ChatGPT

Tổng quan bộ dữ liệu MOOCCubeX

Bộ dữ liệu **MOOCCubeX** là một kho dữ liệu lớn được nhóm nghiên cứu Kiến trúc tri thức trường ĐH Thanh Hoa (THU-KEG) phát triển, hướng đến hỗ trợ nghiên cứu học tập thích nghi trên các khóa học trực tuyến mở (MOOCs). Kho dữ liệu bao gồm thông tin về 4.216 khóa học, 230.263 video bài giảng, 358.265 bài tập, 637.572 khái niệm chi tiết và hơn 296 triệu bản ghi hành vi của 3.330.294 sinh viên [1]. Dữ liệu này được XuetangX cấp phép và cung cấp (có giấy phép GPL-3.0), bao gồm các nguồn liên quan đến khóa học, thông tin giảng viên, trường học, cũng như các thông tin chi tiết về hành vi người dùng (xem khóa học, làm bài tập, bình luận, v.v.) và các khái niệm thu thập từ phụ đề video.

Cấu trúc thư mục

MOOCCubeX được tổ chức thành các thư mục chính sau:

- **docs/**: Chứa tài liệu mô tả định dạng và ý nghĩa của các tệp dữ liệu (bằng tiếng Anh và tiếng Trung), ví dụ `course-en.md`, `user-en.md`, `concept-en.md`...
- **scripts/**: Chứa các kịch bản hỗ trợ (shell, Python) cho phép tải toàn bộ dữ liệu (`download_dataset.sh`), đếm số lượng mẫu (`count.sh`), tìm kiếm thông tin, v.v.
- **entities/**: Chứa các tệp JSON lưu trữ thông tin thực thể (entities) của khóa học, video, bài tập, trường học, giáo viên, người dùng, bình luận, trả lời, khái niệm, v.v. (các tệp `.json` rất lớn có dung lượng từ vài trăm KB đến hàng GB).
- **relations/**: Chứa các tệp quan hệ (thường là file `.txt` hoặc `.json`) biểu diễn liên kết giữa các thực thể như liên kết giữa bài tập và câu hỏi (`exercise-problem.txt`), quan hệ giữa khoá học với lĩnh vực (`course-field.json`), theo dõi video của người dùng (`user-video.json`), quan hệ giữa người dùng và bài tập (`user-problem.json`), v.v. Ngoài ra còn có các liên kết khái niệm đến khóa học/video/bài toán/bình luận/nguồn tài nguyên ngoài (`relations/concept-*.txt`).
- **prerequisites/**: Chứa các tệp JSON mô tả quan hệ tiền đề giữa các khái niệm trong một số lĩnh vực (máy tính, toán, tâm lý học).

Các tệp dữ liệu này có dung lượng rất lớn (ví dụ tệp `entities/problem.json` ~1,2GB, `relations/user-problem.json` ~21GB, `entities/comment.json` ~2,1GB [2] [3]), do đó khi sử dụng cần lưu ý về bộ nhớ. Mọi tệp đều ở định dạng văn bản (JSON hoặc tab-cách nhau `.txt`) và có cấu trúc rõ ràng theo mô tả bên dưới.

Chi tiết các tệp dữ liệu

Thư mục `entities/` (thực thể)

- `entities/course.json` (JSON, ~43MB): Thông tin về mỗi khóa học. Mỗi mục con là một đối tượng JSON chứa các trường chính:
- `id`: Mã khoá học (ví dụ `C_123456`).
- `name`: Tên khoá học.

- `about`: Giới thiệu ngắn về khoá học.
- `field`: Danh sách lĩnh vực thuộc về khoá học.
- `prerequisites`: Mô tả về kiến thức tiên quyết cho khoá học.
- `resource`: Danh sách tài nguyên của khoá học (video hoặc nhóm bài tập). Mỗi tài nguyên là một đối tượng với các trường: `resource_type` (`video` hoặc `exercise`), `resource_id` (ID của video hoặc nhóm bài tập), `chapter` (số chương), `titles` (danh sách tiêu đề các cấp độ của chương/video).
  Đây chính là phần **Course Info** trong dữ liệu (tổ chức video và bài tập của khoá học) [4] [5].

