

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA KHOA HỌC MÁY TÍNH**

**A blue logo with a black background

Description automatically generated with low confidence**

**BÀI TẬP CHƯƠNG 4**

Giảng Viên: Nguyễn Thị Anh Thư

Mã Lớp: CS313.O21

Nhóm 04

| *Trương Văn Khải* | *21520274* |
| --- | --- |
| *Đoàn Nhật Sang* | *21522542* |
| *Lê Ngô Minh Đức* | *21520195* |
| *Lê Minh Quang* | *21522510* |
| *Lê Yến Nhi* | *21522427* |
| *Hoàng Thị Mỹ Hạnh* | *21522044* |
| *Hoàng Tiến Đạt* | *21520696* |

**TP. HỒ CHÍ MINH, 3/2024**

**NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN**

........................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................ ............................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................ .................................................................................................................................... .................................................................................................................................... .................................................................................................................................... .................................................................................................................................... ...................................................................................................................................

**BẢNG 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%** |

Bảng 0.1. Phân công công việc

**MỤC LỤC**

[1. Tên đề tài, thời gian thực hiện, tổng kinh phí. 1](#_1fob9te)

[2. Nhóm thực hiện: chủ nhiệm, nhân lực. 1](#_1fob9te)

[3. Mô tả đề tài: giới thiệu, ứng dụng, các dự án liên quan trong cùng lĩnh vực. 1](#_1fob9te)

[4. Tổng quan: ý tưởng, tính cấp thiết, tính mới. 6](#_v2sabddwptt7)

[5. Input-ouput: 7](#_ma6tqogrv586)

[6. Mục tiêu đề tài. 8](#_1fob9te)

[6.1. Nội dung 1 8](#_u7pv4pjxu79h)

[6.2. Nội dung 2 9](#_1fob9te)

[6.2. Nội dung 3. 10](#_1fob9te)

[6.3. Nội dung 4. 10](#_o5kzi7b93yax)

[7. Kết quả dự kiến, sản phẩm đề tài. 10](#_1fob9te)

[8. Tài liệu tham khảo. 11](#_l22de3xezqio)

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

* Tên đề tài: **HỆ THỐNG KHUYẾN NGHỊ KHÓA HỌC CHO BỘ DỮ LIỆU MOOCCUBEX**.
* Thời gian thực hiện: 8 tuần.
* Tổng kinh phí: Chưa biết.

# **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**

**Lê Minh Quang - 21522510**

**Hoàng Thị Mỹ Hạnh - 21522044**

**Hoàng Tiến Đạt - 21520696**

* 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.**

* **Giới thiệu**:

Các nền tảng học tập trực tuyến (MOOCs - Massive Open Online Courses) ngày nay càng dần được phổ biến, được xây dựng với quy mô ngày càng lớn và đúng một vai trò quan trong trong giáo dục cho học sinh. Để duy trì được sự quan tâm của người dùng đối với MOOCs, các hệ thống khuyến nghị đã được nghiên cứu và triển khai để khuyến nghị các khóa học, video hoặc các khái niệm trong bài học mà một người dùng có thể quan tâm.

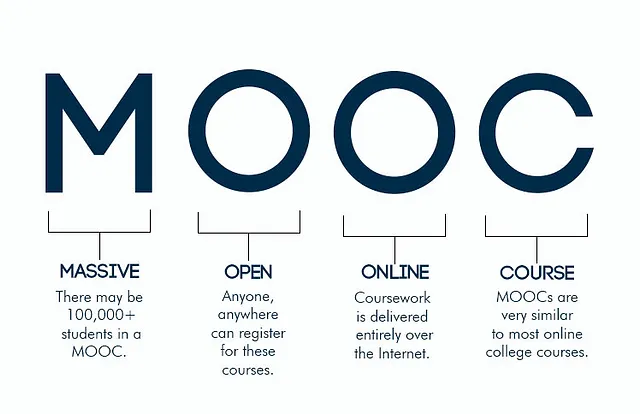
Ở bài này, chúng tôi sẽ tập trung vào nhiệm vụ khuyến nghị các khóa học mà người dùng có thể quan tâm, nhằm giảm được phần nào đó tỷ lệ bỏ các khóa học đăng ký hoặc thúc đẩy hiệu quả hiệu suất học tập của người dùng. Do đó để người đọc dễ tiếp cận, chúng tôi sẽ trình bày về những khái niệm liên quan trong bài toán như sau:

* **MOOC (Massive Open Online Course)**:

MOOC, viết tắt của Massive Open Online Course, là một hình thức giáo dục trực tuyến phổ biến trong thời đại số hóa ngày nay. Được xem là một biểu hiện tiên tiến của sự phát triển công nghệ, MOOC cung cấp cơ hội học tập không giới hạn cho mọi người trên khắp thế giới, bằng cách kết hợp các tài liệu giảng dạy, bài giảng video, bài tập và các hoạt động tương tác trực tuyến.

Với MOOC, học viên có thể tiếp cận vào các khóa học từ các trường đại học và tổ chức giáo dục hàng đầu trên toàn cầu, mà không cần phải di chuyển đến nơi học truyền thống. Điều này không chỉ giúp tiết kiệm thời gian và chi phí cho việc học, mà còn mở ra cánh cửa cho sự đa dạng hóa học vấn đề và phong phú hóa kiến thức.

MOOC cũng tạo điều kiện cho môi trường học tập linh hoạt, cho phép học viên tự điều chỉnh lịch trình học tập theo nhu cầu và tốc độ cá nhân. Đồng thời, qua các diễn đàn trực tuyến và các hoạt động giao tiếp khác, học viên còn có cơ hội chia sẻ ý kiến, kinh nghiệm và học hỏi từ nhau.

