ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA KHOA HỌC MÁY TÍNH

ĐỀ TÀI: HỆ THỐNG KHUYẾN NGHỊ KHÓA HỌC CHO NỀN TẢNG HỌC TẬP TRỰC TUYẾN

THUYẾT MINH ĐỀ TÀI

MÔN HỌC: KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG (CS313)

Nhóm 4

GVHD

ThS. Nguyễn Anh Thư

TP. HO CHI MINH, 6/2024

DANH SÁCH THÀNH VIÊN

STT	Họ và tên	MSSV
1	Đoàn Nhật Sang	21522542
2	Trương Văn Khải	21520274
3	Lê Ngô Minh Đức	21520195
4	Phạm Minh Quốc	22540017
5	Lê Yến Nhi	21522427
6	Hoàng Thị Mỹ Hạnh	21522044
7	Hoàng Tiến Đạt	21520696
8	Lê Minh Quang	21522510

PHÂN CÔNG CÔNG VIỆC

Chi tiết công việc	Trương Văn Khải (NT)	Đoàn Nhật Sang	Lê Ngô Minh Đức	Lê Minh Quang	Lê Yến Nhi	Hoàng Thị Mỹ Hạnh	Hoàng Tiến Đạt
Phân công	✓	✓					
Mục 1, 2, 3	✓						
Mục 4		✓					
Mục 5			✓				
Mục 6 - Nội dung 1				✓			
Mục 6 - Nội dung 2							✓
Mục 6 - Nội dung 3						✓	
Mục 7, 8					✓		
Format docs					✓	✓	✓
Kiểm tra	✓	✓	✓	✓	✓	✓	✓
Mức độ hoàn thành	100%	100%	100%	100%	100%	100%	100%

ĐÁNH GIÁ CỦA GIẢNG VIÊN HƯỚNG DẪN

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……...
……...
……...
……...
……...
……...
……...
……...
……...
……...
……...
……...
……...
……...
……...
……...
……...
……...
……...
……...
……...
……...
……...
……...

TP. HCM, ngày 6 tháng 6 năm 2024

GVHD

Nguyễn Anh Thư

MỤC LỤC

1 Tên đề tài, thời gian thực hiện, tổng kinh phí. 1
2 Nhóm thực hiện: chủ nhiệm, nhân lực 1
3 Mô tả đề tài: giới thiệu, ứng dụng, các dự án liên quan trong cùng lĩnh vực 1
3.1 Giới thiệu 1
3.2 Ứng dụng 2
3.3 Các dự án liên quan 3
4 Tổng quan: ý tưởng, tính cấp thiết, tính mới 4
4.1 Ý tưởng 4
4.2 Tính cấp thiết 5
4.3 Tính mới 5
5 Input – Output 6
6 Mục tiêu đề tài 6
6.1 Nội dung 1 6
6.2 Nội dung 2 7
6.3 Nội dung 3 7
6.4 Nội dung 4 8
7 Kết quả dự kiến, sản phẩm đề tài 9
8 TÀI LIỆU THAM KHẢO 10

1 Tên đề tài, thời gian thực hiện, tổng kinh phí.

- Tên đề tài: HỆ THỐNG KHUYẾN NGHỊ KHÓA HỌC CHO DỮ LIỆU MOOCCUBEX.
- Thời gian thực hiện: 8 tuần.
- Tổng kinh phí: 6.000.000 VND.

2 Nhóm thực hiện: chủ nhiệm, nhân lực

Chủ nhiệm:	Trương Văn Khải	21520274
Thành viên tham gia:	Đoàn Nhật Sang	21522542
	Lê Ngô Minh Đức	21520195
	Lê Yến Nhi	21522427
	Phạm Minh Quốc	22540017
	Hoàng Thị Mỹ Hạnh	21522044
	Hoàng Tiến Đạt	21520696
	Lê Minh Quang	21522510
Giảng viên hướng dẫn:	ThS. Nguyễn Thị Anh Thư	