Trường (field)	Ý nghĩa
`id`	Mã khoá học (định dạng `C_xxxxxx`)
`name`	Tên khoá học
`about`	Giới thiệu khái quát về khoá học
`field`	Lĩnh vực (được gắn thẻ) của khoá học
`prerequisites`	Mô tả kiến thức tiên quyết
`resource`	Danh sách tài nguyên thuộc khoá học (xem chú thích trên)

- `entities/video.json` (JSON, ~580MB): Thông tin chi tiết về từng video bài giảng. Mỗi đối tượng JSON gồm các trường:
- `ccid`: Mã duy nhất của video (XuetangX gọi là `ccid`).
- `name`: Tiêu đề video.
- `start`: Mảng các thời gian bắt đầu của từng câu phụ đề trong video.
- `end`: Mảng thời gian kết thúc tương ứng.
- `text`: Mảng các câu trong phụ đề video.
  Đây là phần **Video** (tên video và phụ đề) của dữ liệu [6] [7].

Trường (field)	Ý nghĩa
`ccid`	Mã video duy nhất
`name`	Tên video
`start`	Danh sách thời điểm bắt đầu mỗi câu phụ đề (giây)
`end`	Thời điểm kết thúc mỗi câu phụ đề
`text`	Nội dung từng câu phụ đề (text theo thứ tự)

- `relations/exercise-problem.txt` (TXT, ~129MB): Mỗi dòng ánh xạ giữa một **nhóm bài tập** và **câu hỏi** (bài toán) của nó. Mỗi dòng định dạng `{exercise_id}\t{problem_id}` (ví dụ `Ex_123\tPm_456`). Đây là mối liên kết giữa *exercise* và *problem* (các bài tập của khoá học) [8].

- `entities/problem.json` (JSON, ~1.2GB): Nội dung chi tiết của mỗi câu hỏi (bài toán) trong các bài tập. Các trường gồm:

- `id`: Mã câu hỏi (`Pm_xxxxx`).
- `exercise_id`: Mã nhóm bài tập chứa câu hỏi (ví dụ `Ex_xxxxx`).
- `language`: Ngôn ngữ bài toán (Tiếng Trung hoặc Anh).
- `title`: Tiêu đề bài tập (mã nhóm).
- `content`: Nội dung đề bài.
- `option`: Đáp án lựa chọn (nếu có).
- `answer`: Đáp án đúng.
- `score`: Điểm của câu hỏi.
- `type` / `typetext`: Loại câu hỏi (ví dụ điền khuyết, trắc nghiệm, v.v.).
- `location`: Vị trí chương của câu hỏi.
- `context_id`: Mảng các *leaf_id* của khái niệm liên quan đến câu hỏi.

Đây là phần **Problem** (nội dung câu hỏi) của khoá học [9] [10].

Trường (field)	Ý nghĩa
`id`	Mã câu hỏi (`Pm_xxxxx`)
`exercise_id`	Mã nhóm bài tập chứa câu hỏi (`Ex_xxxxx`)
`language`	Ngôn ngữ của đề bài (Zh/En)
`title`	Tiêu đề của bài tập
`content`	Nội dung đề bài của câu hỏi
`option`	Các lựa chọn trả lời (cho câu hỏi nhiều lựa chọn)
`answer`	Đáp án đúng
`score`	Số điểm của câu hỏi
`type`	Loại câu hỏi (ví dụ `single choice`, `fill in`, v.v.)
`typetext`	Mô tả kiểu câu hỏi (đầy đủ dạng văn bản)
`location`	Chương (chapter) của câu hỏi
`context_id`	Mảng các **leaf_id** của khái niệm liên quan (fine-grained)

- `entities/school.json` (JSON, ~613KB): Thông tin các trường đại học. Các trường:
- `id`: Mã trường (ví dụ `S_1234`).
- `name`: Tên tiếng Trung của trường.
- `name_en`: Tên tiếng Anh.
- `sign`: Viết tắt tiếng Anh.
- `about`: Giới thiệu về trường.
- `motto`: Khẩu hiệu của trường.

Trường (field)	Ý nghĩa
`id`	Mã trường (`S_xxxxxx`)
`name`	Tên tiếng Trung của trường

Trường (field)	Ý nghĩa
`name_en`	Tên tiếng Anh của trường
`sign`	Viết tắt (ký hiệu) tên tiếng Anh của trường
`about`	Giới thiệu chung về trường
`motto`	Khẩu hiệu của trường

- `entities/teacher.json` (JSON, ~8.7MB): Thông tin giảng viên. Các trường:
- `id`: Mã giảng viên (`T_xxxxxx`).
- `name`: Tên tiếng Trung.
- `name_en`: Tên tiếng Anh.
- `about`: Tiểu sử hoặc thông tin cá nhân ngắn gọn.
- `job_title`: Chức danh, ví dụ giáo sư, tiến sĩ.
- `org_name`: Tổ chức công tác (trường hoặc viện).