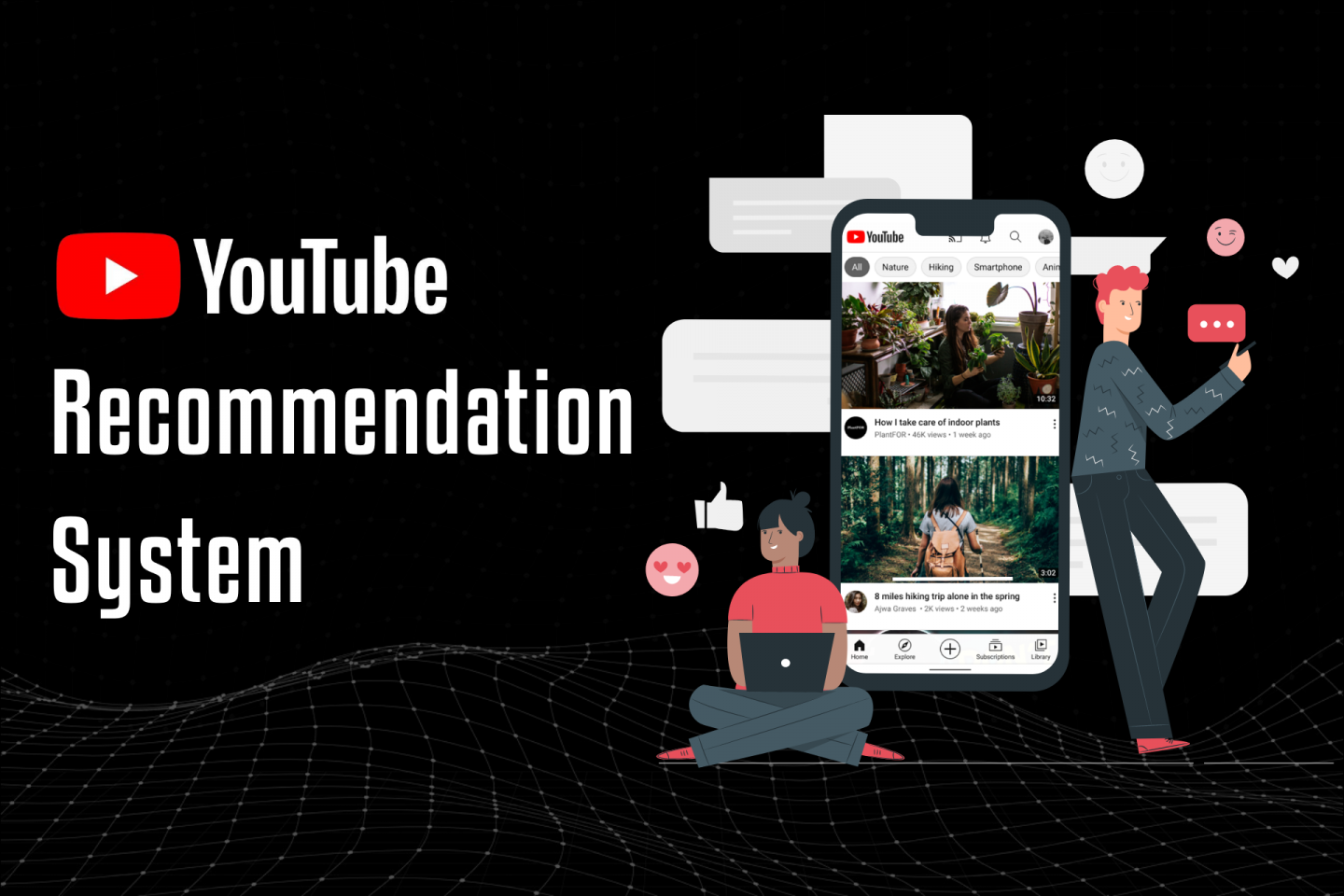


* **Hệ thống khuyến nghị (Recommendation systems)**:

Hệ thống khuyến nghị là một loại hệ thống thông tin được thiết kế để gợi ý hoặc đề xuất các mục tiêu, sản phẩm, hoặc nội dung phù hợp với sở thích, nhu cầu, hoặc hành vi của người dùng. Mục đích chính của hệ thống này là cải thiện trải nghiệm người dùng bằng cách cung cấp thông tin có giá trị và phù hợp.

Hệ thống khuyến nghị thường dựa trên dữ liệu thu thập được về người dùng, bao gồm lịch sử tìm kiếm, mua sắm, hoạt động trước đó, và các đánh giá phản hồi. Các thuật toán phân tích dữ liệu này để tạo ra các gợi ý cá nhân hóa, giúp người dùng khám phá và tiêu dùng sản phẩm, dịch vụ hoặc nội dung một cách hiệu quả hơn.

Hệ thống khuyến nghị được áp dụng rộng rãi trong nhiều lĩnh vực, từ thương mại điện tử, giải trí trực tuyến, đến ứng dụng tin tức và nội dung giáo dục. Các công nghệ mới như học máy và trí tuệ nhân tạo đã giúp cải thiện khả năng của các hệ thống này trong việc đề xuất các gợi ý chính xác và hấp dẫn cho người dùng.



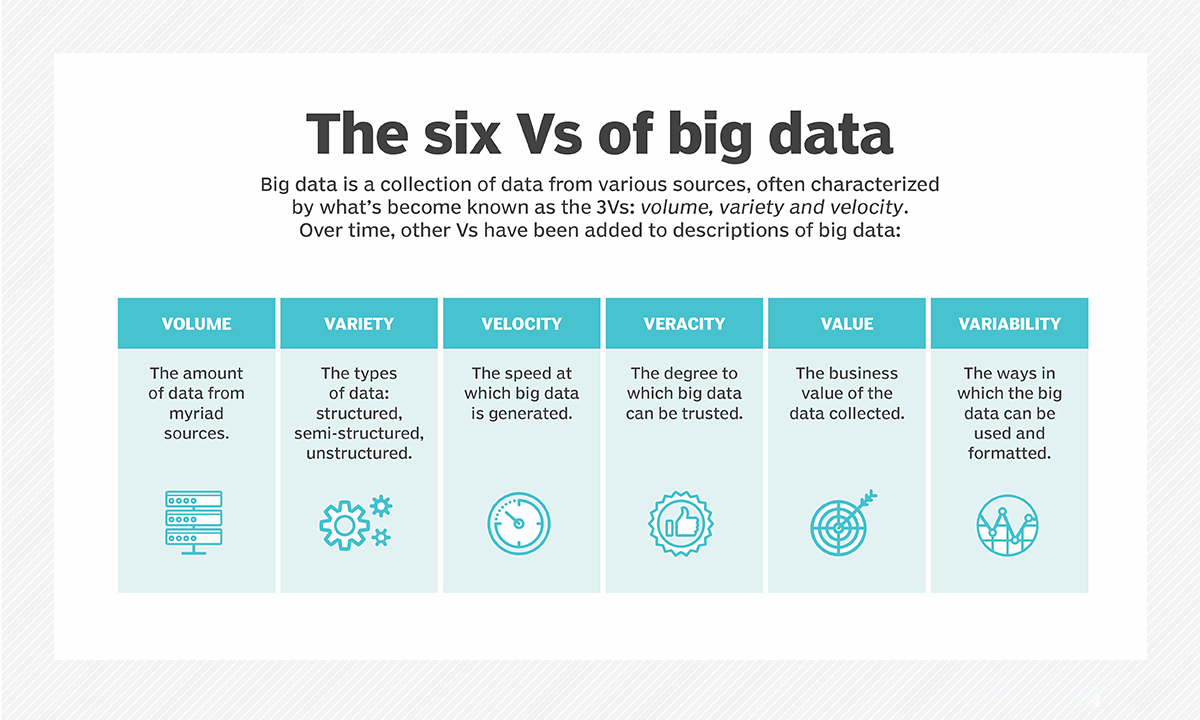
* **Dữ liệu lớn (Big data)**:

Dữ liệu lớn (hay còn gọi là Big Data) là một thuật ngữ chỉ đến việc xử lý, lưu trữ và phân tích các tập dữ liệu với khối lượng lớn và đa dạng, vượt quá khả năng xử lý của các công cụ và phương pháp truyền thống. Các tập dữ liệu này thường được tạo ra từ nhiều nguồn khác nhau, bao gồm dữ liệu từ máy chủ web, thiết bị cảm biến, mạng xã hội, hệ thống giao thông, và nhiều nguồn khác.

Dữ liệu lớn có sáu yếu tố chính được gọi là "6V":

* **Volume (Khối lượng)**: Đây là lượng dữ liệu được tạo ra và thu thập mỗi giây, mỗi phút hoặc mỗi ngày.
* **Velocity (Tốc độ)**: Đây là tốc độ mà dữ liệu được tạo ra và di chuyển qua các hệ thống.
* **Variety (Đa dạng)**: Đây là sự đa dạng của các loại dữ liệu, bao gồm dữ liệu có cấu trúc, bán cấu trúc và không có cấu trúc.
* **Veracity (Tính chắc chắn)**: Đây là sự tin cậy của dữ liệu, bao gồm tính chính xác, độ tin cậy và sự hoàn thiện của dữ liệu.
* **Value (Giá trị)**: Đây là khả năng trích xuất giá trị từ dữ liệu để hỗ trợ ra quyết định, tạo ra giá trị kinh doanh và cải thiện hiệu suất hoạt động.
* **Variability (Biến động)**: Đây là sự thay đổi trong dữ liệu theo thời gian, bao gồm sự biến đổi về cấu trúc, tần suất, và nguồn gốc của dữ liệu.

Dữ liệu lớn mang lại cơ hội lớn cho các doanh nghiệp và tổ chức trong việc tìm ra thông tin quan trọng, hiểu biết sâu sắc về khách hàng và hoạt động của họ, cũng như tạo ra giá trị kinh doanh mới thông qua việc phân tích và ứng dụng dữ liệu.



* **Ứng dụng**:

Xây dựng hệ thống khuyến nghị các nguồn tài nguyên học tập. Chúng tôi thực hiện nhiệm vụ này trên MOOCCubeX để đánh giá khả năng của tập dữ liệu này và thảo luận về cách nó có thể được sử dụng để tiến hành các nghiên cứu liên quan hoặc mở rộng trên các bộ dữ liệu MOOCs khác.

Sau đó có thể sử dụng các bộ dữ liệu học tập của các trường Đại học ở Việt Nam để xây dựng các ứng dụng có liên đến “Cố vấn học tập thông minh tại các trường đại học” hay “Cố vấn học tập thông minh cho các nền tảng học tập trực tuyến”.

* **Các dự án liên quan**:

1. **KGAT: Knowledge Graph Attention Network for Recommendation**[1]: 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.
2. **Truelearn: A family of bayesian algorithms to match lifelong learners to open educational resources** [2]: một phương pháp khuyến nghị giáo dục sử dụng một họ của Mạng Bayesian.
3. **BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer** [3]: một mô hình khuyến nghị tiền huấn luyện điển hình dựa trên kiến trúc Transformer song hướng như BERT.
4. **KB4Rec: A Data Set for Linking Knowledge Bases with Recommender Systems** [4]: Đề xuất phương pháp tổ chức dữ liệu dưới dạng một Đồ thị tri thức (Knowledge Base - KB). Sau đó liên kết tập dữ liệu giữa hệ thống khuyến nghị (RS) với KB.

