专利申请技术交底书

|  |  |  |  |
| --- | --- | --- | --- |
| **提案名称** | 一种基于大语言模型的三层智能组卷方法及系统 | | |
| **技术联系人姓名** | 姚鑫杰 | **部门** | 软件研发部门 |
| **电话号码** | 15366188304 | **邮箱** | 15366188304@189.cn |

**一、初拟的发明名称**

一种基于大语言模型的三层智能组卷方法及系统

**二、名词解释**

大语言模型：是一种参数规模巨大，具有理解、处理和生成自然语言文本数据能力的神经网络模型。

智能组卷：是一种利用人工智能技术，针对教育领域的考试、测试等场景，自动化地生成试题、组织试卷的系统。

OSS（Object Storage Service，对象存储服务）：是一种广泛运用的云存储技术，可以用来存储大量的数据对象，如图片、视频等。

OCR（Optical Character Recognition，光学字符识别）：是一种将印刷体或手写体的纸质文档、图片文件中的文字信息转化为可编辑和可搜索的电子文本数据的技术。

**三、所属技术领域**

本发明涉及组卷系统技术领域，特别涉及一种基于大语言模型的三层智能组卷方法及系统。

**四、背景技术**

在教育场景中，试卷制作是一项繁琐而重要的任务，需要有资历的教师花费大量时间和精力，而且题目质量往往受限于教师个人的经验和能力，带有很大的主观性和局限性。随着人工智能技术的发展，智能化、自动化地生成试题、组织试卷逐渐成为可能，可以极大地提高试卷制作的效率和质量。另外，利用先进的大语言模型技术，教育者可以更好地了解每位学生的学习特点、兴趣爱好和学习进度，从而为他们量身定制个性化的学习路径和教学内容。

目前，国内外针对组卷系统的研究主要分为三类。第一类是人工组卷方式，基于大量教学经验丰富的老师人工挑选知识点契合、难度适中的试题进行组卷，这种组卷方式繁琐、耗时，且主观性较强；第二类是随机组卷方式，根据用户选择的科目、年级和知识点随机选出试题组成试卷，这种方式生成的试卷题目质量不高，知识点覆盖率低且难度不稳定；第三类是利用先进的人工智能技术进行组卷，但目前所提出的智能组卷方法大都算法较为简单，试题解析和知识点标注不够准确，生成的试卷质量良莠不齐，很难契合学生的实际情况。

因此，如何设计一种深度理解用户需求，同时结合固定题库和无题库模型生成的组卷方式，为学生量身定制一套符合特定需求的试卷，是教育场景下组卷需要面临的重要问题。

**五、发明创造的目的**

针对传统组卷系统中存在的主观性和局限性问题，本发明以大模型技术为基础，提供一种实时的三层智能组卷系统。系统从上传文件中分析用户组卷意图，深度理解用户需求和教学内容，能够精准地定制一套符合要求的试卷，确保试卷的知识点和难度与用户上传的文件相匹配。它区别于传统一刀切的教学方式，可根据每个学生的个人情况和需求，生成针对性更强的试卷，为个性化教学提供服务和支撑。

针对学生机械记忆答案而忽略真正理解知识点问题，本发明能实时生成相应题目举一反三的类似题，对原始题目进行改写，以生成一系列类型各异、难度递进的新问题并动态更新题库。通过设置大模型的提示词模板，调整问题中的数字、修改公式中的变量、重新设定假设条件等多种方式，系统边解析边生成与原题类似的改编题目，不断扩展题目范围。此外，这种方法不仅可以增加问题的多样性，还能够激发学生思维，培养学生的创新意识和问题解决能力。

针对传统组卷方法试卷编写时间长、题目质量良莠不齐问题，本发明提出一种结合类似题生成、遗传算法、大模型生成三部分进行组卷的方法，可显著缩短试题编写时间，降低人力成本。此外，结合题库组卷和无题库模型生成方式，不完全依赖于固定的题库，充分利用其训练的大量数据和语言理解能力。通过实时改编类似题，灵活调整题干内容，以确保考试题目的多样性和实用性。同时，利用大型模型生成新题目，不断丰富题库内容，保持题目的新颖性和挑战性。这一创新方法不仅能够满足不同考试需求，也能够适应教学内容和考试大纲的变化，为教育考试行业带来全新的发展方向。

