数据库设计与 AI 实践

舒琨峻 22307130118

2024 年 11 月 28 日

1 数据库设计与 AI 实践

1.1 数据库设计

有关本产品，进行相关数据库和数据表设计，最终的数据 E-R 图见图 (1)，有关数据表的具体描述如下：

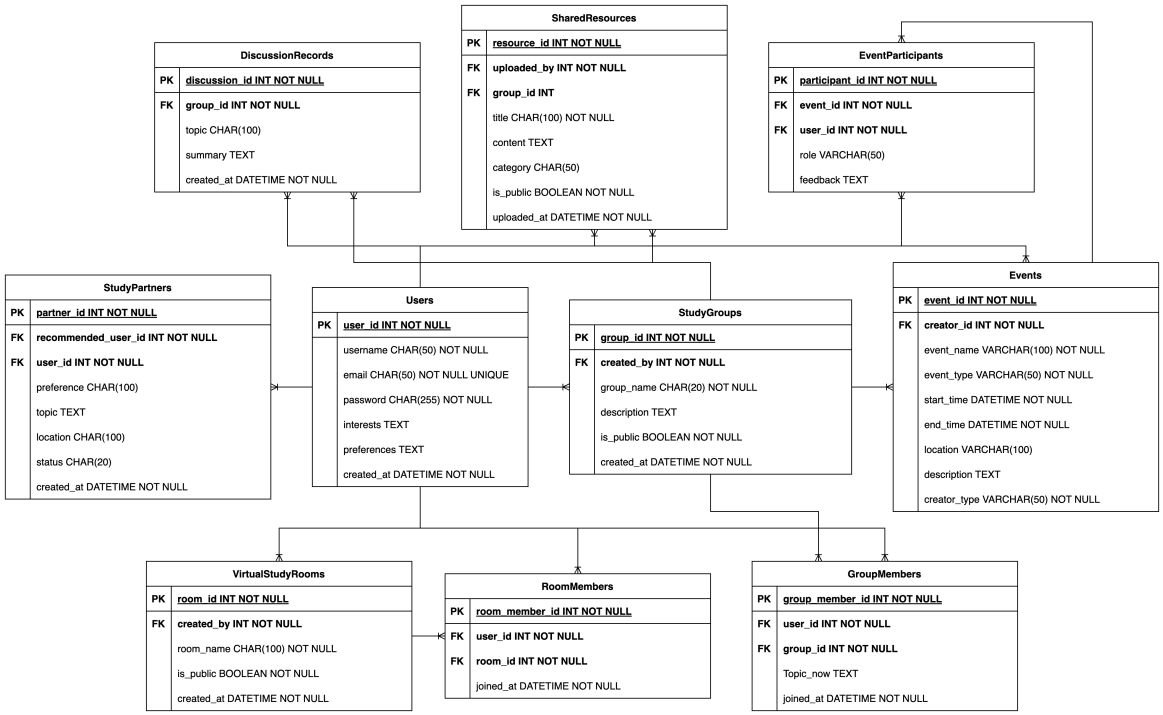

图 1: 数据 E-R 图

Users 表

用户表用于存储用户的基本信息，是数据库的主数据，是学习搭子推荐、学习小组组建、团队创建等功能的基础。

字段	含义	属性
user_id	**主键**	**唯一标识**
username	用户名	字符串，非空
email	电子邮箱	字符串，非空，唯一
password	密码	字符串，非空，加密存储
interests	兴趣	长文本
preferences	偏好设置	长文本
created_at	创建时间	时间戳

StudyPartners 表

学习搭子信息表用于存储学生的偏好信息及其搭子推荐信息，用于学习搭子匹配。与 Users 表关联，存储了两位学生的 user_id，代表本产品推荐此二人为学习搭子。

字段	含义	属性
partner_id	**主键**	**唯一标识**
recommended_user_id	外键	关联 Users.user_id 表示推荐的用户
user_id	外键	关联 Users.user_id 表示当前用户
preference	推荐偏好	字符串
topic	学习或讨论主题	长文本
location	偏好或推荐的场地	字符串
status	记录是否已经推荐了学习搭子	BOOL 类型
created_at	创建时间	时间戳

VirtualStudyRooms 表

虚拟自习室表用于存储参与学生的信息和场地信息，用于学习搭子匹配后的进阶功能。它能够为不希望线下自习的同学提供虚拟教室，督促学生学习。与 Users 表关联，存储了学生的 user_id，代表虚拟自习室参与的同学。

字段	含义	属性
room_id	**主键**	**唯一标识**
created_by	外键	关联 Users.user_id 表示使用的用户
room_name	学习室名称	字符串
is_public	是否为公开学习室	BOOL 类型
created_at	创建时间	时间戳

RoomMembers 表

虚拟自习室的使用用户、以及其中用户的关系。与 Users 表关联，存储了学生的 user_id，代表虚拟自习室参与的同学。与 VirtualStudyRooms 表关联，记录当前自习室。

字段	含义	属性
room_member_id	**主键**	**唯一标识**
user_id	外键	关联 Users.user_id 表示使用的用户
room_id	外键	关联 VirtualStudyRooms.room_id 表示当前自习室
joined_at	用户加入时间	时间戳

StudyGroups 表

学习小组表用于存储组成的小组信息，用于储存展示本小组的相关信息、讨论话题以及是否在平台中公开小组的讨论结论。与 Users 表关联，展示学习小组由某位成员创建。

字段	含义	属性
group_id	**主键**	**唯一标识**
created_by	外键	关联 Users.user_id
group_name	小组名称	字符串，非空
description	小组描述	长文本
is_public	是否公开讨论	BOOL 类型
created_at	创建时间	时间戳