Trường (field)	Ý nghĩa
`id`	Mã giảng viên (`T_xxxxxx`)
`name`	Tên tiếng Trung của giảng viên
`name_en`	Tên tiếng Anh của giảng viên
`about`	Thông tin giới thiệu ngắn (tiểu sử)
`job_title`	Chức danh (như Giáo sư, Phó Giáo sư...)
`org_name`	Nơi công tác (trường, viện...)

- `entities/user.json` (JSON, ~770MB): Hồ sơ người dùng (sinh viên) đăng ký trên nền tảng. Các trường:
- `id`: Mã người dùng (`U_xxxxxx`).
- `name`: Tên người dùng (không bắt buộc duyệt ra).
- `gender`: Giới tính.
- `school`: Trường học người dùng đang học.
- `year_of_birth`: Năm sinh (năm và tháng).
- `course_order`: Mảng các `course_id` (mã khoá học) đã đăng ký.
- `enroll_time`: Mảng thời gian tương ứng khi đăng ký từng khoá học (tương ứng với `course_order`).

Trường (field)	Ý nghĩa
`id`	Mã người dùng (`U_xxxxxx`)
`name`	Tên người dùng
`gender`	Giới tính (nếu biết)

Trường (field)	Ý nghĩa
`school`	Trường học của người dùng
`year_of_birth`	Năm (và tháng) sinh
`course_order`	Danh sách khoá học đã đăng ký
`enroll_time`	Thời gian đăng ký (tương ứng với `course_order`)

- `entities/comment.json` (JSON, ~2.1GB): Thông tin bình luận của người dùng trên các tài nguyên (video hoặc bài tập). Các trường:
- `id`: Mã bình luận (`Cm_xxxxxx`).
- `user_id`: Mã người dùng (`U_xxxxxx`) đã viết bình luận.
- `text`: Nội dung bình luận.
- `create_time`: Thời gian tạo bình luận.

Mỗi bản ghi mô tả một bình luận của người dùng trên một video hoặc bài tập [11].

Trường (field)	Ý nghĩa
`id`	Mã bình luận (`Cm_xxxxxx`)
`user_id`	Mã người dùng viết bình luận (`U_xxxxxx`)
`text`	Nội dung bình luận
`create_time`	Thời điểm viết bình luận

- `entities/reply.json` (JSON, ~50MB): Thông tin trả lời bình luận của người dùng. Các trường:
- `id`: Mã trả lời (`Rp_xxxxxx`).
- `user_id`: Mã người dùng trả lời (`U_xxxxxx`).
- `text`: Nội dung trả lời.
- `create_time`: Thời gian tạo trả lời.

Trường (field)	Ý nghĩa
`id`	Mã trả lời (`Rp_xxxxxx`)
`user_id`	Mã người dùng trả lời (`U_xxxxxx`)
`text`	Nội dung trả lời
`create_time`	Thời điểm tạo trả lời

- `entities/concept.json` (JSON, ~156MB): Tập hợp **khái niệm** (concept) thu được từ phụ đề video. Mỗi khái niệm là một đối tượng với các trường:
- `id`: Mã khái niệm, định dạng `K_{tên khái niệm}_{field}` (ví dụ `K_矩阵变换_数学` cho khái niệm "Ma trận chuyển vị" thuộc lĩnh vực Toán) [12].
- `name`: Tên khái niệm (trùng tên với phần tên trong `id`).
- `context`: Đoạn văn xung quanh xuất hiện khái niệm (trích từ Wikipedia, Baidu Baike, Zhihu, khoảng 50 ký tự trước và sau) [12].

Trường (field)	Ý nghĩa
`id`	Mã khái niệm (định dạng `K_tênKháiNiệm_lĩnhVực`)
`name`	Tên khái niệm
`context`	Ngữ cảnh xung quanh khái niệm (trích từ Wiki/Baike/Zhihu)

Thư mục `relations/` (quan hệ)

- `relations/course-field.json` (JSON, ~62KB): Mối quan hệ giữa khoá học và lĩnh vực được gắn thẻ. Mỗi đối tượng gồm:
- `course_id`: Mã khoá học.
- `course_name`: Tên khoá học.
- `field`: Danh sách lĩnh vực (lĩnh vực học thuật) mà khoá học thuộc về (ghi chú thủ công) [13].

Trường (field)	Ý nghĩa
`course_id`	Mã khoá học
`course_name`	Tên khoá học
`field`	Danh sách lĩnh vực của khoá học

- `relations/course-school.txt`: Liên kết giữa khoá học và trường đại học tổ chức. Mỗi dòng `{course_id}\t{school_id}` (ví dụ `C_12345\tS_6789`) thể hiện khoá học được cung cấp bởi trường nào [14].

