ỦY BAN NHÂN DÂN THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC SÀI GÒN

TRẦN NGUYÊN LỘC

KHẢO SÁT VỀ HỆ THỐNG KHUYẾN NGHỊ TRÊN MỘT SỐ TRANG WEB XEM PHIM

ĐỀ CƯƠNG ĐỒ ÁN CHUYÊN NGÀNH
NGÀNH: CÔNG NGHỆ THÔNG TIN

Thành phố Hồ Chí Minh, tháng 12 năm 2023

ỦY BAN NHÂN DÂN THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC SÀI GÒN

TRẦN NGUYÊN LỘC

KHẢO SÁT VỀ HỆ THỐNG KHUYẾN NGHỊ TRÊN MỘT SỐ TRANG WEB XEM PHIM

ĐỀ CƯƠNG ĐỒ ÁN CHUYÊN NGÀNH
NGÀNH: CÔNG NGHỆ THÔNG TIN

Giảng viên phụ trách
TS. PHAN TẤN QUỐC

Thành phố Hồ Chí Minh, tháng 12 năm 2023

LỜI CẢM ƠN

Trước hết em xin gửi đến lời cảm ơn chân thành và sâu sắc nhất đến thầy TS. Phan Tấn Quốc, người trực tiếp hướng dẫn và tận tình chỉ bảo cho em cho tới khi em hoàn thành đồ án của mình.

Tiếp đến em dành lời cảm ơn đến quý thầy cô khoa Công nghệ thông tin – trường Đại học Sài Gòn đã truyền đạt cho em những kiến thức vô cùng quý báu và bổ ích trong suốt quá trình nghiên cứu và học tập tại trường.

Xin chân thành cảm ơn tới những người bạn đã luôn sát cánh cùng em, những lời động viên, những lần hỗ trợ những lúc cần thiết đã phần nào giúp em hoàn thành đồ án này.

Cuối cùng, em xin cảm ơn đến ba mẹ và người thân trong gia đình đã hỗ trợ và tạo điều kiện thuận lợi cho em trong suốt thời gian học tập và nghiên cứu tại Đại học Sài Gòn.

Mục lục

DANH MỤC HÌNH ẢNH

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.

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 khảo sát và 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) trên các tập dữ liệu thử nghiệm MovieLens, hệ thống đề xuất phim MOVREC.

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

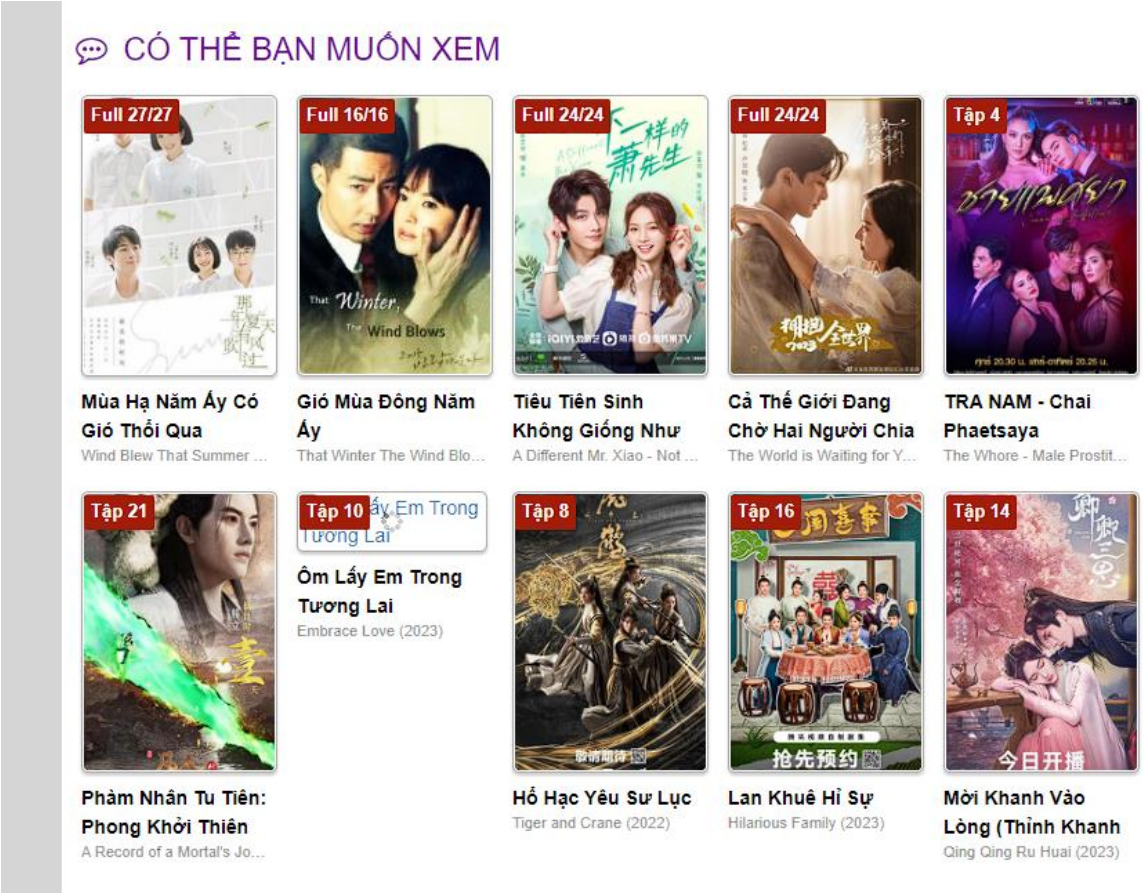
Chương 3 trong đồ án chuyên ngành sẽ tập trung vào việc khảo sát cách thức của một hệ thống khuyến nghị hoạt động trên các trang web xem phim. Các trang web được khảo sát trong đồ án chuyên ngành gồm Netflix, BiliBili và FPT Play. Sau đó ta so sánh cách thức xử lý bài toán khuyến nghị phim giữa 3 trang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

CHƯƠNG 1. TỔNG QUAN VỀ HỆ THỐNG KHUYẾN NGHỊ

1.1. 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 [1].

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 hoặc dịch vụ có liên quan trên các trang web, ứng dụng phổ biến.



Hình 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 đoán 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 về 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 [9]

Không gian người dùng là tập tất cả những người dùng mà hệ thống quan sát được, để thực hiện phân tích, khuyến nghị. Ký hiệu là *U*, $U = \{u_1, u_2, u_3, \dots\ u_i\}$.

Định nghĩa 2: Không gian đối tượng khuyến nghị [9]

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*, $P = \{p_1, p_2, p_3, \dots\ p_j\}$.

Định nghĩa 3: Hàm phù hợp [6]

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 [1]:

Đầu vào (Input):

+ Tập người dùng *U*, mỗi người dùng u_i thuộc *U* và có các đặc điểm I = $\{i_1, i_2, i_3, \dots\ i_k\}$.

+ Một tập sản phẩm, dịch vụ (ở đây ta gọi chung là sản phẩm) P, mỗi sản phẩm p_i có các đặc điểm đặc trưng J = $\{j_1, j_2, j_3, \dots\ j_k\}$.

+ Một ma trận đánh giá tổng quát R = (r_{ij}) 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 r_{ij} là đánh giá của người dùng u_i cho sản phẩm p_i, N và M là lần lượt số người dùng và số sản phẩm.

