô hình. Lấy trung bình các kết quả kiểm tra để đánh giá hiệu suất của mô hình.

+ Leave-One-Out Cross-Validation (LOOCV): Mỗi mẫu dữ liệu sẽ lần lượt được sử dụng làm tập kiểm tra trong khi các mẫu còn lại được sử dụng để huấn luyện. Lặp lại quá trình này cho tất cả các mẫu trong tập dữ liệu. Tính trung bình các kết quả kiểm tra để đánh giá mô hình.

- Mục tiêu:

**Đánh giá Độ Tin Cậy của Mô Hình**: Giúp đưa ra ước lượng chính xác hơn về hiệu suất của mô hình so với việc chỉ sử dụng một tập kiểm tra duy nhất.

**Giảm Nguy Cơ Overfitting**: Bằng cách sử dụng nhiều tập kiểm tra khác nhau, cross-validation giúp đảm bảo rằng mô hình không chỉ học tốt trên dữ liệu huấn luyện mà còn trên dữ liệu mới.

**Sử Dụng Tối Đa Dữ Liệu**: Giúp tận dụng tối đa dữ liệu có sẵn, đặc biệt hữu ích khi có ít dữ liệu.

### 1.4. Quy trình

- Thu thập và tiền xử lý dữ liệu

- Chia dữ liệu: (70-80% tập luyện và 20-30% tập kiểm tra)

- Huấn luyện mô hình

- Đánh giá mô hình

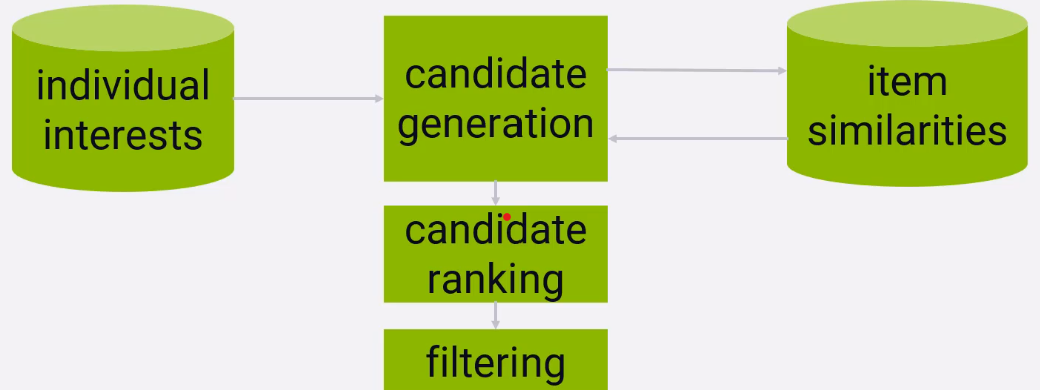
- Triển khai và cải tiến

## 2. Kiến trúc Top-N

### 2.1. Khái niệm

Gợi ý Top-N (Top-N Recommendation) là một loại hệ thống gợi ý trong đó mục tiêu là đề xuất một danh sách ngắn các mục tốt nhất (thường là N mục) mà người dùng có thể quan tâm. Ví dụ, trong một hệ thống gợi ý phim, mục tiêu là đưa ra danh sách N bộ phim mà người dùng có khả năng sẽ thích dựa trên sở thích và hành vi của họ.

### 2.2. Quy trình



- Xác định sở thích cá nhân, dựa vào lịch sử đánh giá phim của người dùng

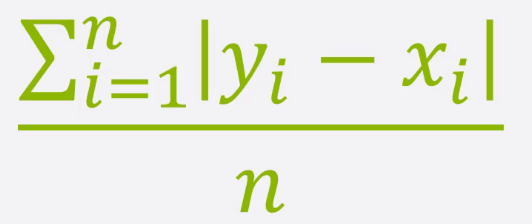
- Tạo danh sách ứng viên (candidate generation): Từ dữ liệu về sở thích cá nhân, hệ thống sẽ tạo ra một danh sách các mục ứng viên có khả năng phù hợp với người dùng dựa trên nội dung và cộng tác.

- Tính toán độ tương đồng: Trong bước này, hệ thống sẽ tính toán độ tương đồng giữa các mục để xác định những mục nào là tương tự nhau.