GroupMembers 表

小组成员表记录用户与学习小组的关系信息，存储了用户加入学习小组的时间、当前用户在小组中讨论的话题，以及关联的用户和学习小组信息。该表通过与 Users 表和 StudyGroups 表关联，描述了每个用户与其加入的小组之间的绑定关系。

字段	含义	属性
group_member_id	主键	唯一标识
user_id	外键	关联 Users.user_id 表示用户
group_id	外键	关联 StudyGroups.group_id 表示学习小组
topic_now	当前用户讨论的话题	字符串
joined_at	加入时间	时间戳

DiscussionRecords 表

讨论记录表存储小组内的讨论记录。用于存储学习小组内部的所有讨论内容，包括讨论的主题、摘要以及记录的创建时间。该表通过与 StudyGroups 表关联，能够清晰地归档每个学习小组的讨论内容，方便成员查看和回溯相关信息。

字段	含义	属性
discussion_id	主键	唯一标识
group_id	外键	关联 StudyGroups.group_id 表示学习小组
topic	讨论主题	字符串
summary	讨论摘要	文本
created_at	创建时间	时间戳

SharedResources 表

共享资料表存储小组内共享的学习资源，以方便本小组和全平台用户学习交流。用于记录学习小组成员上传和共享的学习资源。包括资源的标题、内容、类别、是否公开等信息。该表通过与 Users 表和 StudyGroups 表关联，描述了具体的上传者以及资源所属的学习小组，旨在为小组成员提供共享学习资料的平台。

字段	含义	属性
resource_id	主键	唯一标识
uploaded_by	外键	关联 Users.user_id 表示上传用户
group_id	外键	关联 StudyGroups.group_id 表示学习小组
title	资源标题	字符串
content	资源内容	文本
category	资源类别	字符串
is_public	是否为公开资源	布尔值
uploaded_at	上传时间	时间戳

Events 表

活动表记录相关消息、比赛、讲座等活动。用于存储平台上发布的活动信息，例如系统公告、比赛、讲座、以及其他相关的学习和交流活动。该表包含活动的名称、类型、时间、地点、描述等信息，并通过与 Users 表关联记录活动的创建者及其类型。

字段	含义	属性
event_id	主键	唯一标识
creator_id	外键	关联 Users.user_id 表示事件创建者
event_name	事件名称	字符串
event_type	事件类型	字符串
start_time	事件开始时间	时间戳
end_time	事件结束时间	时间戳
location	事件地点	字符串
description	事件描述	文本
creator_type	创建者类型	字符串

EventParticipants 表

活动参与表记录用户参与的活动，及用户的行为。记录用户参与平台活动的详细信息，包括参与的角色（如主持人、参与者等）以及用户对活动的反馈。该表通过与 Users 表和 Events 表关联，清晰地描述了用户与活动之间的关系，为后续活动评估与改进提供数据支持。

字段	含义	属性
participant_id	主键	唯一标识
event_id	外键	关联 Events.event_id 表示事件
user_id	外键	关联 Users.user_id 表示用户
role	用户在事件中的角色	字符串
feedback	用户对事件的反馈	文本

1.2 AI 实践

结合本产品数据库和学习数据库的数据，我们进行 AI 训练以及生成式 AIGC 的使用，以最大化推荐准确性，提升用户体验感等。

1.2.1 数据处理与模型训练

从数据库中提取有用数据后，需要进行清洗、格式化和转换，例如：

a 缺失值处理：补全用户兴趣、学习偏好等字段，必要时通过推断填充（如基于相似用户或历史记录）。

b 标准化与编码：对数值型数据（如用户偏好评分）标准化，对分类变量（如兴趣类别）进行 One-Hot 编码。

c 时间序列处理：对用户行为记录（如资源上传时间、讨论时间）按时间序列建模，例如根据场地使用时间、自习室使用时间的历史数据推断，以辅助推荐系统。

d 特征提取：提取关键特征，例如用户活跃度、参与过的学习主题、共享资源评分等。

而后训练模型，主要包括学习搭子推荐系统、话题小组推荐系统、学习资源推荐等。可以基于以下方式建模：

a 用户兴趣和偏好，小组成员信息，有关话题的点击率和用户习惯等都可以作为特征输入模型。

b 深度学习模型分析用户的行为（例如，基于 K-means 算法的聚类分析）。

c 训练分类器（如 XGBoost 或神经网络）预测用户在不同主题的参与概率，然后排序进行推荐。

1.2.2 生成式 AIGC 的应用

有关学习搭子推荐，例如，好友推荐卡片——总结用户兴趣，并推荐适合的学习搭子，同时生成趣味性描述，例如“你们都喜欢 AI，这将是一场知识的碰撞！”。学习话题建议——生成交流话题（如“讨论最新的机器学习工具”）。

有关小组讨论的过程，可以生成讨论记录提纲，实时总结小组讨论内容，生成“会议纪要”或后续讨论建议。同时，也可以分层生成讨论主题，从入门到进阶。例如，初级小组可能获得主题“机器学习基础入门”，而进阶小组讨论“Transformer 架构的优化”。

有关学习资料共享平台的整理。例如，自动分类与质量评估。使用生成式 AI 对用户上传的资料进行自动分类（如分为讲义、论文、练习题等）。提取摘要并生成推荐标签。基于用户需求，生成推荐文档或学习资源的个性化描述。创建资料评估和用户学习反馈的总结。

同时，生成式 AIGC 也可以优化之前的 AI 模型。使用历史数据的行为反馈（如用户点击率、参与率），总结并转化为结构化数据，输入到推荐模型，进行调优。