- `relations/course-teacher.txt`: Liên kết giữa khoá học và giảng viên. Mỗi dòng `{course_id}\t{teacher_id}` (ví dụ `C_12345\tT_9876`) thể hiện giảng viên chính của khoá học [14].

- `relations/video_id-ccid.txt`: Ánh xạ giữa **Video ID** và **ccid**. Mỗi dòng `{video_id}\t{ccid}` liên kết ID của từng video (với tiền tố `V_`) với mã ccid của nó [15].

- `relations/exercise-problem.txt` (TXT): Liên kết giữa **nhóm bài tập** và **câu hỏi** (bài toán) của nó. Mỗi dòng `{exercise_id}\t{problem_id}` (ví dụ `Ex_123\tPm_456`) [8].

- `relations/user-video.json` (JSON, ~3.0GB): Dữ liệu hành vi của người dùng khi xem video. Mỗi đối tượng gồm:

- `user_id`: Mã người dùng (`U_xxxxxx`).

- `seq`: Mảng chứa trình tự các phiên xem video của người dùng. Mỗi phần tử của mảng này là một đối tượng bao gồm thông tin như `ccid`, thời gian xem, thời gian bắt đầu, thời gian kết thúc, tốc độ xem... trong một phiên xem video nhất định [16].

- `relations/user-problem.json` (JSON, ~21GB): Dữ liệu hành vi của người dùng khi làm bài tập. Các trường:

- `log_id`: Mã bản ghi (kết hợp `user_id` và `problem_id`).
- `user_id`: Mã người dùng (`U_xxxxxx`).
- `problem_id`: Mã câu hỏi (`Pm_xxxxxx`).
- `is_correct`: Kết quả làm đúng (1) hay sai (0).
- `attempts`: Số lần thử.
- `score`: Điểm đạt được.
- `submit_time`: Thời điểm nộp bài [17].

- `relations/user-xiaomu.json` (JSON, ~9.7MB): Hành vi tương tác của người dùng với Xiaomu (hệ thống hỏi đáp của XuetangX). Các trường:

- `user_id`: Mã người dùng (`U_xxxxxx`).
- `question_type`: Loại câu hỏi mà người dùng hỏi.
- `question`: Nội dung câu hỏi người dùng gửi [18].

- `relations/course-comment.txt`: Liên kết giữa khoá học và bình luận (review). Mỗi dòng `{course_id}\t{comment_id}` (ví dụ `C_12345\tCm_67890`) thể hiện một bình luận thuộc về khoá học nào [19].

- `relations/user-comment.txt`: Liên kết giữa người dùng và bình luận. Mỗi dòng `{user_id}\t{comment_id}` (ví dụ `U_111\tCm_222`) cho biết bình luận nào do người dùng nào viết [20].

- `relations/user-reply.txt`: Liên kết giữa người dùng và trả lời bình luận. Mỗi dòng `{user_id}\t{reply_id}` (ví dụ `U_111\tRp_333`) cho biết trả lời nào do người dùng nào viết [21].

- `relations/comment-reply.txt`: Liên kết giữa bình luận và trả lời (mỗi trả lời thuộc về bình luận nào). Mỗi dòng `{comment_id}\t{reply_id}`.

- `relations/concept-course.txt` (TXT, ~19MB): Danh sách khái niệm liên quan đến khóa học. Mỗi dòng `{concept_id}\t{course_id}` cho biết khái niệm nào xuất hiện trong khóa học nào [22].

- `relations/concept-video.txt` (TXT, ~39MB): Danh sách khái niệm xuất hiện trong video. Mỗi dòng `{concept_id}\t{ccid}` liên kết khái niệm với mã ccid của video [23].

- `relations/concept-problem.txt` (TXT, ~1.3MB): Danh sách khái niệm liên quan đến bài toán. Mỗi dòng `{concept_id}\t{problem_id}` cho biết khái niệm nào liên quan đến câu hỏi nào [24].

- `relations/concept-comment.txt` (TXT, ~1.2MB): Danh sách khái niệm liên quan đến bình luận. Mỗi dòng `{concept_id}\t{comment_id}` cho biết khái niệm nào có trong nội dung bình luận nào [24].

- `relations/concept-other.txt` (TXT, ~19MB): Danh sách khái niệm liên quan đến tài nguyên ngoài khóa học (Wikipedia, Baidu Baike, Zhihu). Mỗi dòng `{concept_id}\t{resource_id}` liên kết khái niệm với ID tài nguyên ngoài (tham khảo `entities/other.json` hoặc tài liệu nói về tài nguyên ngoài) [25].