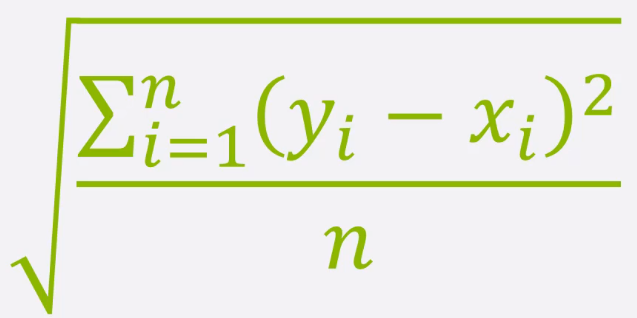
- Lọc kết quả

## 2. Thông số đánh giá

- MAE (Mean Absolute Error): MAE là giá trị trung bình của các lỗi tuyệt đối giữa giá trị dự đoán và giá trị thực tế. Nó đo lường mức độ chênh lệch trung bình giữa các dự đoán của mô hình và các giá trị thực tế. Giá trị MAE càng thấp, mô hình càng chính xác.

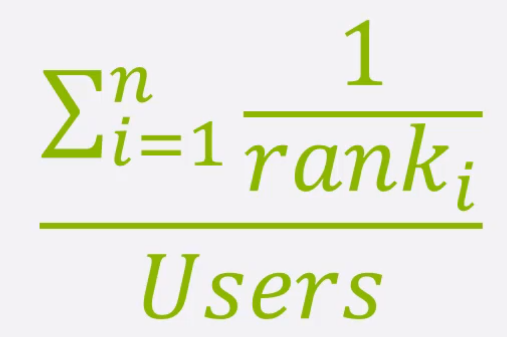


- RMSE (Root Mean Square Error): RMSE là căn bậc hai của giá trị trung bình của các lỗi bình phương giữa giá trị dự đoán và giá trị thực tế. Nó nhạy cảm với các lỗi lớn hơn so với MAE. Giá trị RMSE càng thấp, mô hình càng tốt. RMSE = 0 là tốt nhất.



- HR (Hit Rate): Tỷ lệ trúng đo lường tỷ lệ dự đoán chính xác của hệ thống gợi ý, tức là hệ thống gợi ý một mục và mục đó thực sự được người dùng ưa thích hoặc tương tác. HR càng cao càng tốt.

- AHRH (Average Reciprocal Hit Rate): ARHR là một biến thể của Hit Rate, tính đến vị trí của mục được trúng trong danh sách gợi ý. Nó đánh trọng số cao hơn cho các mục xuất hiện ở các vị trí đầu. ARHR càng cao càng tốt.



- cHR (Cumulative Hit Rate): Cumulative Hit Rate là tổng hợp của các Hit Rate qua các danh sách gợi ý khác nhau. Nó đánh giá khả năng của hệ thống gợi ý trên toàn bộ tập hợp các danh sách gợi ý. cHR càng cao càng tốt.

- rHR (Rating Hit Rate): Rating Hit Rate đo lường tỷ lệ các gợi ý có xếp hạng đúng với mong đợi của người dùng. Nó đánh giá độ chính xác của các xếp hạng trong hệ thống gợi ý. rHR càng cao càng tốt.

- Coverage: Coverage là tỷ lệ các mục được hệ thống gợi ý đến tổng số mục trong tập dữ liệu. Nó đánh giá phạm vi bao phủ của hệ thống gợi ý. Coverage càng cao càng tốt, cho thấy hệ thống có khả năng gợi ý nhiều mục khác nhau.

- Diversity: Diversity đánh giá mức độ khác nhau giữa các mục trong danh sách gợi ý. Nó quan trọng để tránh gợi ý các mục quá giống nhau. Diversity càng cao càng tốt, cho thấy hệ thống gợi ý đa dạng các mục khác nhau.

- Novelty: Novelty đánh giá mức độ mới mẻ của các mục trong danh sách gợi ý, tức là hệ thống gợi ý các mục mà người dùng chưa từng thấy hoặc ít tương tác. Novelty càng cao càng tốt, giúp người dùng khám phá các mục mới.