3 Mô tả đề tài: giới thiệu, ứng dụng, các dự án liên quan trong cùng lĩnh vực

3.1 Giới thiệu

Khai phá dữ liệu và đặc biệt là dữ liệu lớn đang là lĩnh vực được các nhà khoa nghiên cứu quan tâm trong những năm gần đây. Ứng dụng của loại bài toán này khá đa dạng và phong phú và được thực hiện trong nhiều lĩnh vực khác nhau như: kinh doanh, giáo dục, y tế, tài chính, ngân hàng, … Đặc biệt trong những năm gần đây, khai phá dữ liệu hay đặc biệt là khai phá dữ liệu lớn trong lĩnh vực giáo dục đang là đối tượng đang rất được quan tâm nghiên cứu vì tính thiết thực của chúng. Đối với việc giáo dục trực tuyến hiện tại, người dùng hay người học hiện tại sẽ phải sự chủ động và tự giác cao, vì có rất nhiều môn học thuộc các rất nhiều các nhóm ngành học khác nhau

trong danh sách đào tạo. Người dùng sẽ phải phân bổ các môn học các môn học theo các nhóm chuyên ngành học cho từng thời điểm khác nhau để bổ sung hay tự cải thiện vốn kiến thức chuyên ngành cần có của mình. Trên thực tế, vì đây là các nền tảng học tập thực tế nên không có bất kỳ sự ràng buộc trực tiếp nào về mặt thời gian, điểm số nên thường xảy ra các trường hợp như: các môn học không được hoàn thành đúng hạn theo thời gian dự tính hoặc không bao giờ được hoàn thành vì người dùng đã bỏ ngang vì chán nản trong quá trình học tập. Chính vì vậy, công tác cố vấn trực tuyến trên các nền tảng học tập trực tuyến được đặt ra là một công việc quan trọng trong hình thức học tập theo kiểu mới này. Đây cũng là bài toán được sinh ra trong lĩnh vực khai phá dữ liệu trong khi có số lượng dữ liệu lớn về người học cũng như hành vi học tập của người học trong quá trình tham gia các nền tảng học tập trực tuyến nhằm trợ giúp cho việc cải thiện hiệu suất học tập hay gợi ý các môn học thuộc đúng chuyên ngành đang được quan tâm bởi người dùng.

Chính vì những lý do trên, nhóm chúng tôi chọn đề tài: **"Hệ thống khuyến nghị khóa học cho các nền tảng học tập trực tuyến"**. Nhóm chúng tôi hi vọng đồ án sẽ mang tính đóng góp thiết thực vào việc giải quyết các vấn đề mang tính cấp bách và thiết thực trong việc giáo dục trên các nền tảng học tập trực tuyến hiện tại.

3.2 Ứng dụng

Từ những lý do, câu hỏi đề tài và kết quả mong muốn thu được sau khi hoàn thiện đề tài, chúng tôi nhận thấy tính ứng dụng thực tiễn của đề tài là vô cùng lớn, có thể bao gồm:

- Cá nhân hóa trải nghiệm học tập của người dùng:
    - Tối ưu hóa lộ trình học tập: Hệ thống đề xuất các khóa học phù hợp với nhu cầu, sở thích, và mục tiêu của từng người học, giúp họ chọn được những khóa học phù hợp nhất.
    - Gợi ý dựa trên hành vi học tập: Dựa trên dữ liệu hành vi như các khóa học đã hoàn thành, thời gian học, và thành tích, hệ thống có thể đề xuất các khóa học tiếp theo một cách chính xác.

- Nâng cao hiệu suất học tập của người học:
    - Từ những hành vi học tập của người dùng trong quá khứ, hệ thống sẽ dựa và tự động đề xuất các khóa học tương thích nhất với khả năng và kỹ năng của người học để tối ưu hóa nhất hiệu suất học tập của người dùng.
- Tăng cường trải nghiệm và sự hài lòng của người học:
    - Bằng cách đề xuất các khóa học thú vị và phù hợp, hệ thống giữ người học gắn bó hơn với nền tảng, tăng thời gian và mức độ tham gia.
    - Các khóa học phù hợp và hấp dẫn có thể giúp giảm tỷ lệ người học từ bỏ giữa chừng, cải thiện tỷ lệ hoàn thành khóa học.

3.3 Các dự án liên quan

Vì đây là bài toán mang thiên hướng khuyến nghị, nhóm chúng tôi tập trung vào nghiên cứu tìm hiểu và nghiên cứu các bài toán khuyến nghị đã được công bố và đưa vào thực tiễn trước đó, bao gồm:

