# **LỜI MỞ ĐẦU**

**Lý do chọn đề tài**

Với sự phát triển của xã hội hiện nay, việc nhu cầu giải trí của con người ngày càng tăng cao trong số đó phải kể đến là nhu cầu giải trí thông qua phim ảnh. Với lịch sử lâu đời của ngành điện ảnh cùng với kho tài liệu phim ảnh khổng lồ đã được sản xuất qua năm tháng, cùng với đó là sự phát triển của công nghệ hiện đại ngày nay đã cho phép các bộ phim có thể được lưu trữ trong các cơ sở dữ liệu lớn và được sử dụng để phục vụ cho lượng lớn khán giả trong hầu hết các trang web xem phim trực tuyến. Nhưng với số lượng phim khổng lồ và hàng ngàn thể loại phim khác nhau như thế sẽ là một vấn đề đối với các nhà sản xuất phim ảnh, họ có thể sẽ không thể thu lại lợi nhuận từ phim của họ nếu bộ phim mà họ sản xuất không được đến được tay khán giả hoặc không được nhiều người chú ý đến. Do đó, Hệ thống Khuyến Nghị là một trong những giải pháp ứng dụng phù hợp và tốt nhất để giải quyết cho vấn đề trên, nhằm thu hút nhiều khán giả đến với các sản phẩm phim hơn, đồng thời cũng là cầu nối cho khán giả đến với các bộ phim.

Và ngày nay, hệ thống Khuyến Nghị (Recommender System) là một trong những lớp ứng dụng thành công và phổ biến nhất của trí tuệ nhân tạo. Với các nền tảng dịch vụ trực tuyến đang ngày càng phát triển mạnh mẽ trong đời sống hiện nay thì Hệ thống Khuyến Nghị đóng một vai trò rất lớn trong việc ứng dụng vào các ngành Dịch vụ của con người, Ví dụ: Thương mại điện tử, mua bán các sản phẩm dịch vụ trực tuyến, ứng dụng trực tuyến, xem phim/video trực tuyến ..v..vv.

Trên các trang web xem phim, hệ thống Khuyến nghị đóng một vai trò chính trong việc giới thiệu các bộ phim đến với các khán giả. Nó đóng vai trò phân tích và tìm hiểu khối dữ liệu cá nhân của người dùng và từ đó đưa ra những dự đoán, gợi ý, đề xuất phù hợp với sở thích của khán giả. Các trang web xem phim lớn hiện nay ứng dụng Hệ thống Khuyến nghị: Netflix, BiliBili, FPT Play.

**Mục đích nghiên cứu**

Đề tài của đồ án chuyên ngành tập trung phân tích và làm rõ cách thức hoạt động của hệ thống khuyến nghị một số trang web xem phim, đồng thời phân tích chi tiết các thuật toán được ứng dụng để giải quyết bài toán khuyến nghị phim.

**Nhiệm vụ nghiên cứu**

Khái quát và tổng quan về Hệ thống Khuyến nghị.

Phân tích về các thuật toán được sử dụng để giải quyết bài toán khuyến nghị phim.

Khảo sát về Hệ thống khuyến nghị trên các trang web xem phim.

**Đối tượng nghiên cứu và phạm vi nghiên cứu**

Đối tượng nghiên cứu: ác thuật toán được sử dụng để giải quyết cho bài toán khuyến nghị phim trên một số trang web xem phim.

Phạm vi nghiên cứu: Trang web xem phim như Netflix, BiliBili, FPT Play.

**Phương pháp nghiên cứu**

Phương pháp quan sát: Quan sát hành vi thu thập thông tin người dùng của một số hệ thống Khuyến nghị nhằm phân tích bài toán khuyến nghị phim.

Phương pháp điều tra: Tìm hiểu cụ thể đặc điểm và tính chất của bài toán Khuyến nghị phim trên các trang web xem phim nhằm đưa ra một bài viết dễ hình dung cho người đọc và có độ chính xác về mặt nội dung cao.

**Giả thuyết khoa học**

Từ cách hoạt động của hệ thông Khuyến nghị trên các trang web xem phim ta giả thích rằng các hệ thống Khuyến nghị phim đựa theo sở thích, thói quen, hành vi của khán giả để đưa ra những dự đoán và gợi ý phù hợp nhất.

