**TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI**

**BÁO CÁO TIẾN ĐỘ GR1**

**Tuần 1 + 2**

**Thiết kế hệ thống gợi ý phim cho người dùng sử dụng AI**

**PHẠM VIỆT HẢI**

Hai.PV215044@sis.hust.edu.vn

**Ngành Công nghệ Thông tin và Truyền thông Việt – Nhật**

|  |  |
| --- | --- |
| **Giảng viên hướng dẫn:** | TS. Nguyễn Đức Toàn  Chữ ký của GVHD |
|  |  |
| **Trường:** | Công nghệ Thông tin và Truyền thông |
| **HÀ NỘI, 3/2024** | |

**ĐỀ TÀI TỐT NGHIỆP**

Hệ Thống Đề Xuất Phim Sử Dụng Trí Tuệ Nhân Tạo (AI) Cho Ứng Dụng Và Web

Mô Tả:

Trong thời đại số hóa hiện nay, các trang thương mại điện tử đang trở thành một phần quan trọng của cuộc sống hàng ngày, giúp người dùng tìm kiếm và mua sắm sản phẩm một cách thuận tiện. Một trong những yếu tố quan trọng đối với sự thành công của các trang web thương mại điện tử là khả năng đề xuất sản phẩm phù hợp với sở thích và nhu cầu của người dùng.

Đề tài này nhằm mục đích xây dựng một hệ thống đề xuất sản phẩm thông minh sử dụng trí tuệ nhân tạo để giúp người dùng khám phá và mua sắm sản phẩm một cách hiệu quả hơn trên một trang thương mại điện tử.

# I. Tìm hiểu sơ bộ

## 1. Khái niệm:

- Hệ thống gợi ý sản phẩm sử dụng trí tuệ nhân tạo là một ứng dụng máy học được thiết kế để tự động đề xuất các sản phẩm hoặc dịch vụ phù hợp nhất với từng người dùng dựa trên hành vi mua sắm trước đó, sở thích cá nhân, và các yếu tố khác.

## 2. Ví dụ:

- Amazon: Là một trong những trang thương mại điện tử lớn nhất thế giới và sử dụng trí tuệ nhân tạo để gợi ý sản phẩm cho người dùng dựa trên lịch sử mua hàng, đánh giá sản phẩm, và hành vi duyệt trang.

- Netflix: Sử dụng hệ thống gợi ý dựa trên trí tuệ nhân tạo để đề xuất các bộ phim và chương trình truyền hình phù hợp với sở thích của người dùng. Hệ thống này phân tích lịch sử xem phim, đánh giá, và thông tin về thể loại để đưa ra các gợi ý cá nhân hóa.

- YouTube: Sử dụng trí tuệ nhân tạo để đề xuất video cho người dùng dựa trên lịch sử xem video, tương tác (like, dislike, subscribe), và thông tin về thể loại.

## 3. Phân loại:

- Cá nhân hóa:

+ Khái niệm: Gợi ý cá nhân hóa là một phương pháp trong lĩnh vực marketing và trải nghiệm người dùng, mà thông qua đó, các tổ chức và doanh nghiệp cung cấp thông tin, sản phẩm hoặc dịch vụ được tùy chỉnh dựa trên thông tin cá nhân và sở thích của từng cá nhân.

+ Đặc điểm:

* Cá Nhân Hóa: Gợi ý cá nhân hóa tạo ra các đề xuất hoặc nội dung được cá nhân hóa đặc biệt cho từng người dùng, thay vì cung cấp thông tin chung cho tất cả mọi người.
* Dựa trên Dữ Liệu: Gợi ý cá nhân hóa dựa trên việc thu thập và phân tích dữ liệu về hành vi truy cập, tương tác, lịch sử mua hàng, đánh giá, hoặc thông tin cá nhân của người dùng.
* Tùy Chỉnh và Tương Tác: Người dùng thường có thể tùy chỉnh và điều chỉnh các gợi ý cá nhân hóa dựa trên sở thích và phản hồi của họ, tạo ra một trải nghiệm tương tác linh hoạt và được điều chỉnh.
* Tăng Cơ Hội và Hiệu Quả: Gợi ý cá nhân hóa tăng cơ hội tiếp cận và tương tác với thông tin hoặc sản phẩm phù hợp nhất, từ đó tăng hiệu quả và khả năng chuyển đổi.
* Đa Dạng và Linh Hoạt: Các gợi ý cá nhân hóa có thể áp dụng trên nhiều lĩnh vực, từ mua sắm trực tuyến, nội dung giáo dục, tới quảng cáo trực tuyến và dịch vụ cá nhân hóa.

- Phi cá nhân hóa (Mục tiêu):

+ Khái niệm: Gợi ý phi cá nhân hóa đề cập đến việc cung cấp các đề xuất hoặc thông tin một cách không tùy chỉnh dựa trên thông tin chung hoặc không quan trọng về người dùng. Trái ngược với gợi ý cá nhân hóa, gợi ý phi cá nhân hóa không xem xét các thông tin cá nhân hay sở thích cụ thể của từng cá nhân mà thay vào đó tập trung vào một phần thông tin rộng rãi hoặc thông tin mặc định.

+ Đặc điểm:

* + Thông Tin Chung: Gợi ý phi cá nhân hóa cung cấp thông tin hoặc sản phẩm dựa trên tiêu chí rộng rãi hoặc thông tin chung về người dùng, chứ không phụ thuộc vào thông tin cá nhân hoặc sở thích cụ thể.
  + Không Tùy Chỉnh: Không có sự tùy chỉnh hay điều chỉnh dựa trên phản hồi hoặc hành vi của từng người dùng, mà thay vào đó, các đề xuất được cung cấp một cách đồng nhất cho tất cả mọi người.
  + Hiển Thị Đồng Đều: Thông tin hoặc sản phẩm được hiển thị đồng đều cho tất cả người dùng, không có sự phân loại hay ưu tiên dựa trên các yếu tố cá nhân.
  + Tiêu Chuẩn Hóa: Các gợi ý phi cá nhân hóa thường được tạo ra theo tiêu chuẩn hoặc mẫu mực chung, không có sự cá nhân hóa cụ thể.
  + Ít Phản Hồi: Do không có sự tùy chỉnh cá nhân, gợi ý phi cá nhân hóa thường ít gây ra phản hồi tích cực từ phía người dùng.

## 4. Dữ liệu:

- Dữ liệu Rõ Ràng (Explicit Data):

+ Đây là dữ liệu mà người dùng cung cấp một cách rõ ràng thông qua hành động hoặc phản hồi của họ.

+ Ví dụ: Đánh giá sản phẩm, phản hồi từ khách hàng, thông tin về đối tượng như tuổi, giới tính, v.v.

- Dữ liệu Ẩn (Implicit Data):

+ Đây là dữ liệu được thu thập một cách gián tiếp từ hành vi hoặc tương tác của người dùng mà không cần sự tương tác trực tiếp.

+ Ví dụ: Lịch sử mua hàng, lịch sử click hoặc duyệt web, thời gian mà người dùng đã dành cho một sản phẩm cụ thể.

- Thông Tin Về Sản Phẩm:

+ Đây là dữ liệu về sản phẩm được sử dụng để mô tả và phân loại sản phẩm một cách chi tiết.

+ Ví dụ: Cấu trúc phân loại sản phẩm, các thuộc tính của sản phẩm như kích thước, màu sắc, giá cả, mô tả sản phẩm.

# II. Triển khai dự án

## 1. Mô hình

- Đánh giá của khách hàng:

* + - Implicit (ngụ ý): Đánh giá ngụ ý là những dữ liệu về hành vi tương tác của người dùng với hệ thống mà không có sự đánh giá rõ ràng hoặc phản hồi từ phía người dùng. (VD: Thời lượng xem phim, mua phim, ...)
    - Explicit (rõ ràng): Đánh giá rõ ràng là những phản hồi mà người dùng cung cấp trực tiếp về sở thích hoặc chất lượng của một sản phẩm. (VD: Tương tác cảm xúc, đánh giá, bình luận, ...)