1. **MOOCCubeX: A Large Knowledge-centered Repository for Adaptive Learning in MOOCs** [1]: MOOCCubeX là một bộ dữ liệu được duy trì bởi nhóm Kỹ thuật Tri thức của Đại học Thanh Hoa và được hỗ trợ bởi XuetangX [2], một trong những trang web MOOC (Massive Open Online Courses) lớn nhất ở Trung Quốc. Kho lưu trữ này bao gồm 4.216 khóa học, 230.263 video, 358.265 bài tập, 637.572 khái niệm chi tiết và hơn 296 triệu dữ liệu hành vi thô của 3.330.294 sinh viên, để hỗ trợ các chủ đề nghiên cứu về học tập thích ứng trong MOOCs..
2. **KGAT: Knowledge Graph Attention Network for Recommendation** [3]: một phương pháp dựa trên GNN (Mạng Nơ-ron Đồ thị) sử dụng đồ thị tri thức nền (Knowledge Base) để cải thiện khuyến nghị. Phương pháp này được tái tạo dựa trên mạng co-occurrence network.
3. **Amazon.com Recommendations: Item-to-Item Collaborative Filtering** [4]**:** của Amazon năm 2003 giới thiệu một phương pháp cải tiến cho hệ thống

gợi ý sản phẩm, được gọi là lọc cộng tác dựa trên mặt hàng (Item-to-Item Collaborative Filtering).

4. **BPR: Bayesian Personalized Ranking from Implicit Feedback** [5]: một kỹ thuật học máy mới để giải quyết vấn đề xếp hạng cá nhân hóa từ phản hồi ngầm của người dùng (implicit feedback). Phản hồi ngầm không bao gồm xếp hạng rõ ràng mà là các hành vi của người dùng như xem, mua hoặc nhấp vào các sản phẩm.
5. **Fast Context-aware Recommendations with Factorization Machines** [6]: trình bày một phương pháp tiên tiến để cung cấp các khuyến nghị dựa trên ngữ cảnh nhanh chóng, sử dụng Máy phân tích nhân tố (Factorization Machines - FMs).
6. **Neural Factorization Machines for Sparse Predictive Analytics** [7]: giới thiệu một mô hình học máy tiên tiến kết hợp giữa Máy phân tích nhân tố (Factorization Machines - FMs) và Mạng nơ-ron (Neural Networks) để xử lý các bài toán phân tích dự đoán trên dữ liệu thưa.

Qua quá trình đọc và tìm hiểu từng bài toán, chúng tôi nhận thấy được điểm nổi bật của bài báo **KGAT: Knowledge Graph Attention Network for Recommendation** về cả mặt hiệu suất và cách thức tổ chức dữ liệu, nên đây cũng chính là hướng nghiên cứu chính trong đồ án của chúng tôi.

4 Tổng quan: ý tưởng, tính cấp thiết, tính mới

4.1 Ý tưởng

Ý tưởng chính của nhóm là khuyến nghị các nguồn tài nguyên học tập (khóa học) cho người dùng. Chúng tôi định nghĩa nhiệm vụ này là khuyến nghị các khóa học tiếp theo cho học viên dựa trên chuỗi các khóa học lịch sử của học viên. Nhiệm vụ này không chỉ yêu cầu mô hình hóa hành vi của học viên một cách tốt nhất mà còn cần xem xét vai trò của kiến thức được kèm theo trong các khóa học, cấu trúc của các khóa học,...

4.2 Tính cấp thiết

Với sự phổ biến ngày càng nhiều của các nền tảng học trực tuyến, việc xây dựng một hệ thống khuyến nghị khóa học hiệu quả dựa trên một bộ dữ liệu lớn dần trở thành một nhiệm vụ quan trọng và cấp thiết. Các nền tảng học tập trực tuyến này thường yêu cầu người học phải có sự chủ động và tự giác cao, vì có rất nhiều môn học thuộc các nhóm ngành đào tạo khác nhau trên nền tảng. Ngoài ra, thực tế thường thấy các nền tảng học tập này không có bất kỳ sự ràng buộc trực tiếp nào về mặt thời gian, điểm số nên thường xảy ra các trường hợp như: các môn học không được hoàn thành đúng hạn theo thời gian dự tính hoặc không bao giờ được hoàn thành vì người dùng đã bỏ ngang vì chán nản trong quá trình học tập.

Vì những lý do trên, hệ thống khuyến nghị học tập của chúng tôi được tạo ra và huấn luyện dựa trên các bộ dữ liệu lớn được thu thập thu từ các nền tảng học tập trực tuyến trong lĩnh vực giáo dục. Với mong muốn hệ thống này không chỉ giúp người dùng tìm kiếm và lựa chọn các nguồn học tập phù hợp với nhu cầu và mục tiêu cá nhân của họ một cách nhanh chóng và hiệu quả, mà còn cung cấp các gợi ý học tập cá nhân hóa dựa trên sở thích, kỹ năng và tiến trình học tập của mỗi người. Điều này giúp người dùng tối ưu hóa quá trình học tập, nâng cao hiệu suất và đạt được mục tiêu học tập của họ một cách hiệu quả nhất.