Người dùng \ Sản phẩm	1	2	…	*i*	…	*M*
1	5	3	0	1	2	0
2	0	2	0	0	0	4
:	0	0	5	0	0	0
u	3	4	0	2	1	0
:	0	0	0	0	4	0
N	0	0	3	2	0	0
a	3	5	0	?	1	0

Hình 1. 2. Ví dụ ma trận đánh giá tổng quát Rij

Đầu ra (Output):

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

Để giải bài toán này chúng ta cần xây dựng hàm F(u_i, p_i) để đo độ phù hợp của sản phẩm p_i với người dùng u_i, 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 [1]:

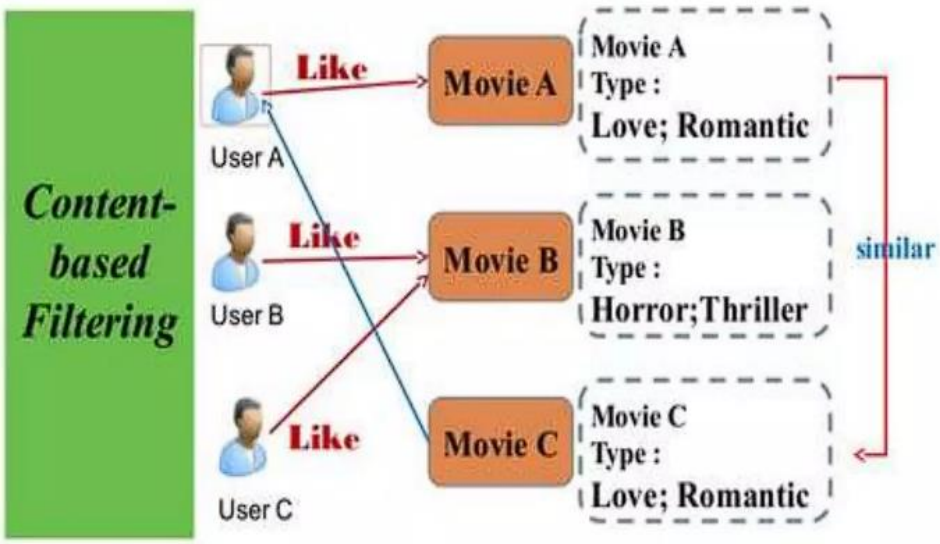
- Đặc điểm của người dùng u_i (lọc theo nội dung người dùng): đặc điểm này được đánh giá chủ quan bởi các quy luật tự nhiên hoặc các quy tắc cơ bản. Ví dụ u_i là nữ thì sẽ có xu hướng chọn mua các sản phẩm của nữ hơn là các sản phẩm của nam [1].
- Đặc điểm của sản phẩm p_i (lọc theo nội dung sản phẩm): 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 … [1]
- Lịch sử giao dịch của người dùng u_i: từ lịch sử giao dịch cũng có thể suy ra sản phẩm mà người dùng u_i 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 [1].
- Những người dùng u_j khác có cùng đặc điểm giống với người dùng u_i: với quan niệm rằng những người dùng giống nhau sẽ thích, đánh giá những sản phẩm giống nhau. Các đặc điểm u_j 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 u_j đã từng quan tâm sẽ đưa ra được những gợi ý phù hợp cho người dùng u_i [1].

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 [1].



Hình 1. 3. Ví dụ mô hình kỹ thuật lọc dựa trên nội dung

Trong hình 1.3 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. Khái quát 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 [2]:

- **Bước 1:** Biểu diện nội dung đối tượng khuyến nghị $p \in P$, ký hiệu *Content(p)*.

- **Bước 2:** Mô hình hóa sở thích người dùng $u \in U$, 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 các cách tiếp cận lọc dựa trên bộ nhớ

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 [7].
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 [10], phân lớp Bayesian [12] và các phương pháp xác xuất như Pazzani và Billsus [13], Mooney và Roy [11], Gemmis và đồng nghiệp [8].

a) 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 $p \in P$ với người dùng $u \in U$ (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 $P_k = \{p_k\}$, $P_k \subseteq P$, hoặc tổng hợp mức độ quan tâm từ tập *k* những người dùng có sở thích tương tự *u*, $U_k = \{u_k\}$, $U_k \subseteq U$. 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 P_k cũng như U_k. 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 [2].

Phương pháp dựa trên bộ nhớ có những ưu điểm và nhược điểm như sau [2]:

- **Ưu điểm:**
  - Đơn giản, dễ thực hiện.
  - Chất lượng khuyến nghị thường tốt hơn do tính toán trên cả tập dữ liệu khi thực hiện khuyến nghị.
- **Nhược điểm:**
  - Tốn bộ nhớ và tốc độ xử lý chậm do phải tính toán, trên cả tập dữ liệu thực khi thực hiện khuyến nghị.
  - Không thể tổng quát hóa tập dữ liệu.

b) 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 [10], phân lớp Bayesian [12] 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 $p \in P$ đối với người dùng $u \in U$ [2]. 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 [12].

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 C_1(u) và C_0(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 $\{k_{1,p}, k_{2,p}, k_{3,p}, \dots\ k_{n,p}\}$. Giá trị hàm phù hợp *f*(*u*,*p*) khi đó được tính như sau [2]:

$$f(u,p) = \mathrm{P}(\mathrm{p} \in C_1(\mathrm{u})) = \mathrm{P}(C_1(u)|k_{1,p}, k_{2,p}, k_{3,p}, \dots\ k_{n,p})$$

Giả sử các từ khóa mô tả tài liệu là độc lập, khi đó xác xuất P(p ∈ C_1(u)) sẽ là [2]:

$$P(p \in C_1(u)) = P(C_1(u)|k_{1,p}, k_{2,p}, k_{3,p}, \dots\ k_{n,p}) = P(C_1(u))\prod_{i=1}^{n} P(k_{i,p}|C_1(u))$$

Nhìn chung phương pháp dựa trên mô hình có ưu điểm và khuyến điểm như sau [2]:

- **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.

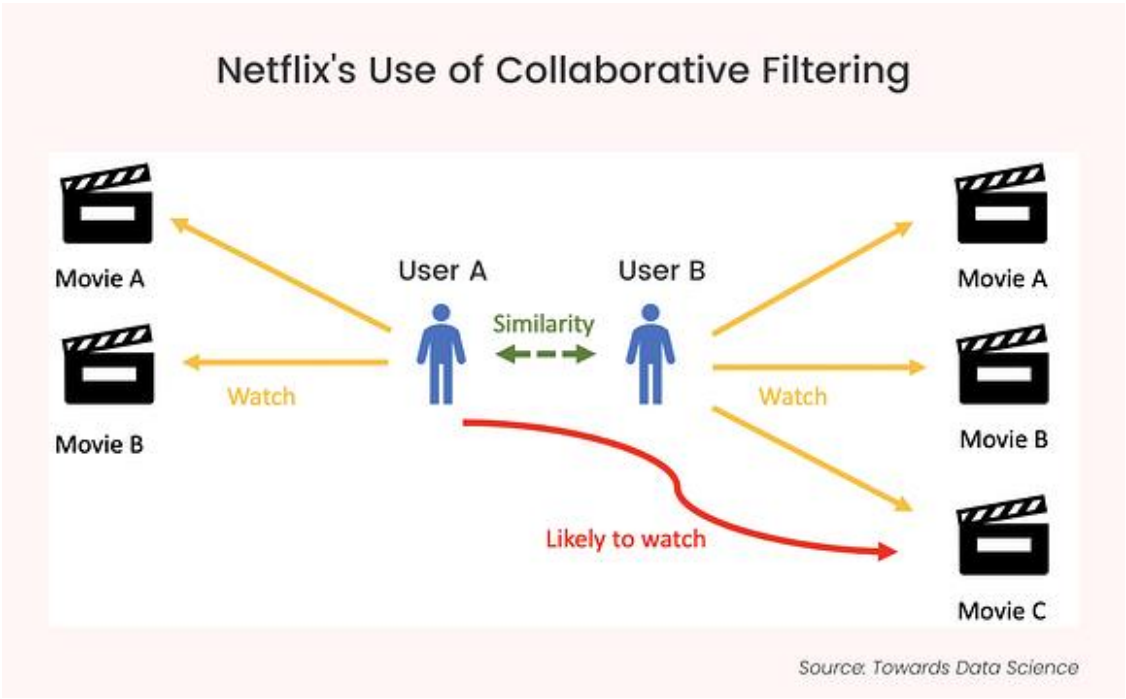
1.3.1.4. Ưu điểm và khuyết điểm của lọc dựa trên nội dung

