TẠP CHÍ KHOA HỌC VÀ CÔNG NGHỆ ĐẠI HỌC DUY TÂN
DTU Journal of Science and Technology 02(75) (2026) 209-219

ĐẠI HỌC DUY TÂN
DUYTAN UNIVERSITY

Hệ thống khuyến nghị tài liệu học tập dựa trên quá trình học tập của sinh viên

Learning Materials Recommendation System Based on Student Learning History

Lê Thanh Long[a*]
Le Thanh Long[a*]

[a]*Khoa Trí tuệ nhân tạo, Trường Khoa học Máy tính và Trí tuệ Nhân tạo, Đại học Duy Tân, Đà Nẵng, Việt Nam*
[a]*Faculty of Artificial Intelligence, School of Computer Science and Artificial Intelligence, Duy Tan University, Da Nang, 550000, Viet Nam*

(Ngày nhận bài: 07/11/2025, ngày phản biện xong: 24/01/2026, ngày chấp nhận đăng: 07/04/2026)

Tóm tắt

Bài báo giới thiệu một hệ thống khuyến nghị tài liệu học tập cá nhân hóa dựa trên lịch sử tương tác của sinh viên trong hệ thống LMS (Learning Management System). Hệ thống triển khai và so sánh hai phương pháp chính: Lọc cộng tác dựa trên người dùng (CF) và Lọc theo nội dung (CBF), đồng thời áp dụng mô hình lai (Hybrid) để nâng cao hiệu quả. Thử nghiệm trên bộ dữ liệu giả lập gồm 500 sinh viên và 10.000 tài liệu cho thấy CF đạt Precision 55%, Diversity 68%, CBF đạt Precision 63%, Diversity 33% trong khi mô hình lai (Hybrid) đạt Precision 57%, Diversity 67% cải thiện chất lượng khuyến nghị. Kết quả chứng minh tính khả thi của giải pháp trong việc nâng cao khả năng tìm kiếm tài liệu phù hợp, đồng thời góp phần cải thiện trải nghiệm học tập trong môi trường giáo dục trực tuyến.

Từ khóa: Hệ thống khuyến nghị, học trực tuyến, lọc cộng tác, lọc dựa trên nội dung, phương pháp lai, đồ thị kiến thức

Abstract

This paper presents a personalized learning materials recommendation system based on students' interaction history in a Learning Management System (LMS.) The system implements and compares two main methods: User-based Collaborative Filtering (CF) and Content-based Filtering (CBF), while also applying a hybrid model to improve the efficiency. Experiments on a simulated dataset of 500 students and 10,000 documents show that CF achieves a Precision of 55%, and a Diversity of 68%; CBF achieves a Precision of 63%, and a Diversity of 33% while the Hybrid model achieves a Precision of 57%, and a Diversity of 67%, improving the recommendation quality. The results demonstrate the feasibility of the solution in enhancing the ability to find suitable documents, while contributing to improving the learning experience in an online education environment.

Keywords: Recommendation system, online learning, collaborative filtering, content-based filtering, hybrid method, knowledge graph

[*]*Tác giả liên hệ:* Lê Thanh Long
Email: lthanhlong@gmail.com

1. Giới thiệu

Hệ thống khuyến nghị (Recommender System – RS) ngày nay đã trở thành thành phần hạ tầng quan trọng trong các nền tảng học tập số, giúp giảm thiểu tình trạng quá tải thông tin và cá nhân hóa quá trình học ở quy mô lớn. Trong lĩnh vực giáo dục, RS được ứng dụng để khuyến nghị khóa học, tài nguyên học tập (video, bài đọc, bài tập) và lộ trình học, đồng thời ngày càng tích hợp các tín hiệu ngữ cảnh và yếu tố sư phạm [1],[2].

Các khảo sát và hội thảo gần đây chỉ ra xu hướng chuyển dịch từ các phương pháp truyền thống như Collaborative Filtering (CF) và Content-based Filtering (CBF) sang các mô hình lai, nhận biết ngữ cảnh, và tận dụng đồ thị tri thức nhằm gắn kết tốt hơn với kết quả học tập và hỗ trợ của giảng viên [1],[3],[4].

Trong bối cảnh học tập trực tuyến phát triển mạnh, sinh viên phải tiếp cận một khối lượng tài liệu lớn, khiến việc tìm kiếm nội dung phù hợp trở nên tốn thời gian và có thể ảnh hưởng đến hiệu quả học tập. RS đóng vai trò giải quyết vấn đề này bằng cách cá nhân hóa danh sách tài liệu dựa trên hành vi, sở thích và năng lực của từng người học.

Nghiên cứu này tập trung vào thiết kế và triển khai hệ thống khuyến nghị tài liệu học tập dựa trên lịch sử học của sinh viên, áp dụng hai phương pháp chính là CF – dựa trên sự tương đồng giữa người học hoặc giữa các tài liệu – và CBF – dựa trên đặc điểm nội dung của tài liệu và hồ sơ sở thích của người học; đồng thời xây dựng mô hình Hybrid để kết hợp ưu điểm của hai phương pháp này.

2. Các công trình liên quan

Các đánh giá hệ thống gần đây nhấn mạnh sự phát triển của RS giáo dục, vượt ra ngoài việc chỉ nhắm mục tiêu vào người học, sang các thiết kế đa bên liên quan, đồng thời hỗ trợ giáo viên [1]. Đồ thị kiến thức và mạng nơ-ron cho phép mô hình hóa chi tiết các khái niệm khóa học và trạng thái của người học, cải thiện chất lượng khuyến nghị và khả năng giải thích [2], [5]–[7]. Các mô hình sâu đã được điều chỉnh để khuyến nghị khóa học trực tuyến với học thuộc tính ngữ nghĩa và mô hình hóa hành vi [8], [9]. RS nhận biết ngữ cảnh (CARS) kết hợp rõ ràng các biến tình huống (thời gian, tiến trình, thiết bị, hoạt động) và tính minh bạch, những yếu tố ngày càng được yêu cầu trong môi trường TEL (Technology-Enhanced Learning) [3]. Các biến thể phân tích nhân tử bậc thấp và thưa thớt vẫn có tính cạnh tranh, đặc biệt là trong điều kiện thưa thớt và khởi động nguội, và hiện được bổ sung bằng các phương pháp dựa trên đồ thị có thể giải thích được [6], [10], [11]. Các khảo sát liên miền toàn diện từ năm 2017 đến năm 2024 tổng hợp thêm các lựa chọn mô hình hóa và giao thức đánh giá cho việc triển khai công nghiệp [4].