- Hệ thống gợi ý Top-N: Trong lĩnh vực gợi ý sản phẩm, hệ thống này nhằm mục đích đề xuất cho người dùng một danh sách các sản phẩm hoặc mục được xếp hạng theo mức độ phù hợp cao nhất dựa trên sở thích hoặc hành vi của người dùng. (N ở đây là số sản phẩm đứng đầu)

- Đánh giá hệ thống:

+ Train/Test (Huấn luyện/Kiểm thử): Quá trình chia dữ liệu thành hai phần: một phần được sử dụng để huấn luyện mô hình và phần còn lại được sử dụng để kiểm tra hiệu suất của mô hình. Dữ liệu huấn luyện (train data) được sử dụng để huấn luyện mô hình, tức là để học các mẫu và mối quan hệ trong dữ liệu. Dữ liệu kiểm thử (test data) được sử dụng để đánh giá hiệu suất của mô hình, tức là để kiểm tra xem mô hình có thể dự đoán chính xác trên dữ liệu mới không.

+ Cross-validation (Kiểm thử chéo): Phương pháp để đánh giá hiệu suất của mô hình bằng cách chia dữ liệu thành các tập con, huấn luyện mô hình trên một số tập con và kiểm tra trên các tập con còn lại. Cross-validation giúp đánh giá hiệu suất của mô hình một cách trung thực hơn bằng cách sử dụng tất cả dữ liệu có sẵn để huấn luyện và kiểm tra, và tránh được sự phụ thuộc vào việc chia dữ liệu một cách ngẫu nhiên.

+ Thông số chính xác:

* + - * RMSE(Root Mean Square Error)/MAE(Mean Absolute Error): Hai phép đo độ lỗi phổ biến được sử dụng trong machine learning và các lĩnh vực liên quan.
      * HR (Hit Rate)/ARHR (Average Reciprocal Hit Rate)/cHR (cumulative Hit Rate)/rHR (rating Hit Rate): Đại lượng đo sự chính xác dựa trên tương tác thực của người dùng
      * Coverage (độ phổ cập): Lượng dữ liệu và độ rộng truy cập của thuật toán
      * Diversity (độ đa dạng): Sự đa dạng của dữ liệu đầu ra
      * Novelty (độ mới lạ): Sự mới lạ trong đề xuất
      * Churn (độ thay đổi): Độ thay đổi của đầu ra
      * Resposiveness (độ thích ứng): Tốc độ phản ứng của hệ thống khi gặp sự tương tác của người dùng

## 2. Dữ liệu đầu vào

- User ID: Người dùng được chọn ngẫu nhiên để bao gồm. Các id của họ đã được ẩn danh. Id của người dùng là nhất quán giữa `ratings.csv` và `tags.csv` (tức là cùng một id tham chiếu đến cùng một người dùng trong hai tệp này).

- Movie Ids: Chỉ các bộ phim có ít nhất một đánh giá hoặc một thẻ được bao gồm trong tập dữ liệu. Các id phim này là nhất quán với các id được sử dụng trên trang web MovieLens (ví dụ, id `1` tương ứng với URL <https://movielens.org/movies/1>). Id phim là nhất quán giữa `ratings.csv`, `tags.csv`, `movies.csv`, và `links.csv` (tức là cùng một id tham chiếu đến cùng một phim trong bốn tệp dữ liệu này).