- **Ưu điểm:**
  - Là phương pháp trực quan, dễ dàng hiểu và giải thích được [3].
  - Không bị ảnh hưởng bởi khởi đầu lạnh (cold start) [3].
  - 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:**
  - 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) [4].



Hình 1. 4. 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. Khái quát 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 $U = \{u_1, u_2, u_3, \ldots u_i\}$ và danh sách các sản phẩm $P = \{p_1, p_2, p_3, \ldots p_j\}$ 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 a_{ij} với i ∈ 1…N và j ∈ 1…M. Những giá trị a_{ij} này thể hiện mức độ phù hợp của đối tượng p_j với người dùng u_i. Giá trị a_{ij} 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 a_{ij} 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 u_i chưa thể hiện đánh giá với đối tượng p_j thì giá trị $a_{ij} = \emptyset$ và cần được thu thập hoặc tính toán [2] (Ví dụ minh họa hình 1.5).

	p_1	p_2	p_3	p_4	p_5	…	p_m
u_1	1	?	5	?	4	?	?
u_2	?	?	4	?	5	?	?
u_3	?	4	?	5	?	?	?
u_4	?	?	?	4	?	?	?
u_5	?	?	?	5	?	?	?
…	?	?	?	?	?	?	?
u_n	?	3	?	?	?	?	5

Hình 1. 5. 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 [2].

1.3.2.3. Phân loại các cách tiếp cận lọc cộng tác

Các phương pháp lọc cộng tác được phân thành hai nhóm chính [2]:

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.

a) 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 [2]. 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.

b) Tiếp cận lọc cộng tác dựa trên mô hình

Phương pháp lọc cộng tác với cách tiếp cận dựa trên mô hình chủ yếu phát triển các 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 [2]. 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 $a_{u,p}$ cho một đối tượng khuyến nghị p, P($a_{u,p}|u,p$). 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 [14].

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 [2]. 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 [14], Thừa số hóa ma trận (Matrix Factorization)…

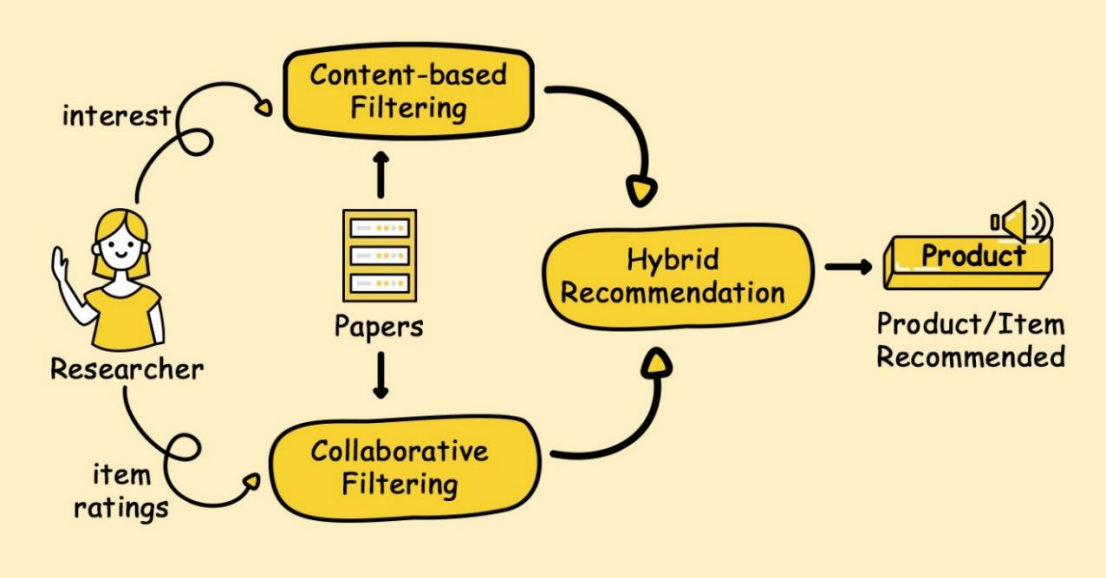
1.3.2.4. Ưu điểm và nhược điểm của lọc cộng tác

- **Ưu điểm:**
  - 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:**
  - 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 [4]. 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].



Hình 1. 6. Ví dụ minh họa cho phương pháp tiếp cận lai

Trong tiếp cận lai ta có một số cách kết hợp các phương pháp như sau [1]:

- 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 [15]: 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.

a) 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 $p \in P$ với người dùng $u \in U$. Tiếp cận lai có trọng số sẽ tính toán giá trị của hàm phù hợp $f_{hybird}(u,p)$ dựa trên kết quả của tất cả $f(u,p)$ 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 nhau trong hệ thống [2].

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.

b) 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 [2].

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 [2].

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.

c) 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 [2].

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.

d) 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 [2].

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.

e) 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 [2].

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.

f) 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 [2].

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.

g) 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 [2].

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

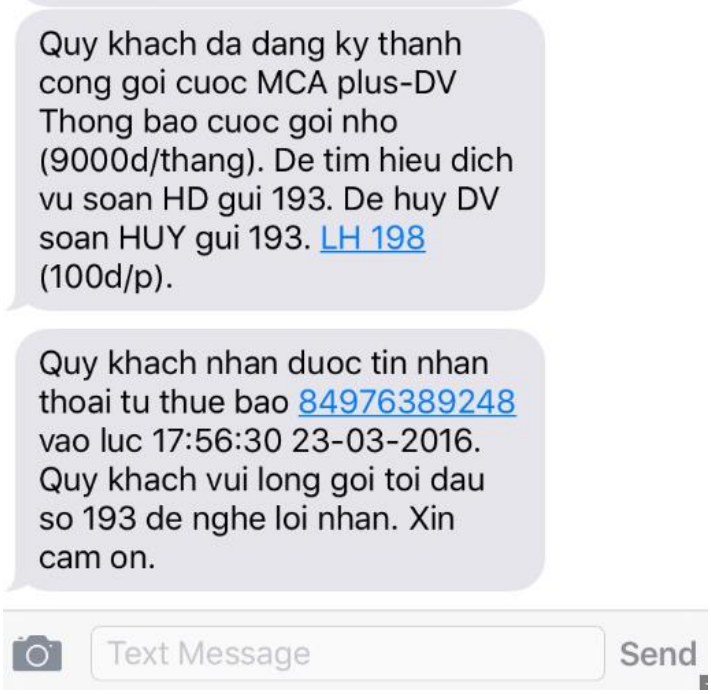
1.4.1. Nghiên cứu và xây dựng hệ thống khuyến nghị cho bài toán dịch vụ giá trị gia tăng trong ngành viễn thông

1.4.1.1. Tổng quan về công trình nghiên cứu

Đề tài "Nghiên cứu và xây dựng hệ thống khuyến nghị cho bài toán dịch vụ giá trị gia tăng trong ngành viễn thông" là một công trình luận văn thạc sĩ của tác giả Kiều Xuân Chấn, được viết và trình bày 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, nghiên cứu 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.