**Cấu trúc đồ án chuyên ngành**

Cấu trúc của đồ án chuyên ngành gồm 3 phần chính:

**Chương 1.** **Tổng quan về hệ thống khuyến nghị**

Chương 1 trong đồ án chuyên ngành sẽ giới thiệu tổng quan về hệ thống khuyến nghị, lý thuyết của bài toán khuyến nghị và các phương pháp tiếp cận của hệ thống khuyến nghị. Bên cạnh đó, đồ án chuyên ngành còn giới thiệu thêm về một số đề tài có liên quan của hệ thống khuyến nghị. Sau cùng sẽ giới thiệu bài toán khuyến nghị phim thường được sử dụng trong các hệ thống khuyến nghị trên các trang web xem phim.

**Chương 2. Phân tích bài toán khuyến nghị phim**

Chương 2 trong đồ án chuyên ngành sẽ tập trung phân tích vào các thuật toán được sử dụng phổ biến trong bài toán khuyến nghị phim, trình bày chi tiết các thuật toán PMF (Probabilistic Matrix Factorization), BPMF (Bayesian Probabilistic Matrix Factorization), ALS (Alternating Least Squares).

**Chương 3. Khảo sát hệ thống khuyến nghị phim trên một số trang web xem phim**

Chương 3 trong đồ án chuyên ngành sẽ tập trung vào việc khảo sát cách thức của một hệ thống khuyến nghị hoạt động trên các trang web xem phim. Các trang web được khảo sát trong đồ án chuyên ngành gồm Netflix, BiliBili và FPT Play. Sau đó ta so sánh cách thức xử lý bài toán khuyến nghị phim giữa 3 trang web xem phim trên.

**CHƯƠNG 1. TỔNG QUAN VỀ HỆ THỐNG KHUYẾN NGHỊ**

* 1. **Giới thiệu chung**

Hệ thống khuyến nghị (có tên gọi tiếng anh là Recommender System) hay còn được gọi là hệ thống tư vấn là một hệ thống có nhiệm vụ chọn lọc thông tin nhằm dự đoán sở thích, mức độ phù hợp, mối quan tâm và nhu cầu của người dùng để đưa ra một hoặc nhiều mục, sản phẩm, dịch vụ mà người dùng sẽ quan tâm với xác suất lớn nhất.

Và hiện nay, mọi hệ thống/ứng dụng có hiển thị quảng cáo trên internet đều sử dụng hệ thống khuyến nghị để đưa ra quảng cáo, đề xuất tốt nhất có thể đến cho người dùng… Một vài ví dụ phổ biến và dễ gặp nhất đó là gợi ý sản phẩm/dịch vụ có liên quan trên các trang web/ứng dụng phổ biến.



1.1.1. Ví dụ về hệ thống khuyến nghị của một trang web xem phim

Và để làm được điều đó, hệ thống khuyến nghị đã sử dụng những thuật toán để phân tích và dự doán đựa trên dữ liệu hành vi người dùng được lưu lại. Nhờ đó, những quảng cáo mang tính cá nhân hóa được đưa đến cho người dùng. Hệ thống sẽ biết chính xác từng người dùng sử dụng có nhu cầu gì, muốn gì để từ đó đưa ra khuyến nghị.

Trong thực tế, ý tưởng để những người lập trình xây dựng một hệ thống khuyến nghị không đâu xa lạ chính là xuất phát từ hành vi của người mua hàng và người bán hàng: Khi một người mua hàng có nhu cầu mua một sản phẩm, họ thường sẽ có hành vi hỏi người bán hàng để tư vấn cho họ về sản phẩm mà họ có ý định mua. Người bán hàng sẽ tiến hành thu thập thông tin từ người mua bao gồm: nhu cầu sử dụng, đặc điểm, mức độ phù hợp, chức năng, màu sắc, … đồng thời kết hợp với kiến thức hiểu biết của mình về sản phẩm để đưa ra đề xuất, lời khuyên sản phẩm phù hợp nhất cho người mua. Và ở một mức độ cao hơn, người bán sẽ liên hệ, liên tưởng những người đã từng mua sản phẩm mà có đặc điểm tương đồng vời người mua hiện tại, từ đó họ dự đoán người mua hiện tại có khả năng thích sản phẩm nào nhất để đưa ra khuyến nghị sản phẩm phù hợp nhất cho người mua.