Mặc dù hệ thống khuyến nghị giáo dục đã đạt nhiều tiến bộ, vẫn tồn tại các khoảng trống quan trọng: cá nhân hóa chưa tích hợp đầy đủ phong cách học, mục tiêu nghề nghiệp, trạng thái cảm xúc; khả năng xử lý dữ liệu đa phương tiện còn hạn chế khi tài liệu học tập ngày càng đa dạng; thiếu khuyến nghị thời gian thực trong quá trình học; đánh giá hiệu quả còn thiên về độ chính xác thay vì tác động lâu dài đến kết quả và động lực học tập; và chưa chú trọng đầy đủ tới ngữ cảnh văn hóa – ngôn ngữ đặc thù như trường hợp sinh viên Việt Nam.

3. Cơ sở lý thuyết

Hệ thống khuyến nghị (RS) là một nhánh quan trọng của Trí tuệ nhân tạo (AI) và Khai phá dữ liệu (Data Mining), được thiết kế nhằm dự đoán và cung cấp những mục (items) phù hợp với sở thích hoặc nhu cầu của người dùng. Trong các lĩnh vực như thương mại điện tử (Amazon, Shopee), giải trí (Netflix, Spotify) hay mạng xã hội (YouTube, TikTok), RS đã trở thành một công cụ then chốt giúp tăng mức độ tương tác và trải nghiệm cá nhân hóa. Về mặt kỹ thuật, RS thường dựa trên một trong ba cách tiếp cận chính CBF, CF và Hybrid Systems.

Lọc cộng tác dựa trên người dùng là phương pháp khuyến nghị dựa trên giả định rằng người dùng có lịch sử hành vi tương tự sẽ có xu hướng quan tâm đến các mục giống nhau trong tương lai. Thuật toán thường sử dụng các phép đo tương đồng như cosine similarity, Pearson correlation hoặc adjusted cosine để xác định nhóm "láng giềng" gần nhất của người dùng mục tiêu trong không gian đặc trưng người dung – đối tượng. Giai đoạn dự đoán điểm khuyến nghị (rating prediction) thường áp dụng kỹ thuật weighted sum hoặc mean-centered weighting, có thể tối ưu hóa bằng top-K neighbor selection nhằm giảm chi phí tính toán. Hạn chế chính gồm vấn đề thưa dữ liệu và khả năng mở rộng, đặc biệt khi tập người dùng lớn; do đó, nhiều hệ thống thực tế áp dụng chiến lược lai hoặc phân cụm để cải thiện hiệu năng và chất lượng khuyến nghị.

Lọc dựa trên nội dung CBF là kỹ thuật khuyến nghị dựa trên việc phân tích đặc trưng của các mục (items) mà người dùng đã tương tác trước đó, sau đó tìm các mục mới có đặc trưng tương tự. Hệ thống thường biểu diễn mục dưới dạng vector thuộc tính (feature vector) – ví dụ: từ khóa, thẻ (tags), embedding từ mô hình NLP, hoặc đặc trưng trích xuất từ ảnh/âm thanh – và áp dụng các phép đo tương đồng như cosine similarity, Euclidean distance hoặc dot product để xếp hạng. Điểm mạnh của CBF là khả năng hoạt động tốt ngay cả khi không có dữ liệu từ cộng đồng (giảm cold-start cho người dùng mới), nhưng hạn chế là chỉ giới hạn trong không gian đặc trưng đã biết, khó khám phá các mục hoàn toàn mới về nội dung.

Hybrid Systems: Hệ thống lai kết hợp CBF và CF nhằm khai thác đồng thời ưu điểm của cả hai phương pháp. CBF tận dụng đặc trưng nội dung để khuyến nghị cho những đối tượng mới hoặc ít tương tác, trong khi CF khai thác sự tương đồng hành vi giữa người dùng để phát hiện sở thích tiềm ẩn khó quan sát từ nội dung. Việc tích hợp này giúp tăng độ chính xác, mở rộng phạm vi khuyến nghị và giảm thiểu các hạn chế như cold-start hoặc sparsity, đặc biệt hữu ích trong các hệ thống khuyến nghị quy mô lớn và dữ liệu đa dạng.

Ngoài các phương pháp nền tảng, nhiều nghiên cứu đã mở rộng bằng cách tích hợp các yếu tố như chấm điểm theo ngữ cảnh (context-aware scoring) để cá nhân hóa khuyến nghị dựa trên thời điểm, thiết bị hoặc trạng thái; tăng cường bằng đồ thị tri thức (knowledge-graph augmentation) nhằm khai thác quan hệ giữa khái niệm, môn học và tài nguyên học tập; và học trên đồ thị (graph learning) cho mạng hai phía (user – item) để nâng cao độ chính xác cũng như khả năng giải thích kết quả khuyến nghị [2], [3], [5].

4. Phương pháp nghiên cứu

4.1. Nguồn và đặc điểm dữ liệu

Dữ liệu nghiên cứu được sinh giả lập với 500 sinh viên và 10.000 tài liệu học tập. Thông tin lịch sử học tập bao gồm tài liệu đã truy cập, thời gian và số lần truy cập, và được sinh giả lập hơn 1 triệu đánh giá của sinh viên (thang 1–5). Dữ liệu được chuẩn hóa thành User–Item Rating Matrix, trong đó User là mã sinh viên, Item là mã tài liệu, còn Rating biểu thị mức độ tương tác hoặc điểm đánh giá (quy đổi từ thời gian học, số lần truy cập hoặc đánh giá sao).