**六、发明创造的技术方案以及具体实施例**

为实现上述目的，本发明提供一种基于大语言模型的三层智能组卷方法及系统，三层架构包括试题解析层、知识点和难度标注层和智能组卷层。首先，在第一层中系统进行试题分解入库，基于OSS存储、正则表达式匹配和OCR识别技术对试题进行解析，形成标准试题库；其次，在第二层中完成知识点和难度标注，利用大模型训练微调技术获取题目对应的知识点和难度，并分析出用户组卷意图；最后，在第三层中完成组卷，结合类似题推荐、遗传算法和组卷大模型，综合考虑题型、知识点覆盖、难度等因素，实时生成符合教学需求、满足用户定制化需求的试卷。

简要流程请参阅图1，包括以下步骤：

**1）步骤1**：试题解析层，该层的主要功能是完成试卷题目的解析入库，主要包括四个模块，分别是图片存储模块、正则处理模块、OCR识别模块和数据存储模块。首先，对用户上传的试卷文件进行数据预处理，自动将其中的图片上传到对象存储服务。其次，利用正则表达式匹配技术对标题和各个题型进行分割，得到科目和题型字段数据。然后，针对每个题型分别运用正则表达式和OCR技术，解析题目中的题干、答案和分析。最后，将科目和题型字段与解析后的数据合并，存储到数据库中。

**2）步骤2**：知识点和难度标注层，该层的主要功能是获取每题的知识点和难度，主要包括三个模块，分别是训练微调模块、知识点标注模块和难度标注模块。首先，依据各科教材版本分别制定标准的三级知识点和五级难度体系架构。然后，利用大量的试卷数据训练微调知识点和难度标注大语言模型，分别对模型结果进行规范化处理。最后，将题干内容批量输入模型，得到对应的知识点和难度，形成最终的标准试题库，同时系统可根据上传的试卷数据分析出用户的组卷意图。

**3）步骤3**：智能组卷层，该层的主要功能是生成满足用户需求、契合学生实际情况的试卷，主要包含三个模块，分别是类似题生成模块、遗传组卷模块和大模型组卷模块。首先，利用大模型技术改编用户上传的原题，生成类似但不同难度和类型的举一反三题目，并将其纳入题库。其次，采用遗传算法得到与学科内容及学生能力匹配最佳的题目。接下来，根据用户的组卷意图，将题型、知识点和难度参数输入经过微调的试卷大模型，生成对应的试题。最后，将举一反三的类似题、遗传算法的组卷和大模型生成的试题相结合，按用户需求进行自动排版，生成全新的试卷供用户下载。

本发明附加的方面和优点将在下面的描述中部分给出，部分将从下面的描述中变得明显，或通过本发明的实践得到了解。

**具体实现方案**

为了使本发明的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本发明进行进一步详细说明。下面通过参考附图描述的实施例是示例性的，旨在用于解释本发明，而不能理解为对本发明的限制。

如图1所示，本发明的技术实现主要包括三层：试题解析层、知识点和难度标注层和智能组卷层。这三层架构使得整个系统分工明确，耦合度降低。现存问题分析及解决方案示意图如图2所示。

1、试题解析层

该层的主要功能是完成试卷题目的解析入库，主要包含四个模块，分别是图片存储模块、正则处理模块、OCR识别模块和数据存储模块。各个模块的具体功能及实现如下：

图片存储模块：对已有试卷进行数据预处理，自动将其中的图片上传到OSS（Object Storage Service，对象存储服务）中，并标记图片所在位置。OSS是一种广泛运用的云存储技术，可以用来存储大量的数据对象，如图片、视频等。该模块选用阿里云的OSS存储，设置存储桶bucket名称为“paper”，地域Region为“华东1（杭州）”，读者权限为“公共读写”，其余为默认设置。然后获取访问密钥AccessKey Id和AccessKey Secret，将各个参数配置到python代码中,以实现文件的上传和下载。