**1.2. Tổng quan bài toán khuyến nghị**

Để có thể xây dựng được một hệ thống khuyến nghị hoàn chỉnh cho từng lĩnh vực cụ thể, các nghiên cứu trước đã phát biểu được lời giải chung cho bài toán khuyến nghị như sau.

Phát biểu bài toán:

**Đầu vào (Input):**

**+** Cho tập người dùng *U*, mỗi người dùng thuộc *U* và có các đặc điểm I = {}.

**+** Một tập sản phẩm, dịch vụ (ở đây ta gọi chung là sản phẩm) P, mỗi sản phẩm có các đặc điểm đặc trưng J = {}.

+ Một ma trận đánh giá tổng quát R = ( với i = 1, … N và j = 1…M, thể hiện mối quan hệ giữa tập người dùng U đối với tập sản phẩm P. Trong đó là đánh giá của người dùng cho sản phẩm , N và M là lần lượt số người dùng và số sản phẩm.



1.2.1. Ví dụ ma trận đánh giá tổng quát Rij

**Đầu ra (Output):**

Danh sách các sản phẩm thuộc *P* có độ phù hợp với người dùng thuộc *U* nhất.

Để giải bài toán này chúng ta cần xây dựng hàm F(, để đo độ phù hợp của sản phẩm với người dùng , từ đó ta sẽ lấy được danh sách các sản phẩm, dịch vụ phù hợp (là các sản phẩm, dịch vụ mà có khả năng được người dùng chọn) nhất.

Và cũng tùy thuộc vào phương pháp sử dụng mà ta có nhiều các xây dựng hàm F khác nhau, các cách xây dựng hàm F phụ thuộc chủ yếu bởi các yếu tố sau:

+ Đặc điểm của người dùng (lọc theo nội dung người dùng): đặc điểm này được đánh giá chủ quan bởi các quy luật tự nhiên hoặc các quy tắc cơ bản. Ví dụ là nữ thì sẽ có xu hướng chọn mua các sản phẩm của nữ hơn là các sản phẩm của nam.

+ Đặc điểm của sản phẩm (lọc theo nội dung sản phẩm): cũng giống như lọc theo nội dung người dùng, các sản phẩm có đặc điểm giống nhau thì cũng có khả năng được người dùng đánh giá như nhau. Ví dụ đặc điểm của các món đồ công nghệ có thể là đặc điểm, tính năng, nhu cầu sử dụng …

+ Lịch sử giao dịch của người dùng : từ lịch sử giao dịch cũng có thể suy ra sản phẩm mà người dùng quan tâm đến, do đó các sản phẩm cùng thể loại, lĩnh vực sẽ có độ liên quan cao hơn. Ví dụ một người đã từng mua áo, giày bóng đá thì có thể dự đoán được được người là một người đam mê bóng đá, thích thể thao. Từ đó suy ra được người này sẽ có khả năng sử dụng dịch vụ hoặc mua các sản phẩm thể thao cao hơn các dịch vụ, sản phẩm khác.

+ Những người dùng khác có cùng đặc điểm giống với người dùng : với quan niệm rằng những người dùng giống nhau sẽ thích, đánh giá những sản phẩm giống nhau. Các đặc điểm bao gồm tập đặc điểm *I*ban đầu, kết hợp với các đặc điểm cộng tác như cùng mua mặt hàng nào đó hoặc có các hành vi mua hàng giống nhau… Việc tìm hiểu những mặt hàng, dịch vụ mà đã từng quan tâm sẽ đưa ra được những gợi ý phù hợp cho người dùng .

**1.3. Các hướng tiếp cận của bài toán khuyến nghị**

**1.3.1. Phương pháp lọc dựa trên nội dung**

**Tổng quan:** Lọc dựa trên nội dung (tên tiếng anh là Content-Base Filtering) là phương pháp thực hiện dựa trên việc so sánh nội dung thông tin hay mô tả hàng hóa, để tìm ra những sản phẩm tương tự với những gì mà người dùng đã từng quan tâm để giới thiệu cho họ những sản phẩm này

A diagram of a movie

Description automatically generated

1.3.1.1 Ví dụ mô hình kỹ thuật lọc dựa trên nội dung

Trong hình 1.3.1.1 ta có thể thấy mô hình minh họa cho việc lọc theo nội dung như sau: User A đánh giá thích Movie A và Movie A có thể loại Love, Romantic. Do đó, phương pháp lọc theo nội dung sẽ dựa theo Type (Thể loại) của Movie A và từ đó khuyến nghị Movie C có cùng Type (Thể loại) với Movie A cho User A.