4.2. Quy trình tiền xử lý dữ liệu

Trước khi áp dụng thuật toán khuyến nghị, dữ liệu được xử lý qua các bước gồm: làm sạch dữ liệu bằng cách loại bỏ bản ghi trùng lặp và xử lý dữ liệu thiếu thông qua kỹ thuật giá trị trung bình hoặc loại bỏ nếu tỷ lệ thiếu vượt quá 50%; chuẩn hóa đánh giá bằng cách quy đổi các chỉ số tương tác về thang điểm chuẩn (ví dụ 1–5) và áp dụng z-score normalization khi cần thiết; mã hóa dữ liệu bằng index encoding để biểu diễn sinh viên và tài liệu dưới dạng chỉ số số học; cuối cùng là tách dữ liệu thành tập huấn luyện và kiểm tra bằng train-test split tỷ lệ 80% train – 20% test.

4.3. Kỹ thuật khuyến nghị

4.3.1. Kỹ thuật Lọc cộng tác

Các mô hình hiện đại như Matrix Factorization (SVD, ALS, NMF) và Neural Collaborative Filtering (NCF) học vector tiềm ẩn (latent factors) đại diện cho sở thích user và đặc trưng item, cho phép khái quát hóa trong không gian thưa. Trong bài báo này, tác giả sử dụng mô hình Matrix Factorization (SVD) nhằm biểu diễn người dùng và đối tượng trong không gian tiềm ẩn (latent space):

$$R \approx PQ^T \ (1)$$

Trong đó:

- $P \in R^{m \times k}$: ma trận vector tiềm ẩn của người dùng
- $Q \in R^{n \times k}$: ma trận vector tiềm ẩn của sản phẩm
- k: số chiều của không gian đặc trưng tiềm ẩn

Dự đoán rating: $r_{ui} = {}^{pT}{}_u q_i$, với p_u, q_i lần lượt là vector tiềm ẩn của user u và item i. Việc huấn luyện tối ưu $\min_{P,Q} \sum_{u,i) \in K} (r_{ui} - P_u^T q_i) + \lambda(||p_u||^2 + ||p_u||^2)$, trong đó K là tập các cặp (u,i) có rating quan sát được.

4.3.2. Kỹ thuật Lọc nội dung

CBF là phương pháp gợi ý dựa trên mô hình hoá nội dung item và hồ sơ sở thích user trong cùng không gian đặc trưng. Mỗi item i được biểu diễn bằng vector đặc trưng nội dung: $x_i = [x_{i1}, x_{i2}, \ldots, x_{id}]^T \in R^d$, trong đó x_{ik} là giá trị đặc trưng thứ k (có thể từ *TF–IDF*, embedding, one-hot, v.v.). Hồ sơ người dùng u được xây dựng từ các item mà họ đã đánh giá cao hoặc tương tác tích cực. Có hai cách phổ biến xây dựng hồ sơ người dùng là phương pháp Trung bình đơn giản (Average Profile) và phương pháp Trung bình có trọng số (Weighted Profile). Mức độ phù hợp hoặc điểm gợi ý r_{ui} giữa user u và item i thường được đo bằng độ tương đồng cosine. Trong bài báo, thay vì dùng độ đo cosine tuyến tính, tác giả áp dụng hàm ánh xạ phi tuyến $r_{ui} = f(p_u, x_i)$, trong đó f là mô hình RF nhằm mô hình hoá mối quan hệ phức tạp giữa các đặc trưng của người dùng và đối tượng. Thay vì huấn luyện một cây duy nhất dễ bị quá khớp (overfitting), RF huấn luyện nhiều cây ngẫu nhiên và lấy trung bình hoặc bỏ phiếu kết quả của chúng.

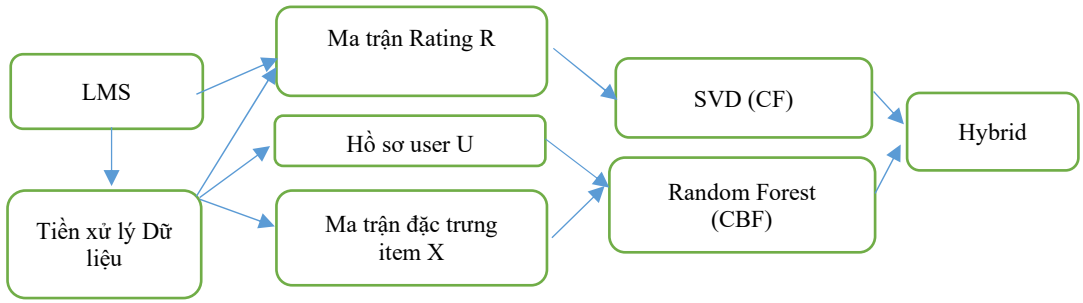
Trong bài báo này, RF được lựa chọn vì phù hợp với dữ liệu dạng bảng (tabular), gồm các đặc trưng rõ ràng như độ tương đồng nội dung, độ phổ biến và chỉ báo danh mục. So với Neural Network hay Deep Learning, Random Forest ổn định hơn với dữ liệu không quá lớn, ít rủi ro overfitting và dễ huấn luyện. Ngoài ra, mô hình này có khả năng diễn giải tốt hơn thông qua phân tích tầm quan trọng đặc trưng, phù hợp với yêu cầu minh bạch trong hệ thống khuyến nghị giáo dục.

4.3.3. Mô hình lai

Mô hình lai là phương pháp kết hợp CF và CBF, nhằm khai thác ưu điểm của từng mô hình và khắc phục nhược điểm riêng lẻ. Trong cấu hình lai tuyến tính điển hình, điểm gợi ý được tính theo công thức: $Score(u,i) = \alpha \cdot CF(u,i) + (1-\alpha) \cdot CBF(u,i)$, trong đó α điều chỉnh mức ưu tiên giữa tín hiệu cộng đồng (CF) và nội dung (CBF).