- Churn: Churn đo lường mức độ thay đổi trong danh sách gợi ý qua các lần gợi ý khác nhau. Nó quan trọng để đảm bảo rằng hệ thống gợi ý không quá tĩnh. Churn vừa phải là tốt, quá cao hoặc quá thấp đều không tốt.

- Responsiveness: Responsiveness đánh giá tốc độ phản hồi của hệ thống gợi ý khi có thay đổi trong hành vi người dùng. Responsiveness càng cao càng tốt, cho thấy hệ thống có thể điều chỉnh nhanh chóng theo thay đổi của người dùng.

- A/B Tests: A/B Tests là phương pháp so sánh hiệu quả giữa hai phiên bản của hệ thống gợi ý (phiên bản A và phiên bản B) trên một tập hợp người dùng ngẫu nhiên. Dựa vào kết quả A/B Tests, chọn phiên bản có hiệu suất tốt hơn dựa trên các chỉ số đánh giá như HR, MAE, RMSE, etc.

## 3. Content-based Filtering

### 3.1. Khái niệm

Content-Based Filtering (lọc dựa trên nội dung) là một kỹ thuật gợi ý dựa trên các thuộc tính của các mục mà người dùng đã thể hiện sự quan tâm hoặc tương tác trước đó. Hệ thống sử dụng thông tin về nội dung của các mục (như thể loại phim, tác giả, diễn viên, vv.) và tạo ra hồ sơ sở thích của người dùng dựa trên những thông tin này.

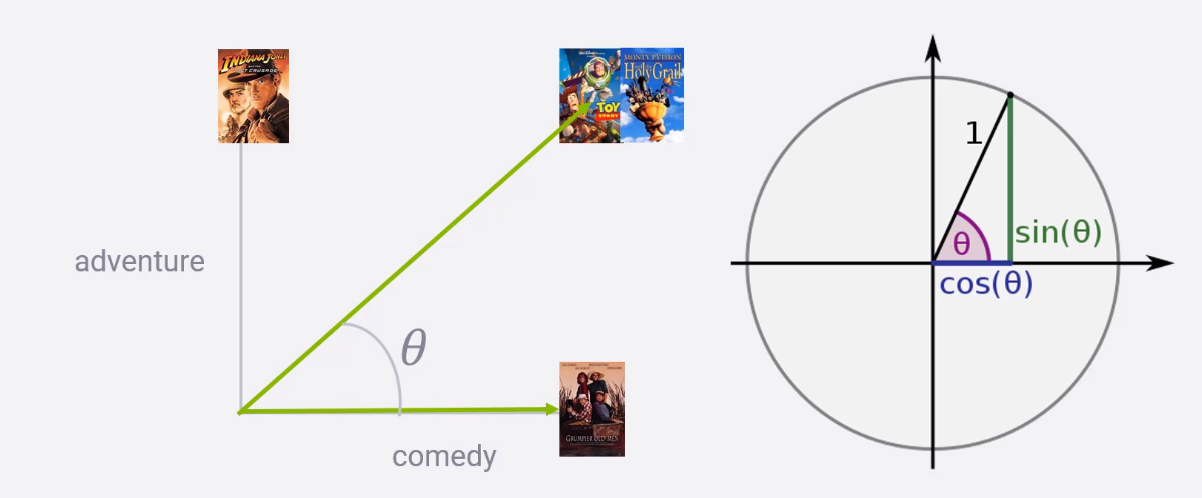
### 3.2. Nguyên lý và ứng dụng

- Biểu diễn nội dung của các bộ phim: Mỗi phim có 18 thể loại (genres) có thể thuộc vào, mỗi thể loại sẽ là 1 chiều của vector