正则处理模块：使用正则表达式匹配技术得到科目和题型字段，并对各个题型进行分割。这种技术是一种非常强大和灵活的文本处理工具，通过简洁的表达方式描述和匹配各种复杂的文本模式，提供了一种高效的方式来处理和操作文本数据。本模块主要依赖python内置库re，首先通过调用compile方法获取pattern匹配对象,然后将其作为参数调用search方法获取匹配结果，进行文本分割。

OCR识别模块：针对每个题型分别使用正则和OCR（Optical Character Recognition，光学字符识别）技术解析题目中的题干、答案和分析。本模块中的OCR技术可将印刷体或图片中的文字信息提取并转换为可编辑和可搜索的电子文本数据，从而实现文档的数字化处理和管理，代码环境包括python 3.10、conda 23.1、opencv-python 4.9.0、numpy 1.24等。

数据存储模块：将OCR识别模块得到的内容与科目、题型字段合存储到数据库中，并获得每一题型对应数量。该模块采用MySQL作为其数据库，它是一种安全可靠、性能高效的关系型数据管理系统，版本号为5.7.24。

2、知识点和难度标注层

该层的主要功能是获取每题的知识点和难度，计算难度平均值，并分析出用户组卷意图。知识点和难度标注的核心技术是利用大量优质试卷数据训练微调一个专用于解决试卷问题的大语言模型，可通过改变提示词（prompt），达到不同效果。将试题解析模块中得到的题干输入大模型，从而输出对应结构的知识点和难度值。它主要分为训练微调模块、知识点标注模块和难度标注模块。具体功能及实现如下：

训练微调模块：大模型的训练微调主要分为四个步骤，分别是数据预处理，模型微调，验证和测试以及输出结果规范化，训练微调流程如图3所示。数据预处理是在使用大型神经网络模型进行训练之前对原始数据进行的一系列处理，包括分词、序列截断和填充、数据增强等，旨在使原始数据适应模型的需求，从而提高训练效率和模型性能。模型微调是指在一个预先训练好的大型神经网络模型基础上，通过在特定任务的数据集上进行额外训练，以使模型适应该任务，包括全量微调、逐层微调、对抗训练等策略。系统中该模块采用Lora微调，其核心思想是引入关注度调整因子，动态调整每一层在训练中的贡献，使模型更专注于对任务更重要的层。微调的环境包括python 3.10、torch 2.0.1、transformers 4.33.1、accelerate 0.22、peft 0.5、jieba 0.42.1等。调整合适参数，如学习速率learning\_rate、批量大小batch\_size、梯度累积步数gradient\_accumulation\_steps等，得到效果最好的模型。模型训练完成后需使用测试数据集进行测试和评估，并根据评估结果对模型进行调整和优化，按照特定的格式输出结果。

知识点标注模块：在大模型训练完成后，设置提示词prompt模板如下，【假设你是一个题目知识点分析师，你的工作是根据用户提供的题干内容{questionStem}，分析出这道题目所属的知识点。其中知识点格式是依据各科教材制定的三级架构，例如[数与代数—数的运算—整数的四则运算]】。接着将上一层解析得到的题干内容输入模型中，替换模板中的{questionStem}，得到对应的知识点并将其存储到数据库中。

难度标注模块：设置该模块提示词prompt为如下，【假设你是一个题目难度分析师，你的工作是根据用户提供的题干内容{questionStem}，分析出这道题目的难度。其中难度划分为五级，分别是容易（0.9~1.0）、较易（0.8~0.9）、适中（0.5~0.7）、较难（0.2~0.4）和困难（0.1）】。然后将题干内容输入模型中，大模型根据训练结果得到题目难度，并将其存储到数据库中。此外，系统计算出所有题目的难度平均值作为用户上传的该套试卷整体难度。

至此，数据库中已形成包含七个指标的标准试题库，分别是题干、答案、分析、题型、科目、知识点和难度。系统根据上传的试卷数据分析出用户的组卷意图，包括希望练习的题型和知识点、各个题型对应数量、试卷整体难度。

3、智能组卷层

该层的主要功能是利用先进的人工智能算法完成组卷，生成满足用户需求、契合学生实际情况的试卷。它主要包括三个模块，分别是类似题生成模块、遗传组卷模块和大模型组卷模块。三层组卷方法具体流程如图4所示，各个模块的功能及实现如下：