Vậy trước khi bắt đầu nội dung của công trình, ta cần hiểu thế nào là dịch vụ giá trị gia tăng VAS. 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*" [1] 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.



Hình 1. 7. Minh họa cho dịch vụ MCA của nhà mạng Viettel

Công trình nghiên cứu của tác giả Kiều Xuân Chấn đã chỉ ra rằng dịch vụ giá trị gia tăng VAS trong ngành viễn thông là một trong những dịch vụ tiện ích mang lại lợi ích lớn và mối quan hệ dài lâu của khách hàng và các nhà cung cấp, ví dụ: Tiện ích của các nhà mạng, các dịch vụ trên nền data ... Cụ thể hơn trong công trình đã chỉ ra rằng người dùng (user) trong bài toán khuyến nghị VAS chính là các thuê bao di động và thông tin (profile) của người dùng ảnh hưởng tới việc sử dụng dịch vụ đặc trưng bởi các thông tin như: Loại thuê bao, thông tin nhân thân, gói cước thuê bao, tiêu dùng hằng tháng của thuê bao, thông tin địa điểm dịch vụ, lịch sử giao dịch …

	ISDN	PRODUCT_CODE	BIRTH_DATE	GENDER	PROVINCE	STA_DATETIME
1	983655	TOM50	26-FEB-1993 00.00.00	F	K060	26-AUG-2012 11.34.34
2	9899	TOM50	25-JUL-1987 00.00.00	F	T008	17-NOV-2004 19.12.24
3	96974	ECOM1	16-OCT-1988 00.00.00	M	N030	27-AUG-2012 00.32.45
4	983367	ECO096_25K	01-JUN-1988 00.00.00	M	D061	28-AUG-2012 18.49.00
5	9899	TOMA1	30-DEC-1973 00.00.00	M	H004	03-JAN-2005 12.05.22
6	989	TOMA1	01-DEC-1960 00.00.00	M	T008	03-NOV-2004 00.00.00
7	981	ECOM1	01-OCT-1967 00.00.00	F	N038	11-JAN-2005 14.32.20
8	9899	ECOM1	30-APR-1984 00.00.00	F	B056	15-DEC-2004 17.33.30
9	16445	TOSV1	27-AUG-1991 00.00.00	F	N351	31-MAR-2011 23.15.06
10	975	TOMA1	11-AUG-1955 00.00.00	F	T008	20-NOV-2007 17.37.05
11	989	TOM50	16-SEP-1992 00.00.00	M	H031	12-SEP-2005 08.54.51

Hình 1. 8. Hình ảnh được lấy từ công trình nghiên cứu minh họa cho thông tin người dùng viễn thông

1.4.1.2. Đề xuất thuật toán

Trong công trình của mình, tác giả Kiều Xuân Chấn đã đề cập đến một số kỹ thuật thường dùng trong hệ thống khuyến nghị cùng với lời giải thích như sau:

1) **Lọc cộng tác dựa trên bộ nhớ**
    a. **Phương pháp tự tính độ đo tương tự** là phương pháp tìm ra độ đo tương tự giữa người dùng và sản phẩm, công trình cũng đã chỉ ra một số phương pháp tính độ do tương tự như Khoảng cách manhattan, hhoảng cách Euclidean, hệ số tương quan Person, hệ số tương tự Cosine
    b. **Phương pháp K-Láng giềng gần nhất (KNN)** là phương pháp cổ điển phổ biến, đơn giản được sử dụng trong phương pháp lọc cộng tác dựa trên bộ nhớ và phương pháp này cũng chia ra thành 2 phương pháp cơ bản là KNN dựa trên người dùng và KNN dựa trên sản phẩm.

2) **Lọc cộng tác dựa trên mô hình:** Sử dụng dữ liệu đã đánh giá của người dùng để huấn luyện và xây dựng một mô hình đánh giá, từ đó tính toán ước lượng đánh giá của người dùng cho các sản phẩm chưa được đánh giá.
3) **Mô hình nhân tố ẩn**
    a. **Phương pháp thừa số hóa ma trận (Matrix Factorization – MF)** là một trong những phương pháp thành công nhất của mô hình nhân tố ẩn (theo tác giả thì mô hình nhân tố ẩn là mô hình biến đổi người dùng vào các mục không gian đặc trưng tiềm ẩn, có thể hiểu rằng nó xác định các yếu tố ẩn của người dùng và cả yếu tố ẩn của sản phẩm). Nó cho phép kết hợp các thông tin đã có với các thông tin bổ sung và khi thông tin phản hồi rõ ràng không có sẵn, hệ thống tư vấn có thể suy ra sở thích của người dùng bằng cách sử dụng thông tin phản hồi ngầm hoặc gián tiếp phản ánh ý kiến bằng cách quan sát hành vi người dùng.
    b. Phương pháp sử dụng các đặc trưng ưu tiên (Biased Matrix Factorization) là một biến thể liên quan đến phương pháp thừa số hóa ma trận. Công trình đã chỉ ra rằng "*nhiều biến thể được quan sát thấy trong các giá trị xếp hạng là do các hiệu ứng liên quan đến người dùng và sản phẩm, được gọi là đặc trưng ưu tiên và các đặc trưng này không phụ thuộc vào bất kì sự tương tác nào. Ví dụ: trong một hệ thống lớn, một số người dùng có xếp hạng cao hơn những người khác và đối với sản phẩm có xu hướng được xếp hạng cao hơn sản phẩm khác. Do đó một số sản phẩm có thể được xem là tốt hơn hoặc tệ hơn*". Chính vì thế ta có thể hiểu rằng thành phần đặc trưng ưu tiên này là một phần không thể thiếu đối với đặc trưng của người dùng và đặc trưng của sản phẩm để mô hình hóa.
4) Tiêu chuẩn đánh giá
    a. Mean absolute error (MAE) là một phương pháp đơn giản để đo chất lượng khuyến nghị bằng cách đo lường sai số tuyệt đối trung bình (MAE), đôi khi còn được gọi là độ lệch tuyệt đối. Phương pháp này chỉ đơn giản mang ý nghĩa của sự khác biệt tuyệt đối nằm giữa dự đoán và

xếp hạng cho tất cả các xếp hạng được giữ lại của người dùng trong tập kiểm tra.

b. **Root mean square error (RMSE)** là biện pháp liên quan có ảnh hưởng của việc nhấn mạnh nhiều hơn vào các lỗi lớn. Nó được tính như MAE nhưng bình phương lỗi trước khi cộng tổng lại.

c. **Normalized Mean absolute error (NMAE)** là phương pháp được sinh ra để giải quyết lỗi thiếu hụt của MAE và tác giả cho rằng "*MAE có cùng tỷ lệ đánh giá ban đầu, ví dụ đánh giá ở thang 5 sao được biểu diễn bằng số nguyên trong đoạn [1,5], một MAE là 0,7 có nghĩa là thuật toán trung bình bị giảm 0,7 sao. Điều này hữu ích cho việc hiểu kết quả trong một ngữ cảnh cụ thể, nhưng làm sao cho việc so sánh các kết quả trên các bộ dữ liệu rất khó khăn vì chúng có các phạm vi đánh giá khác nhau (sai số 0,7 sẽ có ý nghĩa hơn khi xếp hạng ở [1,5] hơn khi chúng ở [-10,10])"* do đó việc sai số như trên sẽ làm ảnh hưởng đến xếp hạng cho các tập dữ liệu và làm ảnh hưởng ít nhiều đến vấn đề xử lý dữ liệu. Do đó NMAE đôi khi được các nhà phát triển sử dụng để giải quyết trường hợp này, phương thức này chuẩn hóa lỗi bằng cách phân chia phạm vi xếp hạng đơn giản nhất có thể (kết quả là một giá trị trong khoảng [0,1] cho hầu hết thang đánh giá hiện tại).

