# **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 thành công hệ thống Khuyến nghị như 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.



Hình 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. Phát biểu 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:

**Định nghĩa 1: Không gian người dùng**

Không gian người dùng là tập tất cả nhữg người dùng mà hệ thống quan sát được, để thực hiện phân tích, khuyến nghị. Ký hiệu là *U, U =* {}.

**Định nghĩa 2: Không gian đối tượng khuyến nghị**

Không gian đối tượng khuyến nghị là tập tất cả những đối tượng sẽ được khuyến nghị cho người dùng. Tùy vào ứng dụng cụ thể, đối tượng khuyển nghị có thể là sản phẩm, dịch vụ hoặc con người … Ký hiệu là *P, P =* {}.

**Định nghĩa 3: Hàm phù hợp**

Hàm phù hợp *F* là ánh xạ *F : U x P R*, dùng để ước lượng độ phù hợp của p P với u U. Với R là một ma trận có thứ tự các số nguyên hoặc thực trong một khoảng nhất định.

**Phát biểu bài toán khuyến nghị:**

**Đầu vào (Input):**

**+** 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.



Hình 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**

**1.3.1.1. Định nghĩa**

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

Hình 1.3.1.1.1 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.

**1.3.1.2. 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.*

**1.3.1.3. Phân loại**

Phương pháp lọc dựa trên nội dung có thể được chia ra làm 2 nhóm chính:

1. 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.
2. 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 như: phân lớp SVM, phân lớp Bayesian và các phương pháp xác xuất như Pazzani và Billsus, Mooney và Roy, Gemmis và đồng nghiệp.
3. **Tiếp cận nội dung dựa trên bộ nhớ**

Tiếp cận nội dung dựa trên bộ nhớ(hay còn gọi là phương pháp dựa trên bộ nhớ) là phương pháp thường được thực hiện với việc ước lượng mức độ phù hợp của đối tượng huyến nghị với người dùng (tức giá trị hàm phù hợp *f*(*u,p*)) dựa trên việc tổng hợp mức độ quan tâm *u* đối với tập k đối tượng có nội dung tương tự với *p*, ký hiệu là = {}, *P*, hoặc tổng hợp mức độ quan tâm từ tập *k* những người dùng có sở thích tương tự *u*, = {}, *U.* Tùy thuộc vào cách biểu diễn nội dung đối tượng dữ liệu và hồ sơ người dùng, chúng ta sẽ có một hàm phù hợp để tính độ tương tự và xác định tập cũng như . Thông thường, các nghiên cứu dùng mô hình không gian vector độ đo Consine để biểu diễn nội dung và tính độ tương tự giữa các đối tượng.

Phương pháp dựa trên bộ nhớ có những ưu điểm và nhược điểm như sau:

* **Ưu điểm:**

+ Đơn giản, dễ thực hiện.

* **Nhược điểm:**

+ Chất lượng khuyến nghị thường tốt hơn do tính toán trên cả tập dữ liệu khi thực hiện khuyến nghị.

1. **Tiếp cận nội dung dựa trên mô hình**

Với phương pháp dựa trên bộ nhớ, hệ thống thường sẽ tính giá trị hàm phù hợp dựa trên các độ do như Cosine, Euclide. Đối với các phương pháp dựa trên mô hình, một mô hình sẽ được huấn luyện từ dữ liệu để phân các đối tượng khuyến nghị thành những đối tượng được người dùng quan tâm hay không quan tâm và quan tâm nhiều hay ít dùng các phương pháp học máy giám sát: phân lớp SVM, phân lớp Bayesian và một số phương pháp xác xuất khác. Nói cách khác, mô hình huấn luyện giúp tiên đoán giá trị hàm phù hợp *f*(*u,p*) của đối tượng khuyến nghị đối với người dùng . Chẳng hạn, phân lớp Bayesian là một phương pháp dựa trên mô hình khá phổ biến, được dùng trong khai thác dữ liệu, phân lớp Bayesian có thể dùng để ước lượng xác xuất đối tượng khuyến nghị *p* phù hợp với *u* như thế nào. Hay nói cách khác, *p* được *u* quan tâm không hay quan tâm nhiều hay ít.