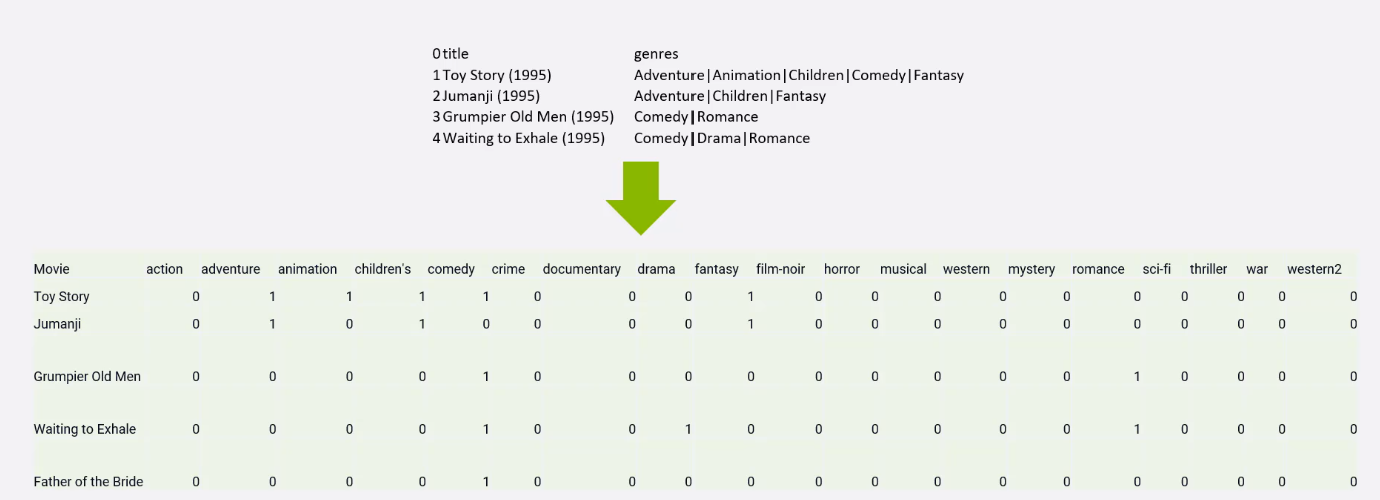


- Biểu diễn người dùng: Mỗi người dùng được biểu diễn bằng một hồ sơ (profile) của các bộ phim họ đã tương tác trước đó.

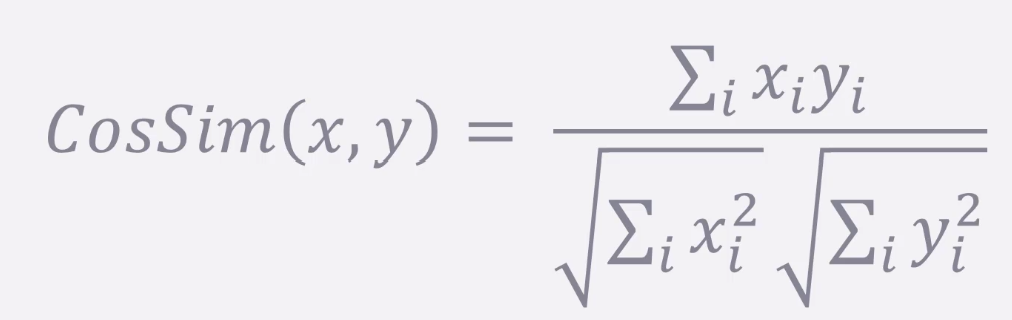
- Tính toán độ tương tự: Độ tương tự giữa các mặt hàng được tính bằng cách so sánh đặc trưng hoặc thuộc tính của chúng. Phương pháp tính độ tương tự được sử dụng bao gồm cosine similarity:



Chuyển phim thành các vector:



Công thức tính độ tương tự:



Thời gian phát hành của phim:



- Đề xuất mặt hàng tương tự: k-nearest-neighbors (đưa ra đánh giá dự đoán của một người dùng cho một bộ phim họ chưa đánh giá)

+ Tính toán độ tương tự của bộ phim đang xét đối với tất cả các bộ phim đã được người dùng đánh giá

+ Các bộ phim gần nhất có độ tương đồng cao nhất

+ Lấy trung bình độ tương đồng với đánh giá của bộ phim nó tương đồng với

## 4. Neighborhood-based Collaborative Filtering

### 4.1. Khái niệm

Collaborative Filtering (lọc cộng tác) là một kỹ thuật gợi ý dựa trên hành vi và sở thích của người dùng khác. Phương pháp này không cần thông tin về nội dung của các mục, mà dựa trên tương tác (ví dụ: xếp hạng, click) giữa các người dùng và các mục.