5. Triển khai hệ thống

5.1. Kiến trúc hệ thống



Hình 1. Kiến trúc hệ thống khuyến nghị sử dụng CF, CBF và hybrid

Hệ thống (Hình 1) được thiết kế theo kiến trúc hai nhánh song song CF và CBF kết hợp tại lớp hybrid. Dữ liệu từ LMS được tiền xử lý để tạo ma trận rating R và ma trận đặc trưng nội dung item X. Nhánh CF sử dụng SVD để sinh ma trận dự đoán R_{CF} dựa trên hành vi cộng đồng. Nhánh CBF xây dựng hồ sơ người dùng và dùng Random Forest để tạo ma trận dự đoán R_{CBF}. Hai nguồn dự đoán được kết hợp theo $R_H=0.8R_{CF}+0.2R_{CBF}$.

Hệ thống được xây dựng bằng Python với các thư viện như scikit-learn và Surprise, triển khai cả thuật toán CF và CBF, đồng thời kết hợp hai phương pháp để tạo ra mô hình lai (hybrid) nhằm tối ưu hóa chất lượng gợi ý. Sau đó tiến hành so sánh, đánh giá độ chính xác và độ đa dạng giữa 3 phương pháp.

5.2. Luồng xử lý dữ liệu và khuyến nghị

Quy trình khuyến nghị được triển khai qua các bước sau: đầu tiên hệ thống sinh dữ liệu giả lập; tiếp đến dữ liệu được tiền xử lý thông qua các bước làm sạch, chuẩn hóa điểm đánh giá và tính toán TF-IDF cho tài liệu; sau đó tiến hành tính toán mô hình khuyến nghị, trong đó CF xây dựng ma trận tương đồng người dùng, còn CBF tạo vector đặc trưng nội dung tài liệu và hồ sơ sinh viên; bước tiếp theo là tạo danh sách khuyến nghị bằng cách dự đoán điểm cho những tài liệu chưa học rồi sắp xếp theo thứ tự giảm dần; cuối cùng, danh sách khuyến nghị được hiển thị cho người dùng.

5.3. Mã minh họa

```
#-------------------------------------------------------------------------
#Thuật toán 1: Lọc cộng tác CF (SVD)
#-------------------------------------------------------------------------
Đầu vào:
  Tập đánh giá R
  Số users n_u , Số items n_i
Đầu ra:
  Ma trận dự đoán R̂
Phương pháp:
Chuẩn bị dữ liệu
  Xây dựng tập dữ liệu huấn luyện D_train ← build_full_trainset(R)
Khởi tạo mô hình
  Khởi tạo mô hình SVD.
Huấn luyện mô hình
  Huấn luyện SVD trên D_train
  SVD←Fit(D_train)
Khởi tạo ma trận dự đoán
  R̂ ← 0n_u x n_i
Dự đoán Rating
  for each user u=1 to n_u  do
for each item i=1 to n_i  do
   R̂ ←SVD.predict(u,i)
```

```
        end for
  end for
    Return R̂
```

Dữ liệu của tập đánh giá R được chuyển thành ma trận D_{train}, bằng build_full_trainset(). Mô hình SVD được huấn luyện trên ma trận D_{train}. Từ đó, hệ thống dự đoán điểm cho từng cặp (u, i) bằng predict(u, i), và lưu kết quả ma trận dự đoán $\hat{R}$ làm rating ước lượng, phục vụ cho việc xếp hạng và gợi ý.

```
    #-------------------------------------------------------------------------
    # Thuật toán 2: Xây dựng đặc trưng nội dung của Item
    #-------------------------------------------------------------------------
    Đầu vào:
      Bảng thông tin item items_df, Dữ liệu đánh giá ratings_df, Số lượng item n_i
    Đầu ra:
      Ma trận đặc trưng item chuẩn hóa X, Vector đặc trưng của toàn bộ item g
    Phương pháp:
    Trích xuất đặc trưng văn bản (TF-IDF)
      Áp dụng TF-IDF lên mô tả văn bản của từng item
    Mã hóa đặc trưng danh mục (One-Hot Encoding)
      Chuyển danh mục của mỗi item thành vector one-hot, Thu được ma trận X
    Tính toán đặc trưng độ phổ biến
      Xác định các đánh giá lớn hơn 3 là "like".
      Đếm số lượt "like" cho mỗi item.
      Chuẩn hóa số lượt "like" về khoảng [0,1].
    Kết hợp đặc trưng (Feature Fusion)
      Nhân mỗi loại đặc trưng với trọng số tương ứng.
      Ghép đặc trưng văn bản, danh mục và độ phổ biến thành một vector duy nhất cho mỗi
item.
    Chuẩn hóa biểu diễn item
      Chuẩn hóa mỗi vector đặc trưng của item về độ dài đơn vị.
    Tính vector đặc trưng toàn cục
      Lấy trung bình của tất cả vector đặc trưng item thành vector g.
    Trả về kết quả
      Xuất ma trận X và vector g.
```