Ví dụ, xác xuất một tài liệu *p* được một người dùng *u* nào đó quan tâm là bao nhiêu? Tức là, giá trị hàm phù hợp *f*(*u,p*) khi đó được tính dựa trên việc ước lượng xác xuất *p* thuộc lớp (u) và (u) (u quan tâm và không quan tâm đến p) là bao nhiêu, khi cho trước một tập các từ khóa mô tả tài liệu p là {}. Giá trị hàm phù hợp *f*(*u,p*) khi đó được tính như sau:

*f*(*u,p*) = P(p (u)) = P(|)

Giả sử các từ khóa mô tả tài liệu là độc lập, khi đó xác xuất P(p (u)) sẽ là:

P(p (u)) = P(|) = P()

Nhìn chung phương pháp dựa trên mô hình có ưu điểm và khuyến điểm như sau:

* **Ưu điểm**

+ Khả năng đáp ứng tốt khi tập dữ liệu được gia tăng.

+ Một mô hình biểu diễn tốt thế giới thực sẽ giúp tránh được vấn đề khớp (overfitting) so với phương pháp dựa trên bộ nhớ.

+ Nhanh hơn so với phương pháp dựa trên bộ nhớ do không phải tính trên cả tập dữ liệu mà chỉ dựa vào mô hình đã xây dựng để khuyến nghị.

* **Khuyết điểm**

**+** Phải xây dựng và cập nhật lại mô hình khi có sự thay đổi. Đây là quá trình gây tốn tài nguyên.

**+** Chất lượng tiên đoán thấp hơn so với các phương pháp dựa trên bộ nhớ vì không được tính toán trên cả tập dữ liệu. Tuy nhiên, nò tùy thuộc vào chất lượng của mô hìnnh được được xây dựng có phản ánh tốt thế giới thực hay không, tức là có đúng với thực tế hay không.

**1.3.1.4. Ưu điểm và khuyết điểm**

* **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**

**1.3.2.1. Định nghĩa**

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).



Hình 1.3.2.1.1 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).

**1.3.2.2. 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* được gọi là ma trận A(*U*,*P*). Ma trận A có kích thước là i x j và chứa các giá trị đánh giá với i 1…N và j 1…M. Những giá trị này thể hiện mức độ phù hợp của đối tượng với người dùng . Giá trị có thể là giá trị nguyên hay thực trong 1 khoảng tùy vào bài toán. Thông thường, giá trị đánh giá mức độ phù hợp trong hầu hết hệ thống ứng dụng phổ biến nhận giá trị từ 1 (không phù hợp) đến 5 (rất phù hợp). Nếu người dùng chưa thể hiện đánh giá với đối tượng thì giá trị = và cần được thu thập hoặc tính toán (Ví dụ minh họa hình 1.3.2.2.1).

A grid of red and black question marks

Description automatically generated

Hình 1.3.2.2.1. 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 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.

**1.3.2.3. Phân loại**

Các phương pháp lọc cộng tác được phân thành hai nhóm chính:

1. Lọc cộng tác với cách tiếp cận dựa trên bộ nhớ (Memory-Based) như các thuật toán tính toán lân cận, tương tự.
2. 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.
3. **Tiếp cận lọc cộng tác dựa trên bộ nhớ**

Các hệ thống lọc cộng tác dựa trên bộ nhớ thường dùng các kỹ thuật thống kê để tìm kiếm những người dùng, hoặc các đối tượng khuyến nghị tương tự nhau dựa trên thông tin đánh giá, hành vi quá khứ của người dùng từ ma trận đánh giá. Tiếp cận lọc cộng tác dựa trên bộ nhớ tìm cách ước lượng giá trị hàm phù hợp *f(u,p)* của đối tượng khuyến nghịp với người dùng *u* dựa trên những đánh giá của những người đồng sở thích của *u* đối với *p* (lọc dựa trên người dùng), hoặc dựa trên những đánh giá của *u* với các đối tượng khuyến nghị *p’* tương tự với *p* (lọc dựa trên đối tượng khuyến nghị). Về cơ bản, thì các thuật toán, kỹ thuật tỉnh toán cho lọc cộng tác dựa trên người dùng và lọc dựa trên đối tượng khuyến nghị từ ma trận đánh giá là tương tự nhau. Có khác chăng là kích thước của không gian người dùng và không gian đối tượng khuyến nghị sẽ ảnh hưởng đến tốc độ tính toán khi xác định nhóm các đối tượng tương tự. 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:

* **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.
* **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.