4.3 Tính mới

Dựa trên sự thành công và tiềm năng mà KGAT đã thể hiện trong lĩnh vực thương mại điện tử, nhóm đã áp dụng thành công nghiên cứu này để hoàn thiện một quy trình hoàn chỉnh trong việc khai phá dữ liệu lớn cho một bộ dữ liệu quan trọng trong lĩnh vực giáo dục, cũng như phát triển thành công một hệ thống khuyến nghị khóa học hiệu quả dựa trên KGAT.

Đồng thời, nhóm đã tiến hành thử nghiệm và cải tiến bằng cách sử dụng các thông tin bổ sung về mối quan hệ giữa Người dùng-Video bài học, Khóa học – Giảng viên,… để nâng cao hiệu suất dự đoán của hệ thống. Nhóm cũng đã hiện thực hóa các phương pháp khuyến nghị truyền thống và tiên tiến khác như FM, NFM, BPR và

Content-based filtering trên cùng bộ dữ liệu để có cơ sở so sánh và đánh giá hiệu suất giữa các phương pháp.

Hơn thế nữa, phần lớn quy trình thực hiện đã được triển khai hiệu quả trên một nền tảng dữ liệu lớn là MS Azure nhằm cải thiện hiệu suất công việc cũng như gia tăng khả năng mở rộng và phát triển đề tài trong tương lai. Cuối cùng, một web service với nhiều tính năng tương tác và giao diện người dùng thân thiện dựa trên mô hình KGAT, đã được nhóm phát triển nhằm thể hiện rõ ràng hơn tính thực tế của dự án mà nhóm hướng đến.

5 Input – Output

- **Input**: Nguồn dữ liệu lớn trong các nền tảng học tập trực tuyến: Thông tin người học, thông tin khóa học, hoạt động học tập của người dùng.
- **Output**: Đề xuất top k (10) các khóa học phù hợp nhất với người dùng.

6 Mục tiêu đề tài

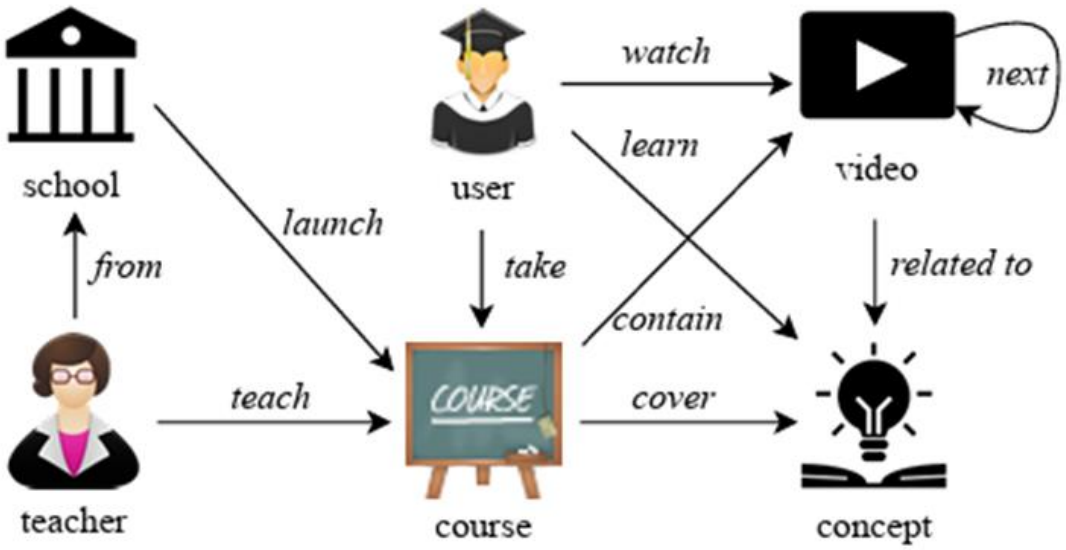
6.1 Nội dung 1

1. **Mục tiêu 1**: Tiền xử lý cho các bảng dữ liệu trong bộ dữ liệu MOOCCubeX phục vụ nhiệm vụ Khuyến nghị khóa học cho người dùng.
2. **Phương pháp 1**: Vì phục để dữ liệu phù hợp với phương pháp khuyến nghị mà chúng tôi thực hiện, chúng tôi phải tiến hành xử lý dữ liệu để tạo ra các bộ dữ liệu training, validation và testing cho mô hình.