Thuật toán này xây dựng đặc trưng nội dung (content features) cho từng item bằng cách kết hợp ba nguồn thông tin chính: TF-IDF (60%) biểu diễn nội dung văn bản mô tả item, sử dụng n-gram (1–2) và chuẩn hoá theo log-scale để giảm trọng số của các từ xuất hiện quá thường xuyên; One-Hot (30%) mã hoá thể loại item dưới dạng vector nhị phân, giúp mô hình phân biệt các nhóm danh mục khác nhau; Popularity (10%) phản ánh mức độ phổ biến dựa trên số lượt thích (rating > 3), được chuẩn hoá về khoảng [0,1] để đảm bảo tỷ lệ cân bằng. Ba thành phần này được ghép lại tạo thành ma trận đặc trưng **X**. Vector trung bình g được tính làm chuẩn tham chiếu toàn cục cho các người dùng chưa có lịch sử tương tác.

```
    #-------------------------------------------------------------------------
    Thuật toán 3: Xây dựng hồ sơ người dùng (User Profiles)
    #-------------------------------------------------------------------------
    Đầu vào:
      Tập dữ liệu đánh giá R, Ma trận đặc trưng item X, Vector đặc trưng g
    Đầu ra:
      Ma trận hồ sơ người dùng U
    Phương pháp:
      Khởi tạo hồ sơ người dùng
       Tạo ma trận hồ sơ người dùng U với kích thước n_u×d_n ban đầu chứa toàn số 0.
      Với mỗi người dùng u=1 đến n_u, thực hiện:
        Xác định tập item mà người dùng yêu thích
        Xây dựng hồ sơ ban đầu của người dùng
           Nếu người dùng có ít nhất một item yêu thích thì U được tính bằng
           trung bình đặc trưng của các item đó. Ngược lại, sử dụng vector đặc
           trưng toàn cục g làm hồ sơ ban đầu.
        Làm trơn hồ sơ người dùng với đặc trưng toàn cục
        Chuẩn hóa hồ sơ người dùng
        Lưu hồ sơ người dùng
           Gán vector hồ sơ đã chuẩn hóa vào hàng tương ứng trong ma trận U.
        Trả về kết quả
           Xuất ma trận hồ sơ người dùng U.
```

Đoạn mã xây dựng hồ sơ người dùng U (user profiles): Với mỗi user *u*, lấy các item có rating > 3, tính trung bình vector đặc trưng của chúng để biểu diễn sở thích nội dung. Nếu user chưa có tương tác, dùng $\boldsymbol{g}$ làm thay thế. Trộn $0.9 \times profile + 0.1 \times global\_vec$ để giảm nhiễu, rồi chuẩn hoá L2. U là vector sở thích chuẩn hoá của user *u*, dùng để tính độ tương đồng nội dung trong CBF.

```
#-------------------------------------------------------------------------
Thuật toán 4: Lọc nội dung sử dụng Random Forest (CBF-RF)
#-------------------------------------------------------------------------
Đầu vào:
  Tập dữ liệu đánh giá R, Ma trận đặc trưng item X
  Ma trận hồ sơ người dùng U, Vector độ phổ biến của item
  Tập danh mục yêu thích của từng người dùng user_fav
  Số lượng người dùng n_u , số lượng item n_i
Đầu ra:
  Ma trận dự đoán R̂_CBF
     Phương pháp:
  Lấy mẫu dữ liệu huấn luyện từ tập R
  Xây dựng đặc trưng huấn luyện M
Với mỗi bản ghi được lấy mẫu (u,i,r_ui)
      Tính độ tương đồng giữa hồ sơ người dùng U_u và đặc trưng item X_i.
      Xác định xem danh mục của item i có thuộc danh mục yêu thích của người
        dùng u hay không.
      Lấy giá trị độ phổ biến của item.
      Tạo vector đặc trưng gồm: độ tương đồng nội dung, độ phổ biến của item,
        chỉ báo danh mục yêu thích.
      Sử dụng rating thực tế r_ui làm nhãn huấn luyện.
  Huấn luyện mô hình hồi quy
      Huấn luyện mô hình Random Forest Regressor trên tập đặc trưng M.
  Khởi tạo ma trận dự đoán
      Tạo ma trận rỗng R̂_CBF  kích thước n_u×n_i .
  Sinh dự đoán cho toàn bộ user-item
 Với mỗi người dùng u=1 đến n_u:
        Tính độ tương đồng giữa hồ sơ U_u và tất cả các item trong X.
        Với mỗi item, xây dựng vector đặc trưng.
        Dùng mô hình Random Forest để dự đoán rating cho tất cả item.
        Cắt giá trị dự đoán về khoảng [0,5].
        Lưu kết quả vào hàng u của R̂_CBF.
```

Đoạn mã huấn luyện CBF dựa trên ba tín hiệu chính**:** độ tương đồng nội dung (*sim*), độ phổ biến (*popularity*), và mức độ phù hợp thể loại (*fav*). Mô hình Random Forest học mối quan hệ phi tuyến giữa các đặc trưng này và rating thực tế, sau đó sinh ma trận dự đoán $\hat{R}_{CBF}$ cho toàn bộ user–item, làm nền cho phần hybrid.

```
#-------------------------------------------------------------------------
# Thuật toán 5: Mô hình khuyến nghị lai (Ưu tiên CF)
#-------------------------------------------------------------------------
Đầu vào:
  Ma trận dự đoán CF R̂_CF, Ma trận dự đoán CBF R̂_CBF
  Số lượng người dùng n_u , số lượng item n_i
  Trọng số kết hợp α=0,8 (ưu tiên CF)
Đầu ra:
  Ma trận dự đoán lai R̂_H
Phương pháp:
  Khởi tạo ma trận lai
    Tạo ma trận R̂_H kích thước n_u×n_i , ban đầu chứa toàn số 0.
  Kết hợp dự đoán cho từng người dùng
    Với mỗi người dùng u=1 đến n_u:
       Tính dự đoán lai bằng cách kết hợp tuyến tính giữa R̂_CF và R̂_CBF:
             Dự đoán của CF được gán trọng số α.
             Dự đoán của CBF được gán trọng số 1-α.
       Lưu kết quả vào hàng tương ứng trong R̂_H.
  Trả về kết quả
    Xuất ma trận dự đoán lai R̂_H .
```

Mô hình lai hybrid = α·CF + (1−α)·CBF kết hợp ưu thế của cả hai phương pháp CF và CBF. $\hat{R}_{CF}$ là ma trận dự đoán từ CF, trong khi $\hat{R}_{CBF}$ được sinh từ CBF. Trước khi kết hợp, dự đoán của cả SVD và Random Forest đều được chuẩn hóa về cùng thang điểm rating [0,5] thông qua bước clipping, đảm bảo hai ma trận $\hat{R}_{CF}$ và $\hat{R}_{CBF}$ có cùng đơn vị và biên độ. Nhờ đó, phép

cộng trọng số trong lớp hybrid phản ánh đúng mức độ ưu tiên của từng mô hình mà không bị lệch do khác biệt thang đo. Tham số α = 0.8 ưu tiên trọng số cho CF nhờ khả năng khai thác quan hệ cộng đồng, còn (1−α) = 0.2 dành cho CBF để bù đắp các trường hợp thiếu dữ liệu người dùng hoặc đối tượng (cold-start). α=0.8 được chọn nhằm nhấn mạnh vai trò của CF trong tăng tính đa dạng, dù *Precision* có thể giảm nhẹ so với CBF.