Trong đề tài của mình, tác giả Kiều Xuân Chấn đã đề xuất một số giải thuật để giải quyết bài toán dịch vụ giá trị gia tăng trong ngành viễn thông, tác giả đã tập trung phân tích ,thực nghiệm và đánh giá đối với **giải pháp KNN** và **giải pháp thừa số hóa ma trận (MF)** trên tập dữ liệu mô phỏng thuê bao di động đăng kí dịch vụ VAS. Đồng thời tác giả cũng cho rằng "*KNN là phương pháp đơn giản và chạy nhanh, nó tỏ ra hiệu quả khi tập dữ liệu lớn và có nhiều thông tin. Phương pháp thừa số hóa ma trận thì có độ chính xác cao và phù hợp với các tập dữ liệu thưa*".

1.4.1.3. Kết quả thực nghiệm của công trình

Kết quả thực nghiệm của công trình đã chỉ ra rằng "*Phương pháp KNN cho sai số RMSE rất lớn, điều đó cho thấy dữ liệu tiêu dùng thuê bao (thoại, sms, vas,*

data) không phải là yếu tố có giá trị đối với việc thuê bao đó đăng ký sử dụng dịch vụ VAS hay không". Và trong kết quả thực nghiệm của công trình đã cho rằng **Phương pháp thừa số hóa ma trận (MF) cho kết quả tốt hơn nhiều so với phương pháp KNN**. Do đó phương pháp MF phù hợp để xây dựng hệ thống khuyến nghị dịch vụ VAS.

1.4.1.4. Tóm tắt công trình nghiên cứu và hướng đi cuối

Tóm lại, Kết quả đạt được của công trình luận văn thạc sĩ của tác giả Kiều Xuân Chấn trên đã chỉ ra được tầm quan trọng của dịch vụ VAS cũng như đã chỉ ra được phương pháp tối ưu nhất cho việc xây dựng một hệ thống khuyến nghị trong ngành viễn thông. Và cũng theo tác giả, hướng đi tiếp theo cho luận văn sẽ là thử nghiệm nhiều đặc trưng của bài toán khuyến nghị VAS trên thuật toán KNN để tìm kiếm kết quả tốt hơn, đồng thời kết hợp nhiều phương pháp lọc lai cùng với deep learning để xử lý bài toán và tìm kiếm kết quả thực tế tốt hơn.

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

1.4.2.1. Tổng quan về công trình nghiên cứu

Đề tài "Giải quyết vấn đề phân phối trong hệ thống khuyến nghị dựa trên đặc trưng nội dung của đối tượng" là một công trình luận văn thạc sĩ của tác giả Nguyễn Văn Đạt, công trình này được viết và trình bày vào năm 2021 tại trường Đại học Công nghệ thuộc Đại học Quốc gia Hà Nội. Trong công trình của mình, tác giả đã đề cập tới 2 vấn đề trong phương pháp lọc dựa trên nội dung rằng "*Mặc dù thuật toán lọc dựa trên nội dung (Content Base – CB) là một thuật toán tốt. Tuy nhiên, trong một số trường hợp, tính chất bắt buộc khác nhau do đó kết quả gợi ý từ thuật toán lọc dựa trên nội dung vẫn chưa đáp ứng được độ chính xác cao khi bài toán liên quan đến độ tương tự về phân phối giữa các thành phần giữa các thuộc tính của đối tượng. Thêm nữa, các phương pháp để đo mức độ tương đồng giữa các sản phẩm cũng là một vấn đề quan trọng ảnh hưởng đến độ chính xác của các thuật toán lọc dựa trên nội dung trong các bài toán về độ tương đồng giữa các phân phối*". Và để giải quyết 2 vấn đề

đã được đề cập trên, tác giả của công trình đã đề xuất một thuật toán lọc dựa trên nội dung mới dựa trên mô hình hỗn hợp Gaussian(Gaussian Mixture Model – GMM) nhằm tăng độ chính xác cho đầu ra. Ngoài ra, tác giả còn đề xuất mô hình thực nghiệm trên một bộ dữ liệu về rượu bao gồm 6 mùi vị, dữ liệu tag mô tả về rượu và một số trường hợp thông tin khác.

Như vậy, công trình luận văn của tác giả hướng đến việc xây dựng một một thuật toán khuyến nghị dựa trên nội dung mới nhằm khắc phục những nhược điểm của các thuật toán lọc dựa trên nội dung đã được công bố trước đó đối với dạng đặc trưng phân phối. Đầu vào nhận bộ sản phẩm có đặc trưng chính là phân phối thuộc tính, thuật toán cần phát triển một *Bộ biểu diễn đặc trưng* thích hợp có khả năng chọn ra các đặc trưng tốt nhất làm tiền đề được sử dụng trong các thuật toán cốt lõi trong *Bộ lọc sản phẩm* [5]. Các thuật toán này cần tận dụng và phát huy tối đa được các đặc điểm đặc biệt, quan trọng của đặc trưng phân phối so với các loại đặc trưng khác như văn bản, số nguyên, dữ liệu rời rạc [5]. Đồng thời, tác giả đã khẳng định rằng "*Đối với thuật toán đề xuất, nó không chỉ hoạt động tốt trên bộ dữ liệu thực nghiệm, mà hoàn toàn còn có thể áp dụng trên các bộ dữ liệu khác có độ tương tự về đặc trưng phân phối đối với bộ dữ liệu được phân tích*" nhằm đảm bảo rằng công trình của ông có thể hoạt động tốt đối với nhiều dạng dữ liệu khác nhau.