1. **Tiếp cận lọc cộng tác dựa trên mô hình**

Phương pháp lọc cộng tác với cách tiếp cận dựa trên mô hìnhchủ 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. Theo quan điểm xác xuất, thì các thuật toán lọc cộng tác dựa trên mô hình cần tính toán xác suất mà người dùng *u* đánh giá cho một đối tượng khuyến nghị *p*, P(). Quá trình đó có thể xem như việc tính toán giá trị kỳ vọng cho đánh giá của người dùng *u* với đối tượng khuyến nghị *p*.

Khác với lọc cộng tác dựa trên bộ nhớ, các thuật toán lọc cộng tác dựa trên mô hình dùng tập các đánh giá có sẵn trong ma trận A để học một mô hình đánh giá cho mỗi người dùng. Sau đó, mô hình học được sẽ dùng để tiên đoán các đánh giá khác. Một số thuật toán lọc cộng tác dựa trên mô hình được sử dụng phổ biến như Thuật toán lọc cộng tác gom cụm, Thuật toán lọc cộng tác dựa trên xác xuất Bayes, Thừa số hóa ma trận (Matrix Factorization)…

**1.3.2.4. Ưu điểm và nhược điểm**

* **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**

Phương pháp tiếp cận lai (tên tiếng anh là Hybird Filtering) là phương pháp kết hợp các kỹ thuật khuyến nghị khác nhau. Hầu hết các hương pháp tiếp cận lai đều đưa ra các dự đoán dựa trên nội dung và dựa trên cộng tác một cách riêng biệt và sau đó kết hợp chúng lại với nhau. Bằng cách thêm các tính năng của lọc dựa trên nội dung vào lọc cộng tác (có thể làm ngược lại) ta có thể đưa ra các dữ liệu thực nghiệm từ một số nghiên cứu thực nghiệm đã được chứng minh rằng hiệu suất của phương pháp tiếp cận lai có thể đưa ra các kết quả khuyến nghị chính xác hơn các phương pháp tiếp cận thuần túy.



1.3.3.1. Hình minh họa phương pháp tiếp cận lai

Trong tiếp cận lai ta có một số cách kết hợp các phương pháp như sau:

* Sử dụng cả hai phương pháp lọc dựa trên nội dung và phương pháp lọc cộng tác, sau đó dùng hai kết quả thu được để quyết định:

+ Sử dụng kết quả của phương pháp nào tốt hơn (tùy thuộc vào từng thời điểm).

+ Dùng cả hai kết quả để đánh giá.

* Xây dựng hệ thống lọc dựa trên nội dung sử dụng các đặc trưng của lọc cộng tác.
* Xây dựng hệ thống lọc cộng tác sử dụng các đặc trưng của lọc dựa trên nội dung.
* Xây dựng hệ thống kết hợp cả lọc dựa trên nội dung và lọc cộng tác (Hoạt động chia làm nhiều pha, mỗi pha mỗi phương pháp hoạt động độc lập với nhau).

Ngoài ra, trong thực tế các phương pháp tiếp cận lai thường được sử dụng rất đa dạng, các phương pháp lai cho hệ khuyến nghị gồm có 7 phương pháp tiến cận lai phổ biến: Lai có trọng số (Weighted Hybrid); Lai chuyển đổi (Switching Hybrid); Lai trộn (Mixed Hybrid); Lai kết hợp đặc trưng (Feature Combination Hybrid); Lai theo đợt (Cascade Hybrid); Lai tăng cường đặc trưng (Feature Augmentation Hybrid); Lai meta (Meta-Level Hybrid). Phần tiếp theo trong đồ án chuyên ngành sẽ trình bày sơ lược về các phương pháp lai đã được mêu tả trên.

1. **Lai có trọng số (Weighted Hybird)**

Mỗi phương pháp khuyến nghị phải đi tìm và xác định giá trị hàm phù hợp *f(u,p)* của một đối tượng *p* *P* với người dùng *u* *U*. Tiếp cận lai có trọng số sẽ tính toán giá trị của hàm phù hợp dựa trên kết quả của tất cả của các phương pháp khuyến nghị khác tồn tại trong hệ thống. Thông thường thì hình thức lai có trọng số đơn giản nhất là kết hợp tuyến tính các giá trị phù hợp tính được từ các phương pháp khác nhua trong hệ thống.

Nhìn chung ưu điểm và nhược điểm của lai có trọng số là như sau:

* **Ưu điểm:** Tất cả khả năng, phương pháp khác nhau của hệ thống được tham gia vào quá trình khuyến nghị một cách minh bạch, tự nhiên và dễ dàng thực hiện, dễ dàng điều chỉnh.
* **Nhược điểm:** Việc ước lượng trọng số lớn hay nhỏ cho phù hợp với những phương pháp khác nhau.