   Đầu tiên chúng tôi sẽ phải loại bỏ các người dùng, items có tương tác ít hơn 10 lần, sau đó chúng tôi sẽ tạo một Knowledge Graph (KG) chứ các cặp bộ ba (head, relation, tail), tiếp theo sẽ phải tiến hành loại bỏ những thực thể có số lần xuất hiện ít hơn 10 lần, số quan hệ xuất hiện ít hơn 50 trong các cặp bộ ba. Cuối cùng chúng tôi sẽ thực hiện phân chia dữ liệu thành tập huấn luyện, tập xác thực và tập kiểm tra.
3. **Sản phẩm 1**: Thu được một bộ dữ liệu mới để tiến hành huấn luyện và đánh giá hiệu suất của mô hình.

6.2 Nội dung 2

1. **Mục tiêu 2**: Xây dựng mô hình học sâu với bộ dữ liệu MOOCCubeX với nghiệm vụ Khuyến nghị khóa học cho người dùng.
2. **Phương pháp 2**: Tập trung vào việc sử dụng mô hình KGAT [3] để dự đoán và khuyến nghị các khóa học có thể thu hút sự quan tâm của người dùng trên các nền tảng MOOCs. Điều này sẽ dựa trên lịch sử tương tác giữa người dùng và các khóa học (tức là, một người dùng đã tương tác với một khóa học nếu người dùng đã đăng ký khóa học đó). Kết quả trả về của mô hình sẽ là top-k khóa học được đề xuất cho người dùng.



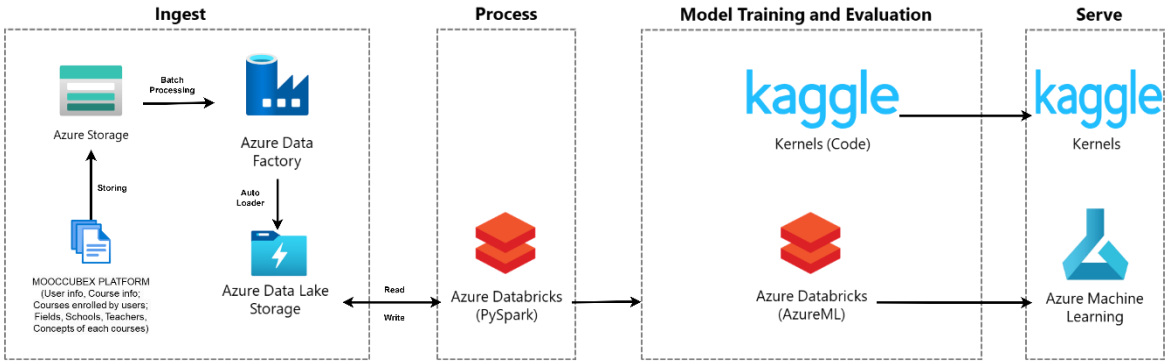
Hình 6.1 Hình minh họa các thức tổ chức dữ liệu theo phương pháp tiếp cận học sâu KGAT

3. **Sản phẩm 2**: pretrained mô hình KGAT được huấn luyện trên bộ dữ liệu MOOCCubeX [1] với nhiệm vụ Khuyến nghị khóa học.

6.3 Nội dung 3

1. **Mục tiêu 3**: Tìm hiểu và lựa chọn nền tảng đám mây phù hợp cho lưu trữ, xử lý dữ liệu lớn (Microsoft Azure) cũng như xây dựng và huấn luyện mô hình học máy (Microsoft Azure, Kaggle).

2. **Phương pháp 3**: Tận dụng các dịch vụ mà Microsoft Azure cung cấp cho việc lưu trữ và xử lý dữ liệu lớn: Azure Blob Storage, Azure Data Factory, Azure Data Lake Storage, Azure Databricks cùng các dịch vụ từ MS Azure và Kaggle cho quá trình xây dựng và huấn luyện mô hình học máy: Azure Machine Learning.