Trong công trình của mình, tác giả cũng đã liệu kê các thuật toán mà ông đã phân tích như sau:

- Thuật toán khuyến nghị CFRS
- Thuật toán gợi ý dựa trên nội dung (CB)
- Thuật toán so sánh độ tương đồng
    - Euclidean Distance (ED)
    - Gaussian Mixture Model (GMM)
    - Word Embeddings (WE)

1.4.2.2. Đề xuất thuật toán

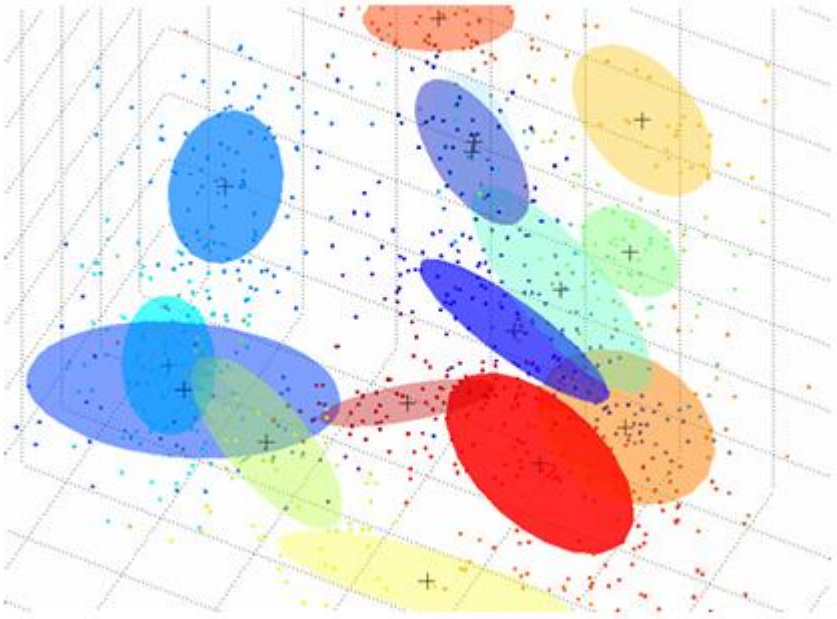
Trong công trình của mình, tác giả đề xuất và thu bộ dữ liệu đầu vào làm dữ liệu thực nghiệm cho bài toán. Bộ dữ liệu mà tác giả thu thập để thực nghiệm là một

bộ dữ liệu về rượu, cụ thể hơn là rượu Sake và được thu thập từ bộ dữ liệu của Sakenowa (Là một trang web uy tín và nổi tiếng chuyên bán rượu sake tại Nhật). Bên cạnh đó, tác giả còn phải đánh giá lại bộ dữ liệu dựa trên đặc tính như tên, thương hiệu, năm sản xuất… và cũng đồng thời nêu ra các khó khăn trong việc chuẩn hóa bộ dữ liệu thực nghiệm cho bài toán, "*Đây thực sự là một bộ dữ liệu với nhiều thách thức, với kích thước nhỏ, thiếu dữ liệu hoặc dữ liệu không đồng đều ở một số thuộc tính dẫn đến sự rời rạc trong dữ liệu Cụ thể hơn, hơn 30% các giá trị trường 6 chỉ số là rỗng, ~2% không tồn tại chỉ số mùi vị tags. Thêm nữa, nhiều giá trị tags là không chính xác, không tin cậy cần được tiền xử lý và xoá bỏ nhiễu*" - theo tác giả chia sẻ. Tiếp đó, Xây dựng thành phần thuật toán gồm 3 phần chính: Bước 1 chọn các đặc trưng quan trọng để xây dựng các vector biểu diễn cho từng sản phẩm, và khi đã có 1 tập vector thì thay vì sử dụng các công thức từ thuật toán để ra luôn kết quả xử lý thì ma trận vector này dẽ được đưa vào một mô hình phân cụm GMM (Gaussian Mixture Model) để tìm ra các cụm riêng biệt, nơi mà các sản phẩm có xu hướng giống nhau về đặc trưng vector sẽ được chọn. Bước 2 và cũng là bước cuối cùng sẽ trả về các kết quả gợi ý đối với mỗi sản phẩm truy vấn. Các giai đoạn thực hiện được mô tả như sau:

1) Tiền xử lý dữ liệu đối với dữ liệu thực nghiệm, theo tác giả đã đề cập trong công trình thì việc khai thác dữ liệu văn bản (text mining) là vô cùng quan trọng trong mọi bài toán liên quan đến văn bản, và các thuật toán lọc dựa trên nội dung cũng tương tự. Tác giả đã đề xuất phương pháp đối với bộ dữ liệu thực nghiệm của mình như sau "*thuật toán chỉ chọn tags mùi vị và 6 chỉ số mùi vị như là các đặc trưng chính cho việc tính toán mức độ tương đồng giữa các sản phẩm. Trong đó, tags mùi vị là một tập các văn bản được viết bằng tiếng Nhật và cần được làm sạch và cấu trúc lại trước khi đưa vào mô hình tính toán. Thuật toán chuyển đổi 6 chỉ số mùi vị thành số thực, và cần thực hiện một số thuật toán làm sạch và cấu trúc lại cho dữ liệu văn bản như tokenization, stemmings, stop word removal, tìm và thay thế từ đồng nghĩa, lemmatization, ... trước khi sử dụng. Trong đó, trường chỉ số tags mùi vị đã được tách thành các từ có nghĩa, vì vậy, thuật toán có thể bỏ qua bước*

tokenization và thực hiện các bước tiếp theo", việc đề xuất trên sẽ giúp cho dữ liệu được sạch hơn và tái cấu trúc giúp cho việc thao tác với dữ liệu thực nghiệm một cách dễ dàng hơn.

2) Tiếp đến là phân cụm dữ liệu, tác giả cho rằng "*các thuật toán lọc dựa trên nội cổ điển sẽ xây dựng một vector biểu diễn cho các thuộc tính được chọn của từng loại rượu, rồi tận dụng các ông thức so sánh độ tương đồng như Cosine hoặc Euclidean để sắp xếp và trả về top m kết quả*". Tuy nhiên trong thực tế, đối khi nội dung của các bộ dữ liệu không được chính xác, có thể bị nhiễu nhiều, hoặc cũng có thể bị ảnh hưởng bởi các yếu tố ngoại lại dẫn đến việc làm ảnh hưởng đến kết quả cuối cùng và tác giả cũng nhấn mạnh rằng nếu áp dụng công thức so sánh trên toàn bộ tập dữ liệu sẽ làm giảm hiệu xuất, tốc độ tính toán của thuật toán, nhất là trong các bộ dữ liệu lớn. Do đó, đề xuất của tác giả trong trường hợp này là nhóm tất cả sản phẩm dựa theo phân phối của chúng (trong bộ dữ liệu thực nghiệm của tác giả là 6 mùi vị khác nhau) thành các nhóm riêng biệt, làm tiền đề cho giai đoạn tiếp theo.



Hình 1. 9. Hình ảnh được lấy từ công trình minh họa cho việc dữ liệu được phân cụm

3) Sau khi phân cụm dữ liệu ở giai đoạn trước, để trả về kết quả gợi ý cho sản phẩm truy vấn nhiệm vụ cần làm là áp dụng một thuật toán sắp xếp lên các

cụm nơi các sản phẩm truy vấn được tìm ra từ mô hình huấn luyện GMM. Tác giả công trình đã phân tích bộ dữ liệu này như sau "*Đối với 6 chỉ số mùi vị, có thể nhận thấy chúng tuân theo phân phối Gaussian, tận dụng lợi thế này đồng thời dựa trên nguyên lý của hàm Gaussian Filter để tạo ra một công thức tính trọng số giữa các dãy 6 chỉ số mùi vị trong cụm. Còn đối với tags mùi vị, công thức so sánh độ tương đồng sẽ dựa trên khoảng cách Levenshtein (LD) giữa 2 chuỗi.*". Theo đề xuất trong công trình của tác giả, ông đã đề xuất các công thức tính độ tương đồng như: 1. Độ tương đồng trong phân phối với GFF, 2. Độ tương đồng chuỗi với LD, 3. Công thức sắp xếp tổng hợp.

4) Sau đó để tường minh hơn tác giả đã đưa ra tiến trình xử lý thuật toán đề xuất dưới dạng giả mã để người đọc dễ dàng hiểu và hình dung hơn về toàn bộ thuật toán, giả mã của tác giả được trình bày ở hình dưới 1.10.