Thư mục `prerequisites/`

- `prerequisites/cs.json` (JSON, ~133MB): Dữ liệu đánh dấu (ground truth) và dự đoán quan hệ **tiền đề** giữa các cặp khái niệm trong lĩnh vực Khoa học máy tính. Các trường:
- `c1`: Khái niệm tiền đề (chuỗi ID khái niệm).
- `c2`: Khái niệm hậu tố (tức khái niệm phụ thuộc).
- `ground_truth`: Quan hệ tiền đề thực tế (1 = có, 0 = không).
- `text_predict`: Nhãn dự đoán bằng mô hình dựa trên văn bản.

- `graph_predict`: Độ tự tin của dự đoán dùng thông tin đồ thị.
(Các trường này như trong bảng [26]).

- `prerequisites/math.json` (JSON, ~59MB): Đánh dấu và dự đoán quan hệ tiền đề của các khái niệm trong lĩnh vực Toán học (có cấu trúc giống `cs.json`) [27] [26].

- `prerequisites/psy.json` (JSON, ~87MB): Đánh dấu và dự đoán quan hệ tiền đề của các khái niệm trong lĩnh vực Tâm lý học (có cấu trúc giống `cs.json`) [27] [26].

Lưu ý khi sử dụng

- **Dung lượng và hiệu năng**: Nhiều tệp dữ liệu rất lớn (hàng trăm MB đến vài chục GB), ví dụ `user-problem.json` ~21GB, `entities/problem.json` ~1.2GB, `entities/comment.json` ~2.1GB [28]. Cần có máy có bộ nhớ phù hợp và dùng các phương pháp đọc dữ liệu theo lô (chunk) hoặc cơ sở dữ liệu NoSQL để xử lý hiệu quả.
- **Định dạng**: Các tệp `.json` cần được đọc bằng trình phân tích JSON (có thể dùng Python/Pandas), các tệp `.txt` dùng phân tích tách bằng tab.
- **Bản quyền và trích dẫn**: MOOCCubeX được cấp phép theo GPL-3.0, nên khi sử dụng công khai hoặc trong công bố học thuật cần ghi rõ nguồn. Tham khảo file `CITATION.bib` kèm theo để trích dẫn thích hợp.
- **Công cụ hỗ trợ**: Trong thư mục `scripts/` có sẵn các kịch bản ví dụ để tải và đếm dữ liệu (`download_dataset.sh`, `count.sh`) cũng như tìm kiếm liên quan. Ví dụ, `./scripts/download_dataset.sh` sẽ tải toàn bộ dữ liệu và `./scripts/count.sh` đếm số dòng mỗi tệp, giúp kiểm tra nhanh quy mô dữ liệu.
- **Ngôn ngữ**: Một số trường văn bản (tiêu đề khóa học, nội dung bài toán, phụ đề video, v.v.) có thể bằng tiếng Trung hoặc tiếng Anh, tùy khóa học. Cần lưu ý mã hóa (UTF-8) khi đọc.

Kết hợp trên, báo cáo này đã mô tả đầy đủ từng tệp dữ liệu của MOOCCubeX với ý nghĩa trường, định dạng, và các lưu ý chính khi sử dụng bộ dữ liệu này. Thông tin được tham khảo từ tài liệu chính thức của MOOCCubeX [5] [29] và bảng tổng quan trong README của dự án [4] [30].

---

[1] [4] [6] [9] [22] [23] [24] [25] [30] YWX/MOOCCubeX
https://gitee.com/Yliterature/MOOCCubeX

[2] [3] [28] GitHub - THU-KEG/MOOCCubeX: A large-scale knowledge repository for adaptive learning, learning analytics, and knowledge discovery in MOOCs, hosted by THU KEG.
https://github.com/THU-KEG/MOOCCubeX

[5] [7] [8] [10] [13] [14] [15] docs/course-en.md · YWX/MOOCCubeX - Gitee
https://gitee.com/Yliterature/MOOCCubeX/blob/main/docs/course-en.md

[11] [16] [17] [18] [19] [20] [21] [29] docs/user-en.md · YWX/MOOCCubeX - Gitee
https://gitee.com/Yliterature/MOOCCubeX/blob/main/docs/user-en.md

[12] docs/concept-en.md · YWX/MOOCCubeX - Gitee
https://gitee.com/Yliterature/MOOCCubeX/blob/main/docs/concept-en.md

[26] [27] docs/prerequisites-en.md · YWX/MOOCCubeX - Gitee
https://gitee.com/Yliterature/MOOCCubeX/blob/main/docs/prerequisites-en.md