1. **Lai chuyển đổi (Switching Hybird)**

Các hệ thống khuyến nghị thuộc nhóm lai chuyển đổi thường sử dụng một số điều kiện để chuyển đổi qua lại giữa các phương pháp khuyến nghị khác nhau. Ví dụ: một số nghiên cứu sinh đã thực hiện nghiên cứu một hệ thống khuyến nghị sử dụng phương pháp lai chuyển đổi giữa tiếp cận nội dung và lọc cộng tác. Các tác giả đã áp dụng phương pháp lọc nội dung trước và sau đó những trường hợp là tiếp cận nội dung không thực hiện được (không thể tiếp tục thực hiện khuyến nghị và đưa ra giá trị phù hợp thấp) thì tiếp cận lọc cộng tác sẽ được áp dụng vào thay.

Lọc cộng tác trong phương pháp lai chuyển đổi sẽ giúp hệ thống có thể khuyến nghị được các đối tượng có nội dung, ngữ nghĩa khác với các đối tượng đã được đánh giá cao. Nói cách khác, một đối tượng có thể không được khuyến nghị với cách tiếp cận nội dung nhưng sau khi áp dụng lọc cộng tác thì đối tượng đó có thể được ưu tiên khuyến nghị.

Nhìn chung ưu điểm và nhược điểm của lai chuyển đổi là như sau:

* **Ưu điểm:** Phương pháp tiếp cận này rất “nhạy” với các điểm mạnh và điểm yếu của các phương pháp khác nhau.
* **Nhược điểm:** Tuy “nhạy” với điểm mạnh và điểm yếu của các phương pháp khác nhau, nhưng lai chuyển đổi yêu cầu cần phải xác định điều kiện để chuyển đổi giữa các phương pháp. Điều đó làm quá trình chuyển đổi trở nên phức tạp hơn.

1. **Lai trộn (Mixed Hybird)**

Tiếp cận lai trộn là phương pháp thực hiện các phương pháp khuyến nghị khác nhau một cách độc lập và kết hợp các kết quả từ phương pháp sẽ được chuyển thành danh sách đề xuất và được chuyển đến cho người dùng. Tiếp cận lai trộn có thể tránh được vấn đề người dùng mới (Khởi động lạnh – Cold Start). Giống như trường hợp trên, lọc dựa trên nội dung trong tiếp cận lai trộn cũng giúp đề xuất các đối tượng khuyến nghị mới (là các đối tượng vừa được khởi tạo, chưa có một đánh giá từ phía người dùng) trong danh sách sau cùng dựa trên mô tả nội dung của đối tượng này, trong khi đó phương pháp lọc cộng tác thông thường không thể làm được. Bù lại, lọc cộng tác trong lai trộn chỉ giúp đề xuất các đối tượgn khuyến nghị tiềm năng nhưng lại không tương tự về nội dung.

Nhìn chung ưu điểm và nhược điểm của lai trộn là như sau:

* **Ưu điểm:** Có thể giúp đề xuất các đối tượng tiềm năng mà bản thân một phương pháp riêng biệt không thể xác định được. Trộn lọc nội dung và lọc cộng tác có thể giúp giải quyết được vấn đề khởi động lạnh (Cold Start) và cũng có thể đa dạng hóa khuyến nghị.
* **Nhược điểm:** Vì tiếp cận này sử dụng nhiều đề xuất từ các phương pháp khác nhau. Do đó hệ thống cần được xử lý, lọc các đề xuất đụng độ, trùng lắp từ các phương pháp khác nhau.

1. **Lai kết hợp đặc trưng (Feature Combination Hybird)**

Lai kết hợp đặc trưng là tiếp cận phát triển phương pháp khuyến nghị bằng cách sử dụng kết hợp thông tin đánh giá của người dùng với nội dung của đối tượng khuyến nghị.

Ưu điểm và nhược điểm của lai kết hợp đặc trưng là như sau:

* **Ưu điểm:** Lai kết hợp đặc trưng cho phép hệ thống xem xét dữ liệu cộng tác, nhưng không chỉ phụ thuộc duy nhất vào dữ liệu cộng tác trong ma trận đánh giá. Ngược lại, hệ thống cũng có được thông tin về sự tương tự vốn có giữa các đối tượng khuyến nghị (dựa trên đặc trưng nội dung) mà không bị ảnh hưởng bởi dữ liệu cộng tác.
* **Nhược điểm:** Khó khăn trong việc xác định các đặc trưng cộng tác và đặc trưng nội dung phù hợp.