**Bài toán:** Để thực hiện việc ước lượng xem có hay không một người dùng *u* có thích đối tượng sản phẩm *p.* Ta xây dựng một hàm phù hợp *f(u,p)* của các sản phẩm khuyến nghị *p* với người dùng *u* và ước lượng giá trị phù hợp này. Các phương pháp tiếp cận nội dung thường sẽ thực hiện các bước sau đây:

* **Bước 1:** Biễu diện nội dung đối tượng khuyến nghị *pP*, ký hiệu *Content(p)*
* **Bước 2:** Mô hình hóa sở thích người dùng *uU*, gọi tắt là hồ sơ người dùng (User’s Profile), ký hiệu *UserProfile(u)*
* **Bước 3:** Ước lượng giá trị phù hợp dựa trên độ tương tự nội dung của sản phẩm khuyến nghị *p* với hồ sơ người dùng *u*. Hệ thống sẽ ưu tiên khuyến nghị những đối tượng sản phẩm *p* có nội dung tương tự cao so với hồ sơ người dùng *p.*

Phương pháp lọc dựa trên nội dung có thể được chia ra làm 2 nhóm chính: Một là phương pháp lọc dựa trên bộ nhớ, thực hiện tính toán độ tương tự giữa *Content(p)* và *UserProfile(u)* dùng các độ do lường tương tự Consine, Euclide; Hai là phương pháp lọc dựa trên mô hình, với mô hình được học từ dữ liệu dùng các kỹ thuật học máy giám sát để phân các sản phẩm khuyến nghị thành những sản phẩm được người dùng quan tâm hoặc không quan tâm.

* **Ưu điểm của phương pháp lọc dựa trên nội dung:**

+ Là phương pháp trực quan, dễ dàng hiểu và giải thích được.

+ Không bị ảnh hưởng bởi khởi đầu lạnh (cold start)

+ Không bị ảnh hưởng bởi vấn đề vấn đề thưa thớt dữ liệu.

+ Có thể khuyến nghị những sản phẩm mới hoặc sản phẩm không phổ biến.

+ Có thể khuyến nghị cho những người dùng có sở thích riêng.

* **Khuyết điểm của phương pháp lọc dựa trên nội dung:**

+ Thường gặp các khó khăn liên quan đến phân tích nội dung.

+ Gặp các vấn đề khi khuyến nghị cho người dùng mới (Khởi động lạnh).

+ Không thể đa dạng trong khuyến nghị (bao gồm các đối tượng khuyến nghị ngoài lĩnh vực quan sát).

**1.3.2. Phương pháp lọc cộng tác**

**Tổng quan:** Lọc cộng tác (tên tiếng anh là Collaborative-Filtering) là phương pháp khai thác những khía cạnh liên quan đến thói quen sử dụng sản phẩm của một nhóm người dùng có cùng sở thích trong quá khứ để đưa ra dự đoán các sản phẩm mới phù hợp với người dùng hiện tại (có thể hình dung rằng là lọc cộng tác giả định rằng những người đồng ý trong quá khứ sẽ đồng ý trong tương lai rằng họ sẽ thích các mặt hàng tương tự như các mặt hàng mà họ đã thích trong quá khứ).



1.3.2.1 Ví dụ mô hình kỹ thuật lọc cộng tác

Trong hình minh họa 1.3.2.1 ta có thể thấy phương pháp lọc cộng tác hoạt động như thế nào trên hệ thống Netflix. User A đã xem Movie A và Movie B, User B thì đã xem Movie A, Movie B và Movie C và giữa User A và User B đều có sự tương đồng về sở thích ở một mức độ giống nhau. Suy ra phương pháp lọc cộng tác sẽ cho rằng User A cũng sẽ có thể thích xem Movie C (Movie C là movie mà User B đã xem).