# **4. Tổng quan: ý tưởng, tính cấp thiết, tính mới.**

* **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, v.v.
* **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. Đồng thời, MOOCCubeX [5] chứa đựng một lượng lớn thông tin về hành vi học tập, sở thích và nhu cầu của người dùng, cung cấp một nguồn dữ liệu quý giá để phân tích và hiểu biết sâu rộng về các yếu tố ảnh hưởng đến việc học trực tuyến.

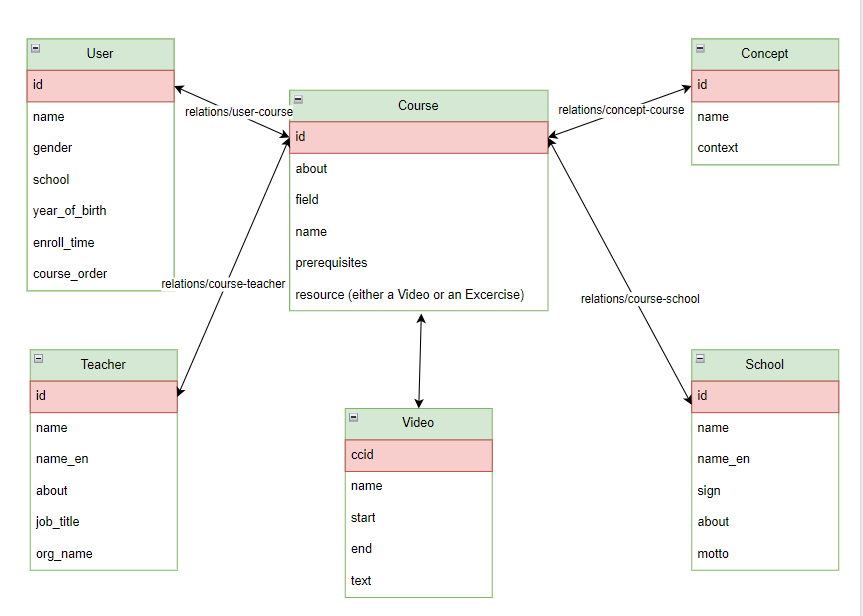
Vì những lý do trên, hệ thống khuyến nghị học tập của chúng tôi được tạo ra và huấn luyện dựa trên bộ dữ liệu MOOCCubex [5], nó không chỉ giúp người dùng tìm kiếm và lựa chọn các nguồn học tập phù hợp với nhu cầu và mục tiêu cá nhân của họ một cách nhanh chóng và hiệu quả, mà còn cung cấp các gợi ý học 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.

* **Tính mới:** Trong quá trình tìm hiểu, chúng tôi nhận thấy chưa có một dự án nghiên cứu nào thực hiện việc khuyến nghị các khóa học dựa trên nguồn tài nguyên MOOCCubeX [5]. Ngoài ra, đề xuất một phương pháp tiền xử lý dữ liệu, một mô hình học sâu để phục vụ nhiệm vụ khuyến nghị trên. Ngoài ra, xây dựng Website phục vụ mục đích sử dụng của người dùng với những trải nghiệm mới như: tính năng tương tác, cho phép người dùng chia sẻ, đánh giá và gợi ý nguồn học tập với cộng đồng.

# **5. Input-ouput:**

* *Input*:
* Danh sách khóa học (**Course**) mà sinh viên (**User**) đăng kí
* Danh sách video mà sinh viên xem, cùng với thời lượng xem từng video
* Mỗi khóa học cần có:
* Danh sách tài nguyên (**Video**) thuộc về khóa học
* Danh sách giảng viên (**Teacher**), trường (**School**) tổ chức khóa học
* Danh sách khái niệm (**Concept**) của khóa học

Để dễ hình dung, các thực thể (User, Course, Teacher, School, Concept, Video) trong tập MOOCubeX mà nhóm sử dụng có quan hệ như sơ đồ:



Các file **relations** là các file được cung cấp sẵn trong MOOCubeX để liên kết khóa chính & khóa ngoại của hai thực thể lại với nhau (có dạng id\_entity\_1 id\_entity\_2). Dễ thấy ta có 4 cặp relations, **suy ra đồ thị KGAT sẽ có tổng cộng 2\*4=8 relation khác nhau.**

* *Output*:
* Top 20 khóa học phù hợp với từng sinh viên
* Top 20 khóa học trending
* Các khóa học mới

# **6. Mục tiêu đề tài.**

## **6.1. Nội dung 1**

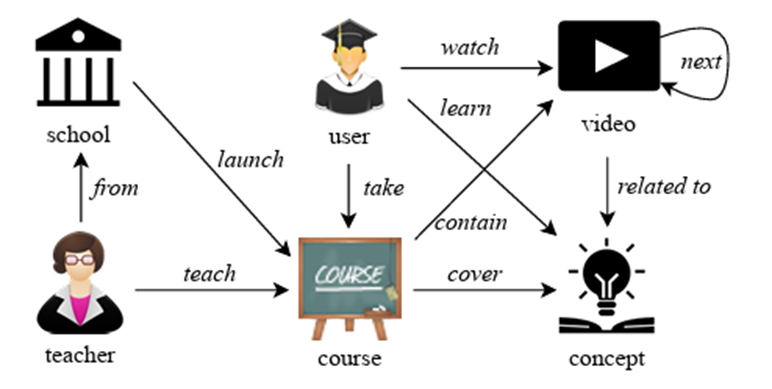
1. **Mục tiêu 1:** Tiền xử lý cho các bảng dữ liệu trong bộ dữ liệu MOOCCubeX [5] phục vụ nhiệm vụ Khuyến nghị khóa học cho người dùng.
2. **Phương pháp 1:** Vì phục để dữ liệu phù hợp với phương pháp khuyến nghị mà chúng tôi thực hiện, chúng tôi phải tiến hành xử lý dữ liệu để tạo ra các bộ dữ liệu 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.

1. **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 12:** Xây dựng mô hình học sâu với bộ dữ liệu MOOCCubeX [5] với nghiệm vụ Khuyến nghị khóa học cho người dùng.
2. **Phương pháp 12**: Tập trung vào việc sử dụng mô hình KGAT[1] 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 vẽ biểu diễn sự đa dạng các mối quan hệ trong bộ dữ liệu MOOCCubeX [5]. Trong đó, một user được coi là quan tâm đến một khóa học nếu user đã đăng ký hoặc có khả năng sẽ đăng ký khóa học đó trong tương lai.

1. **Sản phẩm 2:** pretrained mô hình KGAT được huấn luyện trên bộ dữ liệu MOOCCubeX [5] với nhiệm vụ Khuyến nghị khóa học.