1. **Lai theo đợt (Cascade Hybrid)**

Lai theo đợt là tiếp cận mà các phương pháp khuyến nghị khác nhau được lần lượt áp dụng theo một thứ tự ưu tiên được xác định trước tùy vào mỗi ứng dụng cụ thể.Ví dụ, phương pháp khuyến nghị thứ nhất sinh ra một danh sách xếp hạng các Ứng viên (danh sách thô). Tiếp đó, những phương pháp khác với độ ưu tiên thấp hơn sẽ được áp dụng để lọc lại danh sách thô này. Lai theo đợt giúp phương pháp thứ hai tránh những đối tượng có thể không bao giờ cần khuyến nghị vì những đối tượng này đã được lọc qua phương pháp thứ nhất. Đồng thời, các đối tượng được ưu tiên chọn với phương pháp thứ nhất sẽ được tinh lọc, chứ không bị loại bỏ thông qua phương pháp thứ hai.

Ưu điểm và nhược điểm của lai theo đợt là như sau:

* **Ưu điểm:** So với tiếp cận lại có trọng số (Weighted Hybrid) và một số tiếp cận lai khác thì việc lọc lại danh sách thô làm cho tiếp cận này hiệu quả hơn bởi vì các phương pháp tiếp theo chỉ thực hiện lọc trên một không gian nhỏ hơn, thay vì trên cả không gian tất cả các đối tượng khuyến nghị.
* **Nhược điểm:** Khó khăn trong việc xác định độ ưu tiên giữa các phương pháp khác nhau cho mỗi ứng dụng cụ thể.

1. **Lai tăng cường đặc trưng (Feature Augmentation Hybird)**

Với tiếp cận lai tăng cường đặc trưng, phương pháp đầu sẽ học một mô hình để sinh ra đặc trưng tăng cường cho đầu vào của phương pháp tiếp theo (Nhìn chung nó khá giống lai theo đợt). Nhìn chung, lai theo đợt và lai tăng cường đặc trưng đều là những tiếp cận mà hai phương pháp khác nhau sẽ được thực hiện một cách trình tự, tức là kết quả từ phương pháp thứ nhất sẽ ảnh hưởng đến phương pháp thứ hai. Tuy nhiên, về cơ bản thì hai tiếp cận lai này là hoàn toàn khác nhau. Với lai tăng cường đặc trưng thì những đặc trưng được dùng trong phương pháp thứ hai bao gồm những đặc trưng sinh ra bởi phương pháp thứ nhất, còn đối với lai theo đợt thì phương pháp thứ hai được dùng với độ ưu tiên thấp hơn phương pháp thứ nhất, nhằm lọc lại danh sách ứng viên mà phương pháp thứ nhất đã sinh ra.

Ưu điểm và nhược điểm của lai tăng cường đặc trưng là như sau:

* **Ưu điểm**: Việc tăng cường đặc trưng dùng các phương pháp khác giúp hệ thống có thể cải tiến độ chính xác khuyến nghị mà không thay đổi, ảnh hưởng đến phương pháp khuyến nghị chính.
* **Nhược điểm:** Khó khăn trong việc xác định đặc trưng tăng cường phù hợp.

1. **Lai meta (Meta-Level Hybird)**

Lai meta dùng mô hình được tạo ra bởi phương pháp trước làm đầu vào cho phương pháp sau. Với lại tăng cường đặc trưng (Feature Augmentation Hybird) thì phương pháp đầu sẽ học một mô hình để sinh ra đặc trưng làm đầu vào cho phương pháp tiếp theo. Trong khi lai meta thì cả mô hình của phương pháp thứ nhất sẽ làm đầu vào cho phương pháp thứ hai. Lai meta giữa lọc nội dung và lọc cộng tác phần nào giải quyết được vẫn đề ma trận thưa trong tiếp cận lọc cộng tác bởi vì lai meta sẽ tìm kiếm những người dùng tương tự dựa trên các đặc trưng nội dung trước khi áp dụng phương pháp lọc cộng tác. Đối với những người dùng có quá ít đánh giá thì việc xác định nhóm những người đồng sở thích thông qua lọc cộng khác sẽ không được chính xác.