6. Thực nghiệm và kết quả

6.1. Mục tiêu thực nghiệm

Mục tiêu chính của phần thực nghiệm là: Đánh giá hiệu suất của hai phương pháp khuyến nghị CF và CBF. Sau đó dùng mô hình lai (hybrid) nhằm cải thiện chất lượng khuyến nghị. Mô hình sử dụng công thức (2) để tính độ chính xác *Precision@5* và công thức (3) để tính sự đa dạng *Diversity@5*.

$$Precision@k = \frac{\text{lượng item gợi ý đúng trong top k}}{k} \quad (2)$$

Item gợi ý đúng nghĩa là item đó nằm trong danh sách thật sự được người dùng yêu thích, trong trường hợp này, vì mỗi user được gợi ý k=5 item.

$$Diversity@k = 1 - MeanSim(R_u) \quad (3)$$

Trong đó: $MeanSim(R_u) = \frac{1}{k(k-1)} \sum_{a=1}^{k-1} \sum_{b=a+1}^{k} sim(x_{i_a}, x_{i_b})$ độ tương đồng trung bình giữa các cặp item và $sim(x_{i_a}, x_{i_b}) = \frac{}{||x_{i_a}||.||x_{i_b}||}$ là độ tương đồng cosine giữa hai item.

```
#-------------------------------------------------------------------------
# Thuật toán 6: Đánh giá hiệu suất hệ thống khuyến nghị
#-------------------------------------------------------------------------
Đầu vào:
  Ma trận dự đoán R̂_H , Ma trận đặc trưng item X
  Tập dữ liệu đánh giá R, Kích thước danh sách khuyến nghị Top-k với k=5
  Số lượng người dùng để đánh giá n_eval=1000
Đầu ra:
  Precision@k, Diversity@k, RMSE
Phương pháp:
  Khởi tạo các danh sách rỗng: Prec, Div.
  Với mỗi người dùng u=1 đến n_eval, thực hiện:
    Tạo danh sách khuyến nghị Top-k
    Xác định tập item thực sự liên quan (ground truth)
    Bỏ qua người dùng nếu không có item liên quan
    Tính số lần "trúng" (hits)
        Đếm số item chung giữa danh sách khuyến nghị và ground truth.
    Tính các chỉ số đánh giá
       Precision@5: số hits chia cho 5.
    Tính độ đa dạng (Diversity)
       Lấy đặc trưng của các item trong danh sách khuyến nghị.
       Tính ma trận tương đồng giữa các item này.
       Lấy giá trị trung bình của phần tam giác trên.
       Độ đa dạng được tính bằng: 1 - độ tương đồng trung bình.
    Lưu kết quả
       Thêm Precision và Diversity của người dùng vào Prec, Div.
    Tính giá trị trung bình cuối cùng
       Precision@5 = trung bình của tất cả giá trị trong Prec.
       Diversity@5 = trung bình của tất cả giá trị trong Div.
       Lấy mẫu ngẫu nhiên tối đa 1.000 bản ghi từ tập đánh giá R.
       Với mỗi bản ghi (u,i,r_ui) trong mẫu:
           So sánh rating thực tế r_ui với rating dự đoán R̂_H .
       Tính RMSE trên toàn bộ mẫu.
```

Thuật toán tính toán *Precision@5, Diversity@5* trên tập người dùng, kỹ thuật *MMR* (*Maximal Marginal Relevance*) dùng một hàm mục tiêu (*objective function*) để chọn tập gợi ý tối ưu, sao cho các item được chọn liên quan cao nhất tới người dùng nhưng không quá giống nhau, để tăng đa dạng danh sách gợi ý. *MMR* tìm ra item vừa phù hợp, vừa khác biệt so với những item đã được chọn trước đó đảm bảo cân bằng giữa liên quan và đa dạng.

6.2. Môi trường thực nghiệm

Mô hình chạy trên nền tảng Google Colab (Python 3.10). Thư viện: pandas, numpy, scikit-learn, surprise, matplotlib, seaborn.

6.3. Thiết lập dữ liệu

Dữ liệu được giả lập dựa trên đặc điểm thực tế của khoa Trí tuệ nhân tạo:

- Số sinh viên: 500
- Số tài liệu học tập: 10.000
- Số lượt đánh giá: 1.319.152 (thang điểm 1-5)
- 80% train: huấn luyện mô hình khuyến nghị.
- 20% test: đánh giá độ chính xác dự đoán.

6.4. Kết quả thực nghiệm

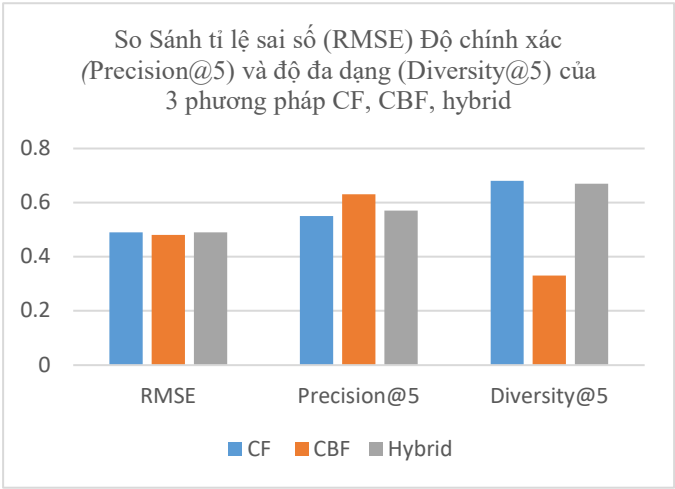
Bảng 1. Lựa chọn trọng số α cho mô hình hybrid