**Bài toán:** Cũng giống như lọc dựa trên nội dung, phương pháp lọc cộng tác cũng xây dựng một ma trận đánh giá gồm danh sách các người dùng *u* và danh sách các sản phẩm *p* nhằm tìm kiếm những giá trị tiên đoán độ phù hợp giữa sản phẩm *p* và người dùng *u.*

A grid of red and black question marks

Description automatically generated

1.3.2.2. Dấu ? là những giá trị cần tiên đoán trong ma trận đánh giá

Ý tưởng chung của phương pháp lọc cộng tác là khai thác thông tin, hành vi quá khứ của người dùng dựa trên các đánh giá sẵn có từ ma trận đánh giá (Hình 1.3.2.2) để tiên đoán, lượng hóa mức độ phù hợp của các đối tượng sản phẩm khuyến nghị mà người dùng chưa biết. Một số nghiên cứu phổ biến đã thực hiện khảo sát, phân loại, cũng như thực nghiệm, đánh giá các thuật toán lọc cộng tác. Các phương pháp lọc cộng tác được phân thành hai nhóm chính: Một là Lọc cộng tác với cách tiếp cận dựa trên bộ nhớ (Memory-Based) như các thuật toán tính toán lân cận, tương tự; Hai là Lọc cộng tác với cách tiếp cận dựa trên mô hình (Model-Based) như các thuật toán gom cụm, phân lớp giám sát, thừa số hóa ma trận (Matrix Factorization).

**Phương pháp lọc cộng tác với cách tiếp cận dựa trên bộ nhớ** có đặc trưng cơ bản là thường sử dụng toàn bộ dữ liệu đã có để dự đoán đánh giá của một người dùng nào đó về sản phẩm mới. Cách tiếp cận dựa trên bộ nhớ thường được chia làm 2 loại: dựa trên người dùng và dựa trên sản phẩm:

a. Dựa trên người dùng

Phương pháp này gồm 2 bước như sau:

- Bước 1: Tìm kiếm những người dùng có đánh giá tương tự với người dùng cần được dự đoán.

- Bước 2: Sử dụng đánh giá từ những người dùng được tìm thấy ở bước 1 để tính toán dự đoán cho người cần được dự đoán.

b. Dựa trên sản phẩm

Phương pháp này gồm 2 bước như sau:

- Bước 1: Xây dựng một ma trận để xác định mối quan hệ giữa các cặp sản phẩm với nhau.

- Bước 2: Kiểm tra thị hiếu của người dùng cần dự đoán bằng cách kiểm tra ma trận và kết hợp dữ liệu của người dùng đó.

**Phương pháp lọc cộng tác với cách tiếp cận dựa trên mô hình** chủ yếu phát triểncác mô hình bằng cách sử dụng các khai phá dữ liệu khác nhau, các thuật toán học máy để dự đoán đánh giá của người dùng về các mặt hàng chưa được đánh giá

* **Ưu điểm của phương pháp lọc cộng tác:**

**+** Có khả năng dự doán sở thích và nhu cầu của người dùng để đưa ra các gợi ý sản phẩm phù hợp với từng khách hàng mà không cần hiểu sản phẩm.

**+** Phù hợp với những hệ thống lớn có nhiều đánh giá từ phía người dùng.

* **Khuyết điểm của phương pháp lọc cộng tác:**

**+** Ma trận đánh giá còn thưa

**+** Giống như lọc dựa trên nội dung, Lọc cộng tác vẫn gặp các vấn đề khi khuyến nghị cho người dùng mới (Khởi động lạnh).

**+** Phương pháp này cũng không thể gợi ý được các sản phẩm mới và các sản phẩm chưa được người dùng đánh giá.

**+** Độ chính xác sẽ kém nếu như sở thích của người dùng thay đổi.

**1.3.3. Phương pháp tiếp cận lai**

**1.4. Một số công trình nghiên cứu khoa học có liên quan**

**1.4.1. Phát triển một số phương pháp khuyến nghị hỗ trợ tìm kiếm thông tin học thuật dựa trên tiếp cận phân tích mạng xã hội**

**1.4.2. Giải quyết vấn đề phân phối trông hệ thống khuyến nghị dựa trên đặc trưng nội dung của đối tượng**

**1.5. Giới thiệu bài toán khuyến nghị phim trên các trang web xem phim**

**1.6. Tóm tắt chương 1**