Thuật toán: Thuật toán đề xuất

Đầu vào: Số cụm k

Đầu ra: Top *m* sản phẩm tương tự nhau của mỗi sản phẩm

Dữ liệu: Bộ dữ liệu *L*

1. Tiền xử lý dữ liệu cho các trường văn bản
2. Xây dựng ma trận vector 6 chiều đại diện cho 6 chỉ số mùi vị (f_1-f_6)
3. Lấy ma trận này như là đầu vào cho *GMM* để phục vụ cho quá trình huấn luyện và lưu giá trị cụm tương ứng cho mỗi sản phẩm vào bộ dữ liệu.
4. For *item in dataset* do:
   - Lấy ra số cụm của sản phẩm
   - Tìm tất cả những sản phẩm có cùng số cụm với sản phẩm truy vấn
   - Áp dụng công thức **(3.4)** để tính ***S(i,j)*** cho mỗi cặp sản phẩm
   - Trả về top *m* sản phẩm tương tự bằng cách sắp xếp theo thứ tự giảm dần

Hình 1. 10. Mô hình giả mã của tiến trình xử lý thuật toán của tác giả Nguyễn Văn Đạt

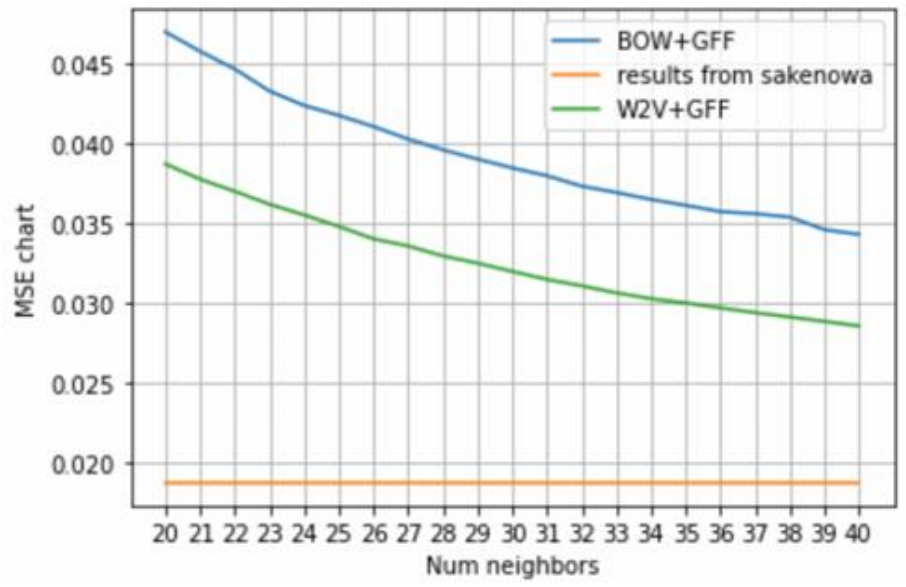
Trong thực tế, giải thuật của tác giả Nguyễn Văn Đạt hiện tại là giải thuật đang được sử dụng để giải quyết cho một hệ thống gợi ý rượu cho một công ty lớn tên là **Asian Frontier** tại Nhật Bản, điều đó cũng là đủ để cho chúng ta thấy độ tin tưởng của thuật toán là như thế nào. Đánh giá đối với thuật toán của chính bản thân mình,

ông khẳng định *"Dựa trên kết quả thực nghiệm so sánh với các thuật toán CB phổ biến, mạnh mẽ khác, thuật toán này không chỉ tốt hơn về chỉ số mùi vị trên kết quả trả về mà còn có khả năng xử lý, phản hồi người dùng nhanh hơn, đáp ứng hoàn toàn điều kiện cho một ứng dụng thời gian thực. Đặc biệt, thời gian được yêu cầu cho việc huấn luyện định kì cho thuật toán sau một khoảng thời gian các sản phẩm mới được thêm vào là nhanh, và không đáng kể so với các mô hình huấn luyện học sâu hiện tại."* đủ để ta thấy lời nói cùng độ tin tưởng với công trình của mình là răng thép.

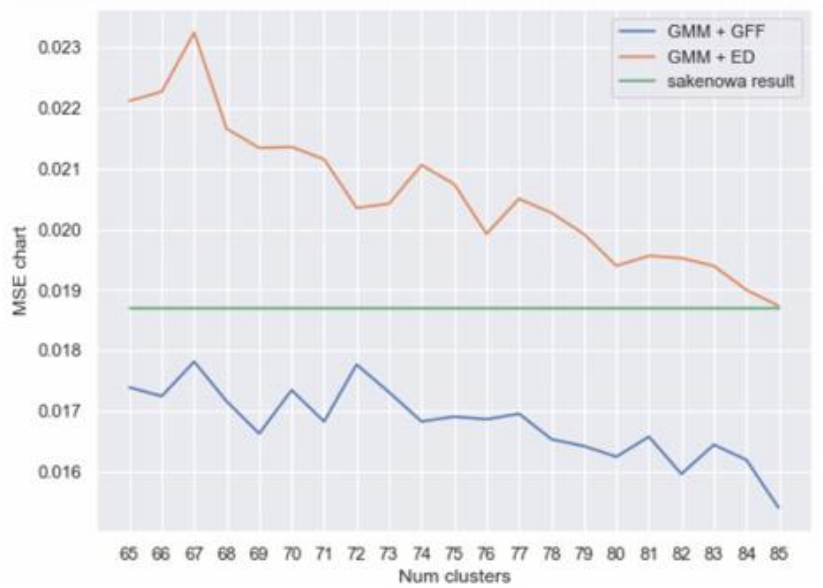
1.4.2.3. Kết quả thực nghiệm của công trình

Để chứng minh cho thuật toán đề xuất của mình hiệu quả đến như thế nào, tác giả đã đề xuất chi tiết kịch bản thử nghiệm đối với thuật toán đề xuất và 3 thuật toán lọc dựa trên nội dung phổ biến khác như BOW + GFF, W2V + GFF, và GMM + ED. Các kết quả thực nghiệm sẽ được tổng hợp, phân tích và so sánh để chứng minh độ hiệu quả, mạnh mẽ để chứng minh độ hiệu quả của thuật toán đã được tác giả đề xuất (GMM + GFF). Và để đánh giá, tác giả sử dụng phương pháp đánh giá MSE (Mean Square Error) để tính trung bình của bình phương lỗi giữa các kết quả để so sánh.

Sau cùng kết quả thực nghiệm trong công trình đã chỉ ra rằng, thuật toán đề xuất (GMM + GFF) cho ra kết quả tốt hơn 3 thuật toán lọc dựa trên nội dung phổ biến còn lại.