Trọng số α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	**0.8**	0.9
Precision@5	0.62	0.61	0.60	0.59	0.58	0.58	0.57	**0.57**	0.56
Diversity@5	0.41	0.45	0.50	0.56	0.61	0.64	0.66	**0.67**	0.67

Bảng 1 trình bày việc lựa chọn trọng số α trong mô hình hybrid. α=0.8 không phải tùy ý, mà được xác định thông qua thực nghiệm quét tham số (grid search) trên tập kiểm tra, với α thay đổi từ 0.1 đến 0.9. Kết quả cho thấy khi α tăng, *Diversity* cải thiện nhưng *Precision* giảm dần; ngược lại, khi α nhỏ, *Precision* cao nhưng *Diversity* thấp. Giá trị α=0.8 được chọn vì đạt sự cân bằng tốt nhất giữa *Precision* và *Diversity*.

Bảng 2. So sánh hiệu suất giữa 3 phương pháp CF, CBF và hybrid

Phương Pháp	***RMSE***	***Precision@5***	***Diversity@5***
CF	0.49	0.55	0.68
CBF	0.48	0.63	0.33
Hybrid	**0.49**	**0.57**	**0.67**



Hình 2. Hiệu suất của ba phương pháp CF, CBF và hybrid

Hình 2 cho thấy cả ba phương pháp đều cho tỉ lệ sai số thấp <= 0,49/5 tỉ lệ dưới 10%. *RMSE* của ba mô hình tương đương, chênh lệch không đáng kể (≤ 0.01). Mô hình lai đạt sự cân bằng tốt hơn giữa Precision (57%) và Diversity (67%), dù Precision thấp hơn CBF.

7. Thảo luận

7.1. Phân tích nguyên nhân kết quả

Kết quả thực nghiệm (Mục 6.4) chứng minh rằng mô hình lai đạt hiệu quả tổng hợp tốt nhất, giúp cân bằng giữa độ chính xác của CBF và đa dạng của CF trong khuyến nghị tài liệu. Nó giúp sinh viên nhận được khuyến nghị các tài liệu sát nhu cầu nhưng vẫn phong phú, đa dạng. Về mặt cơ chế, điều này phản ánh sự bổ sung lẫn nhau giữa hai hướng tiếp cận nền tảng. Cụ thể, CF thể hiện ưu thế trong việc khai thác cấu trúc hành vi tập thể, tận dụng tương quan giữa user để nhận diện các mẫu tiêu thụ tài liệu tiềm ẩn. Khi quy mô và mật độ dữ liệu đủ lớn, CF hình thành được không gian ngữ nghĩa ngầm phản ánh sở thích học tập của cộng đồng. Tuy nhiên, CF vốn phụ thuộc mạnh vào lượng tương tác, nên khi xuất hiện sinh viên hoặc tài liệu mới (vấn đề cold-start), khả năng dự đoán suy giảm rõ rệt. Ngược lại, CBF dựa vào đặc trưng nội dung và hồ sơ cá nhân hóa, cho phép khuyến nghị độc lập với hành vi của người khác, do đó khắc phục phần nào hạn chế cold-start. Dẫu vậy, CBF lại có xu hướng rơi vào hiện tượng khuyến nghị quá hẹp (*overspecialization*) chỉ lặp lại các lựa chọn tương tự, làm suy giảm độ đa dạng và khả năng khám phá. Mô hình lai dung hòa hai cực: sử dụng CF để khai thác tri thức cộng đồng và CBF để phân tích ngữ nghĩa nội dung, từ đó tạo nên một không gian gợi ý có tính tổng hợp cao hơn. Sự phối hợp này không chỉ giảm thiểu rủi ro cold-start mà còn cải thiện độ phủ nội dung, tăng tính đa dạng mà vẫn duy trì độ chính xác dự đoán một cân bằng mà các mô hình đơn lẻ khó đạt được.

7.2. Tác động thực tiễn

Việc tích hợp hệ thống khuyến nghị tài liệu học tập vào môi trường học tập trực tuyến (LMS) mở ra hướng tiếp cận mới trong quản lý và cá nhân hóa hoạt động học tập. Đối với sinh viên, hệ thống không chỉ giúp rút ngắn thời gian tìm kiếm tài liệu mà còn định hướng họ đến các nguồn học liệu phù hợp với năng lực, sở thích và tiến trình học tập, qua đó nâng cao mức độ gắn kết và động lực học. Đối với giảng viên, dữ liệu từ hệ thống khuyến nghị cho phép họ nắm bắt sâu hơn hành vi học tập của từng cá nhân, hỗ trợ việc điều chỉnh nội dung, phương pháp và tốc độ giảng dạy một cách linh hoạt, dựa trên nhu cầu thực tế của người học. Về phía quản lý đào tạo, việc áp dụng hệ thống này mang lại khả năng tối ưu hóa sử dụng tài nguyên học tập, đồng thời cung cấp các chỉ số định lượng về xu hướng và nhu cầu học tập, làm cơ sở cho việc hoạch định chiến lược phát triển chương trình và nâng cao chất lượng đào tạo trên quy mô toàn hệ thống.

Về mặt sư phạm, Precision cao giúp sinh viên nhận được nhiều tài liệu "đúng" hơn với nhu cầu hiện tại, giảm rủi ro lãng phí thời gian. Tuy nhiên, một hệ thống chỉ tối ưu Precision (như CBF) có xu hướng rơi vào hiện tượng overspecialization – chỉ khuyến nghị những tài liệu rất giống với những gì sinh viên đã xem trước đó. Điều này có thể làm thu hẹp phạm vi tiếp cận kiến thức, hạn chế khả năng khám phá (exploration), và khiến sinh viên bỏ lỡ những tài liệu hữu ích nhưng khác biệt về nội dung. Ngược lại, mô hình hybrid, với Precision thấp hơn một chút nhưng Diversity cao hơn đáng kể (67% so với 33% của CBF), tạo ra danh sách khuyến nghị cân bằng hơn giữa khai thác (exploitation) và khám phá (exploration). Điều này đặc biệt quan trọng trong giáo dục, nơi mục tiêu không chỉ là tối đa hóa mức độ phù hợp tức thời, mà còn mở rộng hiểu biết, kích thích tư duy đa chiều và hỗ trợ học tập dài hạn. Việc sinh viên tiếp cận một số tài liệu "không hoàn toàn tối ưu" có thể vẫn mang lại giá trị học tập nếu chúng giúp họ tiếp cận góc nhìn mới hoặc kết nối kiến thức liên ngành.