Hình 6.2 Hình minh họa các dịch vụ trên nền tảng đám mây phục vụ cho việc lưu trữ và xử lý dữ liệu lớn

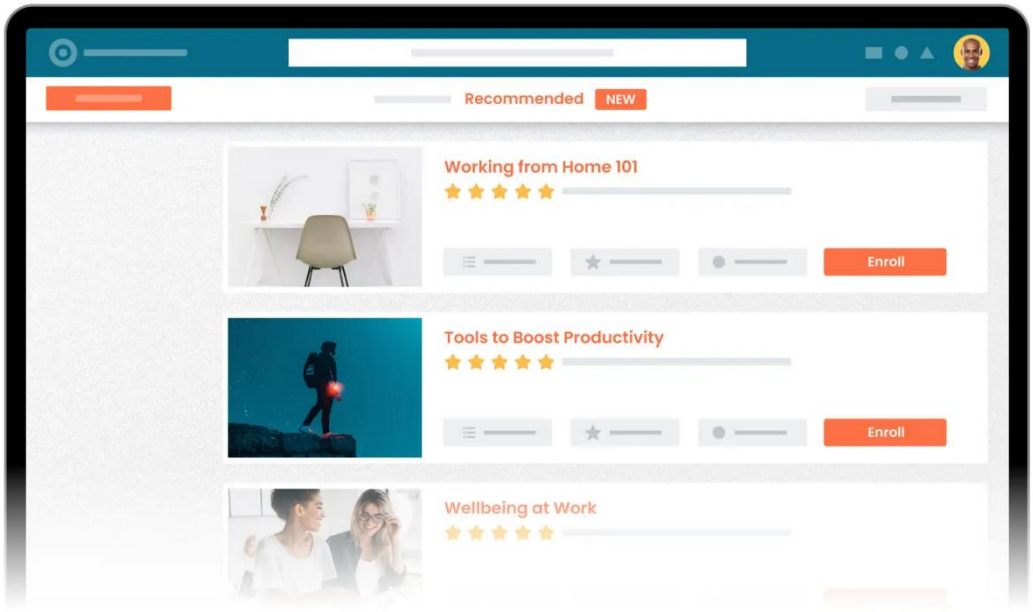
3. **Sản phẩm 3**: Các mô hình máy học được lưu trữ đầy đủ thông số sau quá trình thực nghiệm, phục vụ cho việc phát triển thành các ứng dụng thực tế cho nhiệm vụ khuyến nghị khóa học.

6.4 Nội dung 4

1. **Mục tiêu 4**: Xây dựng ứng dụng Website để phục vụ việc tương tác giữa người dùng và Hệ thống khuyến nghị.
2. **Phương pháp 4**: Sau khi thu được pretrained của mô hình KGAT được huấn luyện trên bộ dữ liệu MOOCCubeX trên nhiệm vụ Khuyến nghị các khóa học cho học viên. Sau đó sử dụng FastAPI (backend), NextJS (front-end) và SQL để triển khai website.
3. **Sản phẩm 4**: Thu được một ứng dụng Website trực quan, dễ dàng sử dụng, cho phép người dùng nhập vào các khóa học đã học, trả về tập gồm top-k các khóa học được hệ thống khuyến nghị.

7 Kết quả dự kiến, sản phẩm đề tài

Hệ thống khuyến nghị khóa học cho người dùng được chạy trên website. Hệ thống được mong đợi sẽ được chạy với kết quả có độ chính xác cao và tốc độ ổn định.



Hình 7.1 Hình minh họa kết quả sản phẩm bài toán của nhóm

8 TÀI LIỆU THAM KHẢO

[1] Yu, J.; Wang, Y.; Zhong, Q.; Luo, G.; Mao, Y.; Sun, K.; Feng, W.; Xu, W.; Cao, S.; Zeng, K.; et al., "MOOCCubeX: A Large Knowledge-centered Repository for Adaptive Learning in MOOCs," trong *In Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021.

[2] "XuetangX: Online Courses from Top Universities," Tsinghua University, [Trực tuyến]. Available: https://www.xuetangx.com/global. [Đã truy cập 28th May 2024].

[3] Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua, "KGAT: Knowledge graph attention network for recommendation," trong *KDD 2019*, 2019.

[4] G.Linden, B. Smith, and J. York, "Amazon.com recommendations: item-to-item collaborative filtering," *IEEE Internet Computing*, tập 7, số 1, pp. 76 - 80, 22 January 2003.

[5] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner,and Lars Schmidt-Thieme, "BPR: Bayesian Personalized Ranking from Implicit Feedback," trong *UAI*, 2009.

[6] Steffen Rendle, Zeno Gantner, Christoph Freudenthaler, and Lars Schmidt-Thieme, "Fast context-aware recommendations with factorization machines," trong *SIGIR*, 2011.

[7] Xiangnan He, Tat-Seng Chua, "Neural Factorization Machines for Sparse Predictive Analytics," trong *SIGIR*, 2017.