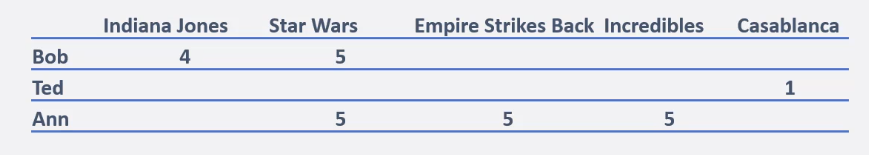
### 4.2. Nguyên lý hoạt động

- User-based collaborative filtering:



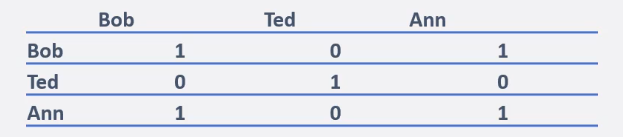
Tìm những người dùng tương đồng lịch sử hoạt động và gợi ý phim chưa tương tác

Ma trận dữ liệu:



Chuyển các user thành các vector với chiều là những bộ phim, trọng số là đánh giá (VD: Bob (4, 5, 0, 0, 0)) => Tương đồng qua công thức cosin

Ma trận tương đồng:



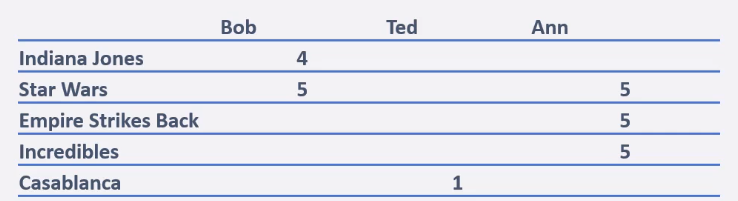
=> Đánh giá các hàng xóm của Bob: Ann - 1.0; Ted - 0

Gợi ý các bộ phim với điểm cao nhất (đánh giá \* trọng số tương đồng \* số lần xuất hiện (trong TH có nhiều người dùng khác tham chiếu)

- Item-based colaborative filtering:

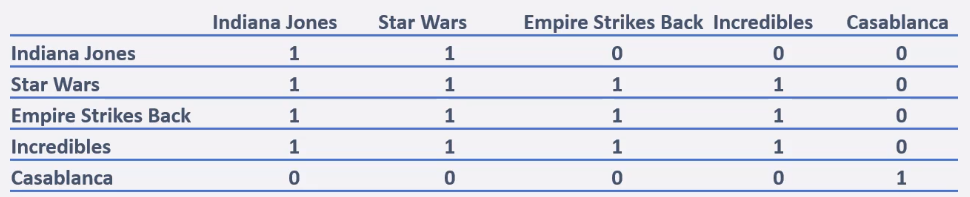
Gợi ý những phim tương đồng với phim đã tương tác

Ma trận dữ liệu:



Chuyển các bộ phim thành vector với chiều là những người dùng, trọng số là đánh giá (VD: Indiana Jones (4, 0, 0)) => Tương đồng qua công thức cosin

Ma trận tương đồng:



\* Chú ý: các chỉ số trong ma trận tương đồng của 2 VD trong thực tế có trọng số ý nghĩa hơn chỉ 0 và 1.

# III. Triển khai

## 1. Công nghệ sử dụng

## 2. Thu thập và tiền xử lý dữ liệu

## 3. Framework đánh giá mô hình

## 4. Content-based Filtering

## 5. User-based Collaborative Filtering

## 6. Item-based Collaborative Filtering

# IV. Đánh giá và nhận xét thực tế

# V. Phân công công việc

|  |  |  |
| --- | --- | --- |
| **Thành viên** | **Nhiệm vụ** | **Đóng góp** |
| Phạm Việt Hải  (Leader) | - Xử lý Backend  - Hỗ trợ FE  - Hoàn thành báo cáo | 25% |
| Trần Bình Minh | - Quản lý tiến độ  - Quản lý cấu trúc, phân chia các chức năng  - Hỗ trợ BE  - Phân chia nhiệm vụ  - Góp ý Design Concepts | 25% |
| Nguyễn Lê Sơn | - Xử lý back-end  - Kiểm thử API  - Góp ý Design Principles  - Hoàn thành báo cáo | 25% |
| Nguyễn Văn Dũng | - Xử lý Frontend chính  - Hoàn thành báo cáo | 25% |

# VI. Phụ lục