8. Kết luận và hướng phát triển

Nghiên cứu này giới thiệu một hệ thống khuyến nghị tài liệu học tập nhằm hỗ trợ sinh viên lựa chọn tài liệu học tập phù hợp và giúp giảng viên theo dõi tiến trình học tập hiệu quả hơn. Hệ thống được xây dựng trên nền tảng quy trình xử lý dữ liệu có kiểm soát và thử nghiệm có hệ thống hai hướng tiếp cận kinh điển CF và CBF. Kết quả cho thấy mô hình lai giữa *CF* và *CBF* đạt hiệu quả cân bằng tối ưu giữa độ chính xác và độ đa dạng thể hiện sự bổ trợ lẫn nhau giữa tri thức cộng đồng và đặc trưng nội dung. Đóng góp chính của nghiên cứu ở việc điều chỉnh và tích hợp hiệu quả các mô hình CF, CBF, và hybrid vào bối cảnh giáo dục đại học Việt Nam, chứng minh tính khả thi và hiệu quả cân bằng Precision, Diversity trong môi trường học trực tuyến. Dù đạt kết quả khả quan, nghiên cứu vẫn còn một số giới hạn: (1) chưa tích hợp yếu tố ngữ cảnh như tiến độ học tập, khung thời gian, hay mục tiêu cá nhân hóa của từng người học vốn có ảnh hưởng mạnh đến tính phù hợp của khuyến nghị; (2) chưa khai thác các mô hình học sâu (Deep Learning) với khả năng học biểu diễn ngữ nghĩa đa chiều, vốn đang chứng tỏ ưu thế vượt trội trong các hệ thống khuyến nghị hiện đại.

Trong tương lai, hệ thống có thể được mở rộng theo các hướng: (1) tích hợp ngữ cảnh học tập, xem xét yếu tố thời gian, tiến độ môn học và mục tiêu học tập để tạo khuyến nghị sát thực tế hơn; (2) áp dụng Deep Learning, sử dụng các mô hình hiện đại như Neural Collaborative Filtering, BERT4Rec hoặc Graph Neural Networks để cải thiện khả năng học đặc trưng và xử lý dữ liệu phức tạp; (3) khuyến nghị đa phương tiện, mở rộng khả năng khuyến nghị sang video, audio, tài liệu tương tác và khóa học trực tuyến; (4) đánh giá tác động lâu dài: đo lường tác động của hệ thống đến kết quả học tập và mức độ gắn bó của sinh viên trong nhiều học kỳ liên tiếp.

Việc triển khai thành công một hệ thống khuyến nghị hiệu quả trong giáo dục không chỉ giúp cá nhân hóa trải nghiệm học tập mà còn góp phần nâng cao hiệu quả sử dụng tài nguyên đào tạo và hỗ trợ ra quyết định cho giảng viên và nhà quản lý. Do dữ liệu thực nghiệm trong bài báo hiện tại là dữ liệu giả lập nhằm kiểm soát và đánh giá có hệ thống mô hình CF, CBF và hybrid, trong tương lai cần kế hoạch thử nghiệm trên dữ liệu thực tế từ hệ thống LMS của Đại học Duy Tân trong tương lai để kiểm chứng tính đúng đắn, khả năng mở rộng và hiệu quả thực tiễn của mô hình.

Tài liệu tham khảo

[1] Henriksen, E.K., Janssen, L.P., & Ha, J.M. (2023). "Recommender systems for teachers: A systematic literature review (2011–2023)". *Education Sciences* (13), 723.

[2] Wang, X., Wang, Z., Xu, G., et al. (2023). "ConceptGCN: Knowledge concept recommendation in MOOCs via knowledge graph and GCN". *Patterns* (4).

[3] CARS Workshop Committee. (2023). Workshop on context-aware recommender systems (CARS). *Proceedings of the RecSys Workshops*.

[4] Vector Institute. (2024). *A survey of recommender systems (2017–2024)* (Technical Report).

[5] Liu, Y., Chen, H., & Zhang, J. (2023). "Knowledge-aware sequence modelling for online course recommendation". *Information Processing & Management* (60).

[6] Zhang, J., Wu, H., & Li, X. (2024). "MOOCs video recommendation using low-rank and sparse matrix factorization". *Information Processing & Management*.

[7] Li, Y., Liang, Y., Yang, R., et al. (2024). "CourseKG: An educational knowledge graph based on course information for precision teaching". *Applied Sciences* (14), 2710.

[8] Koutrika, G., & Ioannidis, C. (2022). "Providing recommendations for communities of learners in MOOCs ecosystems". *Expert Systems with Applications* (205), 117620.

[9] Abdelhamid, S., El-Sappagh, A., & Kwak, K.-S. (2021). "Deep learning-based semantic personalized recommendation for e-learning". *Journal of King Saud University – Computer and Information Sciences* (33), 1189–1201.

[10] Zhao, X., Guo, Y., & Yang, Z. (2024). "An explainable graph-based course recommendation model based on dual graphs and rule-based reasoning". *Expert Systems with Applications*.

[11] Karimi, A., Yin, H., & Chen, L. (2024). "Review-based recommender systems: A survey of approaches and applications". *ACM Computing Surveys*.

[12] Al-Hossain, A.A., et al. (2022). "What is needed to build a personalized recommender system for K–12 education?". *Education and Information Technologies*.