## **6.2. Nội dung 3.**

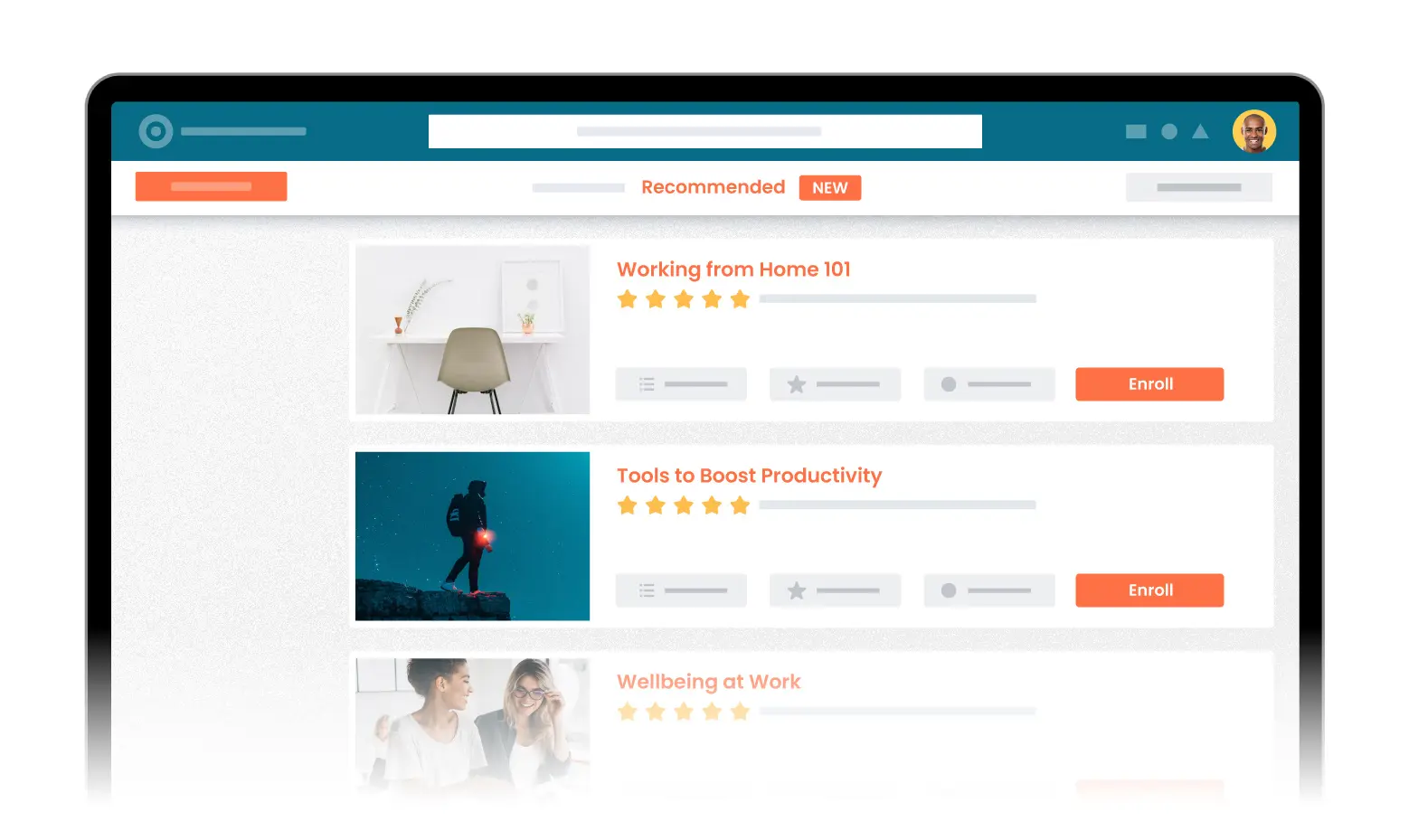
1. **Mục tiêu 3:** Tìm hiểu các hệ thống lưu trữ dữ liệu phù hợp (Microsoft Azure), sau đó tiến hành triển khai lưu trữ bộ dữ liệu của nhóm trên nền tảng này.
2. **Phương pháp 3:** Tận dụng các dịch vụ mà Microsoft Azure cung cấp, cụ thể: **Ingest data** (Nhập dữ liệu) và **Storage services** (dịch vụ lưu trữ).
3. **Sản phẩm 3:** Thu được một **Azure Tables** dùng để lưu trữ các dữ liệu dưới dạng SQL.

## **6.3. 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 [1] được huấn luyện trên bộ dữ liệu MOOCCubeX [5] trên nhiệm vụ Khuyến nghị các khóa học cho học viên. Sau đó sử dụng FastAPI (backend) và ReactJS (front-end) 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.



# **8. Tài liệu tham khảo.**

**[1]**: Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. 2019. Kgat: Knowledge graph attention network for recommendation. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 950–958.

**[2]**: Sahan Bulathwela, Maria Perez-Ortiz, Emine Yilmaz, and John Shawe-Taylor. 2020. Truelearn: A family of Bayesian algorithms to match lifelong learners to open educational resources. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34. 565–573.

**[3]**: Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In Proceedings of the 28th ACM international conference on information and knowledge management. 1441–1450.

**[4]**: WayneXinZhao, Gaole He, Hong-Jian Dou, Jin Huang, Siqi Ouyang, and Ji-Rong Wen. 2018. KB4Rec: A Dataset for Linking Knowledge Bases with Recommender Systems. CoRR abs/1807.11141 (2018).

**[5]**: Jifan Yu, Yuquan Wang, Qingyang Zhong, Gan Luo, Yiming Mao, Kai Sun, Wen zheng Feng, Wei Xu, Shulin Cao, Kaisheng Zeng, et al. 2021. MOOCCubeX: a large knowledge-centered repository for adaptive learning in MOOCs. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 4643–4652.