- Cấu trúc tệp dữ liệu Đánh giá (ratings.csv): Tất cả các đánh giá được chứa trong tệp `ratings.csv`. Mỗi dòng của tệp này sau dòng tiêu đề đại diện cho một đánh giá của một bộ phim của một người dùng, và có định dạng sau: userId, movieId, rating, timestamp. Các dòng trong tệp này được sắp xếp trước theo userId, sau đó, bên trong người dùng, theo movieId. Đánh giá được thực hiện trên một thang điểm 5 sao, với các bước nhảy bán sao (0,5 sao - 5,0 sao). Timestamps đại diện cho số giây kể từ nửa đêm Giờ thống nhất thế giới (UTC) ngày 1 tháng 1 năm 1970.

- Cấu trúc tệp dữ liệu Thẻ (tags.csv): Tất cả các thẻ được chứa trong tệp `tags.csv`. Mỗi dòng của tệp này sau dòng tiêu đề đại diện cho một thẻ được áp dụng cho một bộ phim của một người dùng, và có định dạng sau: userId,movieId,tag,timestamp Các dòng trong tệp này được sắp xếp trước theo userId, sau đó, bên trong người dùng, theo movieId. Tags là dữ liệu metadata do người dùng tạo ra về các bộ phim. Mỗi thẻ thường là một từ đơn hoặc cụm từ ngắn. Ý nghĩa, giá trị và mục đích của mỗi thẻ cụ thể được xác định bởi từng người dùng. Timestamps đại diện cho số giây kể từ nửa đêm Giờ thống nhất thế giới (UTC) ngày 1 tháng 1 năm 1970.

- Cấu trúc tệp dữ liệu Phim (movies.csv): Thông tin về phim được chứa trong tệp `movies.csv`. Mỗi dòng của tệp này sau dòng tiêu đề đại diện cho một bộ phim, và có định dạng sau: movieId, title, genres. Các tiêu đề phim được nhập thủ công hoặc nhập từ <https://www.themoviedb.org/>, và bao gồm năm phát hành trong dấu ngoặc đơn. Có thể tồn tại lỗi và không nhất quán trong các tiêu đề này. Thể loại là một danh sách được phân tách bằng dấu gạch chéo, và được chọn từ các thể loại sau:

* + - Hành động
    - Phiêu lưu
    - Hoạt hình
    - Trẻ em
    - Hài kịch
    - Tội phạm
    - Tài liệu
    - Kịch
    - Huyền bí
    - Film-Noir
    - Kinh dị
    - Âm nhạc

## 3. Thuật toán sử dụng

- KNN (K Nearest Neighbors): Thuật toán máy học có giám sát được sử dụng cho cả các vấn đề phân loại và hồi quy. Ý tưởng chính của thuật toán là dự đoán nhãn của một điểm dữ liệu mới dựa trên nhãn của các điểm dữ liệu lân cận của nó trong không gian đặc trưng. KNN được coi là một thuật toán "lười biếng" (lazy learning) vì nó không học một mô hình cụ thể từ dữ liệu đào tạo mà thay vào đó lưu trữ toàn bộ dữ liệu đào tạo và thực hiện tính toán khi cần thiết.

* + - Xác định số lân cận (K): Đầu tiên, bạn cần xác định số lân cận (K), tức là số điểm dữ liệu gần nhất mà bạn muốn sử dụng để dự đoán nhãn cho một điểm dữ liệu mới.
    - Tính khoảng cách: Sau đó, tính khoảng cách giữa điểm dữ liệu mới và tất cả các điểm dữ liệu trong tập huấn luyện. Các phép đo khoảng cách thông thường bao gồm khoảng cách Euclid và khoảng cách Manhattan.
    - Chọn các lân cận: Chọn K điểm dữ liệu gần nhất (có khoảng cách nhỏ nhất) với điểm dữ liệu mới.
    - Phân loại hoặc dự đoán: Dựa trên nhãn của các điểm dữ liệu lân cận đã chọn, phân loại hoặc dự đoán nhãn của điểm dữ liệu mới. Trong trường hợp phân loại, chúng ta sử dụng đa số phiếu để xác định nhãn của điểm dữ liệu mới. Trong trường hợp hồi quy, chúng ta sử dụng giá trị trung bình của các giá trị nhãn của các điểm dữ liệu lân cận.