Hình 1. 11. Kết quả thực nghiệm từ công trình của tác giả Nguyễn Văn Đạt (MSE áp dụng BOW+GFF, W2V+GFF)



Hình 1. 12. Kết quả thực nghiệm từ công trình của tác giả Nguyễn Văn Đạt (MSE áp dụng GMM+GFF, GMM+ED)

Bên cạnh đó, thời gian xử lí trên mỗi truy vấn của thuật toán đề xuất cũng nhanh hơn 3 thuật toán còn lại:

BOW+GFF	GMM+ED	W2V+GFF	GMM+GFF
0.1856s	0.0174s	0.0251s	**0.0156s**

Hình 1. 13. Thời gian thực hiện truy vấn từ công trình của tác giả Nguyễn Văn Đạt

1.4.2.4. Tóm tắt công trình nghiên cứu và hướng đi cuối

Tóm lại, công trình nghiên cứu của tác giả Nguyễn Văn Đạt đã đề xuất được một thuật toán lọc dựa trên nội dung mới có độ hiệu quả cho các bài toán gợi ý dựa tên phân phối thuộc tính trong các hệ thống khuyến nghị, mấu chốt sử dụng các đặc trưng nội dung với thuật toán GMM và nó cũng đang được áp dụng để giải quyết một hệ thống khuyến nghị rượu hiện đang được triển khai tại Nhật Bản. Thuật toán khuyến nghị dựa trên phân phối này không chỉ đạt được độ chính xác cao, mà còn đạt được tốc độ xử lý nhanh, phù hợp với các ứng dụng thực tế, song nhược điểm của thuật toán là cần huấn luyện lại mô hình định kỳ sau khi có thêm một lượng sản phẩm mới được thêm vào hệ thống. Và cũng theo tác giả Nguyễn Văn Đạt, Hướng nghiên cứu trong tương lai sau luận văn là tìm cách cải thiện mô hình GMM trong khâu phân cụm sản phẩm để đạt được kết quả tốt hơn nữ

1.4.3. Kết luận rút ra từ khảo sát các công trình nghiên cứu có liên quan

Sau khi khảo sát các công trình nghiêm cứu ở 2 phần trên, kết luận đối với hệ thống khuyến nghị được rút ra rằng, các hệ thống khuyến nghị đều dựa trên các phương pháp cổ điển như lọc dựa trên nội dung (Contend Base Filtering – CB), lọc cộng tác (Colaborative Filtering – CF) và cả phương pháp lai (Hybird Filtering) để nhằm mục đích xây dựng bài toán để xử lý các vấn đề gặp phải của các hệ thống cũng như là tiền đề để xây dựng, cải tiến các bài toán khác. Bên cạnh đó các thuật toán đã được xây dựng cũng dựa vào tiền đề là các tiếp cận của các phương pháp trên. Như vậy để có thể hiểu rõ việc nghiên cứu, xây dựng hay khảo sát một hệ thống khuyến nghị hay các bài toán liên quan thì trước tiên ta phải hiểu bản chất của hệ thống khuyến nghị, các vấn đề mà các phương pháp đang gặp phải và cả các thuật toán, thuật giải, phương pháp mà những người xây dựng công trình tiên phong trước kia đã đạt được.

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

Bài toán khuyến nghị phim là một bài toán gồm các thuộc tính là sản phẩm và người dùng. Sản phẩm ở đây chính là các bộ phim và người dùng ở đây chính là những khán giả có nhu cầu thưởng thức bộ phim, và trong mỗi bộ phim đều có các thuộc tính như thể loại, thời lượng, loại phim, đánh giá, tác giả… và đối với mỗi đối tượng khán giả đều có các thuộc tính riêng biệt nhau về sở thích và nhu cầu xem phim. Bài toán khuyến nghị phim không hệ xa lạ nó đã tồn tại rất lâu trong các hệ thống xem phim nhằm hỗ trợ cho khách hàng tìm kiếm thông tin về một bộ phim mà họ yêu thích. Trong đồ án chuyên ngành này sẽ tập trung phân tích và khảo sát các phương pháp giải quyết bài toán khuyến nghị phim này, các phương pháp đó được biết đến như một thuật toán hay hệ thống nhằm giải quyết bài toán khuyến nghị phim. Và các phương pháp đó sẽ được trình bày riêng trong chương tiếp theo trong đồ án chuyên ngành này.

1.6. Tóm tắt chương 1

Chương này đã trình bày tổng quan và bài toán chung của hệ thống khuyến nghị, đồng thời trình bày chi tiết các phương pháp và các hướng tiếp cận cụ thể của từng phương pháp (thông tin được trình bày được tham khảo từ một số tài liệu công trình luận văn thạc sĩ, tiến sĩ trong nước). Bên cạnh đó cũng tiến hành khảo sát cụ thể các công trình nghiên cứu luận văn thạc sĩ của 2 tác giả Kiều Xuân Chấn và Nguyễn Văn Đạt nhằm mở rộng tầm mắt về những gì làm được trong hệ thống khuyến nghị của 2 tác giả trên. Cuối cùng là giới thiệu sơ lược về chủ đề chính của đồ án chuyên ngành này đó là bài toán khuyến nghị phim.

TÀI LIỆU THAM KHẢO

Tiếng Việt

[1] Kiều Xuân Chấn, "Nghiên cứu và xây dựng Hệ thống Khuyến nghị cho bài toán dịch vụ giá trị gia tăng trong ngành viễn thông", Luận văn thạc sĩ, Đại học quốc gia Hà Nội – Đại học Công Nghệ, Hà Nội, 2017

[2] Huỳnh Ngọc Tín, "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", Luận văn thạc sĩ, Đại học quốc gia Thành phố Hồ Chí Minh – Trường Đại học Công nghệ thông tin, Tp.Hồ Chí Minh, 2016

[3] Bùi Văn Minh, "Nghiên cứu, Xây dựng Hệ thống Khuyến nghị phim tự động", Luận văn thạc sĩ, Học viện Bưu chính viễn thông, Hà Nội, 2017

[4] Đỗ Thị Liên, "Phát triển một số phương pháp xây dựng hệ tư vấn", Luận văn tiến sĩ, Học viện Bưu chính Viễn thông, Hà Nội, 2020

[5] Nguyễn Văn Đạt, "Giải quyết vấn đề phân phối trong hệ thống khuyến nghị dựa trên đặc trưng nội dung của đối tượng", Luận văn thạc sĩ, Đại học Quốc gia Hà Nội – Trường đại học Công nghệ, Hà Nội, 2021

Tiếng anh

[6] Gediminas Adomavicius, Alexander Tuzhilin, *Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions*, IEEE Trans on Knowl and Data Eng, 2005

[7] Ricardo Baeza-Yates, Ricardo Baeza-Yates, *Modern Information Retrieval*, Addison-Wesley Longman Publishing Co., Boston – USA, 1999

[8] Marco de Gemmis, Pasquale Lops, Pasquale Lops, Pasquale Lops, *Integrating tags in a semantic content-based recommender*, In proceedings of the 2008 ACM Conference on Recommender System, 2008

[9] Dietmar Jannach, Markus Zanker, Alexander Felfernig, Gerhard Friedrich, *Recommender System: An Introduction*, Cambridge University Press, New York, 2010

[10] Thorsten Joachims, *Text Categorization With Support Vector Machines Learning with Many Relevant Features*, In procedings of the 10th Europen Conference on Machine Learning, 1999

[11] Raymond J. Mooney, Loriene Roy, *Content-based book recommending using learning for text categorization*, In proceedings of the Fifth ACM Conference on Digital Libraries, 2000

[12] Kamal Nigam, Andrew Kachites Mccallum, Sebastian Thrun, Tom Mitchell, *Text Classification from Labeled and Unlabeled Documents using EM*, Machine Learning, 2000

[13] Michael Pazzani, Daniel Billsus, *Learning and Revising User Profiles: The Identification of Interesting Web Sites*, Machine Learning, 1997

[14] John S. Breese, David Heckerman, Carl Kadie, *Empirical Analysis of Predictive Algorithms for Collaborative Filtering*, In Proceeding of the Fourteenth Conference on Uncertainty in Artificial Intelligence, Fan Francisco, 1998

[15] Robin Burke, *Hybrid Recommender Systems: Survey and Experiments*, User Modeling and User-adapted Interaction, 2002