Ưu điểm và nhược điểm của lai meta là như sau:

* **Ưu điểm**: Với lai meta giữa lọc nội dung và cộng tác, phương pháp lọc cộng tác sẽ dễ dàng thực hiện tính toán trên “Dữ liệu dày” hơn so với dữ liệu thô trong ma trận đánh giá.
* **Nhược điểm:** Khó khăn trong việc chọn phương pháp để thực hiện trước. Mỗi phương pháp được chọn vẫn phải gặp những hạn chế vốn có của nó.

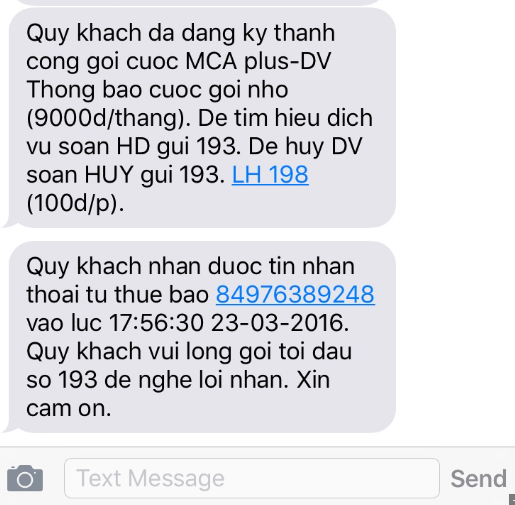
Tóm lại, mỗi phương pháp tiếp cận lai đều có những ưu và nhược điểm vốn có của nó. Tiếp cận lai sẽ giúp giảm bớt phần nào hạn chế của các phương pháp khác nhau và đồng thời cũng được sử dụng cho nhiều trường hợp/mục đích khác nhau.

**1.4. Khảo sát một số công trình nghiên cứu khoa học có liên quan**

**1.4.1. Nghiên cứu và xây dựng hệ thống khuyến nghị cho bài toán dịch vụ giá trị gia tăng trong ngành viễn thông cv**

Đề tài “Nghiên cứu và xây dựng hệ thống khuyến nghị cho bài toán dịch vụ giá trị gia tăng trong ngành” là một công trình luận văn thạc sĩ của tác giả Kiều Xuân Chấn, được viết vào năm 2017 tại trường Đại học Công nghệ thuộc Đại học Quốc gia Hà Nội. Là một công trình nghiên cứu, xây dựng hệ thống khuyến nghị cho bài toán dịch vụ giá trị gia tăng (tên tiếng anh là VAS – Valued Added Service) ứng dụng 3 phương pháp khuyến nghị phổ biến là lọc dựa trên nội dung, lọc cộng tác và cả kết hợp cả hai phương pháp trên, cũng như là sử dụng các kỹ thuật/thuận toán nhằm giải quyết bài toán và xây dựng một hệ thống khuyến nghị dành cho bài toán giá trị dịch vụ gia tăng trong ngành điện tử viễn thông.

Theo đề tài đã mô tả “*Các dịch vụ giá trị gia tăng là một thuật ngữ được sử dụng để chỉ các phụ trợ cho một dịch vụ cơ bản. Dịch vụ giá trị gia tăng thường được giới thiệu đến khách hàng sau khi khách hàng đã mua các dịch vụ cơ bản. Dịch vụ cơ bản đóng vai trò trung tâm và các dịch vụ giá trị gia tăng thường là những dịch vụ phụ thuộc vào nó*” thì dịch vụ giá trị gia tăng (VAS) là một trong những dịch vụ được sinh ra nhằm xây dựng một mối quan hệ mạnh mẽ giữa khách hàng và các nhà cung cấp và các nhà cung cấp cũng xem các dịch vụ này như là một nguồn thu nhập bổ sung cho ngân sách của họ. Một trong những ví dụ dễ hình dung đó là dịch vụ thông báo cuộc gọi nhỡ MCA (Miss Call Alert) của nhà mạng Viettel và Mobifone, đây là một hệ thống cho phép thuê bao di động nhận được tin nhắn SMS thông báo thông tin về các cuộc gọi nhỡ tới số thuê bao của mình khi điện thoại di động của họ đang tắt máy, hết pin hoặc nằm ngoài vùng phủ sóng.



1.4.1.1. Hình minh họa cho dịch vụ MCA của nhà mạng Viettel

**1.4.2. 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 gia tăng trong ngành viễn thông**

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

**1.6. Tóm tắt chương 1**