类似题生成模块：首先设置提示词prompt模板：【假设你是一个出题老师，你的任务是根据用户给出的题干{questionStem}、答案{answer}、分析{analyze}内容，通过调整数字、改变公式、重新设定假设条件等方式对原题进行变化，以生成类似但不同难度和类型的举一反三题目】。然后，将试题解析层得到的题干、答案、分析内容输入大语言大模型，替换模板内的参数，实时生成母题不同类型和难度的类似试题。从选择题到开放式问题，从简单到复杂，不断扩展题目范围，激发学生思维，灵活满足不同教学目标和学生水平的需求。此外，举一反三的类似题区别于母题，解决学生机械记忆答案而忽略理解知识点的问题，培养学生的创新意识和问题解决能力。

遗传组卷模块：利用遗传算法将每套试卷模拟成染色体，每道题目想象成染色体上的基因。通过染色体的交叉变异，即题目的交换和变异，选择契合学生的试题。遗传组卷流程如图5所示，具体步骤如下：

Step1：试卷初始化。设置试卷初始列表，默认大小为30，按用户设置的题目数量随机从数据库中抽取题目得到成套试卷，总共组成30套。

Step2：计算试卷适应度。首先计算每一套试卷知识点覆盖率和题目难度，知识点覆盖率是题目涉及的知识点数目与整套试卷题目数量的比，整套试卷的题目难度为所有题目难度的平均值：

其中，为这套试卷所有题目涉及的知识点数目，为整套试卷的题目数量，为每道题的难度。再根据以下公式得到整张试卷的适应度：

其中，和分别为知识点覆盖率和试卷难度所占权重，默认分别为0.5，可根据用户需求和组卷的实际情况进行动态调整。为预设难度，默认为0.6。

Step3：选择合适试卷。首先生成试卷的选择概率和累积概率。选择概率是每套试卷适应度与所有试卷适应度和的比，累积概率表示每个个体之前所有个体的选择概率之和。

其中，为第i套试卷的适应度，为所有试卷适应度的和。然后采用轮盘赌策略选择m个个体，可以修正陷入局部最优的问题。如果每次都是按照适应度高低排序来选择，算法可能会陷入局部最优的陷阱，而无法获得全局最优解。随机生成一个大小为30的数组，每个元素的取值范围都在0到1之间，并将其按从小到大的方式进行排序。若累积概率大于数组中的元素，则试卷被选中，若小于，则比较下一个，直至选出所有试卷。

Step4：试题交换。随机选择两套试卷进行配对，生成之间的随机数，与交叉概率0.5进行比较，当随机数小于交叉概率时，随机选择一个交叉点位置，互换交叉点后的基因。检查是否存在重复元素，如果存在则使用同配置题目进行替换。

Step5：试题变异。对于试卷中的每一道题，生成一个随机数，与变异概率0.1相比较，如果小于变异概率，则使用同配置题目进行替换，计算变异后的新试卷个体的适应度。

Step6：重复迭代，直到达到最大的迭代次数或者适应度收敛，得到适应度最高的试卷。

大模型组卷模块：

用户上传一份文件，系统分析出组卷意图，为其量身定制一套相同知识点、难度契合的试卷。首先，使用解析入库的大量题目数据对大模型进行训练微调，得到一个专用于出题的试卷大模型。然后，设置提示词prompt模板：【假设你是一个出题老师，你的任务是根据用户提供的指标，包括知识点{knowledge}、难度{difficulty}和题型{questionType}，按要求生成相应的题目】。大模型组卷具体流程如下：

Step1：用户首次使用时，需要上传一份包含各种题型试题的试卷文件，试卷的格式可以包括Word、PDF、Markdown等各种文本形式。根据用户上传的试卷文件，从中解析出题干、答案、分析、题型、科目。

Step2：将问题题干输入训练完成的知识点和难度标注模型，获取题目的知识点和难度，并将所有数据存储到数据库中，系统根据上传的试卷数据分析出用户的组卷意图。

Step3：根据题干、答案、分析内容获得举一反三的类似题。

Step4：使用遗传算法获取题库内适应度最高的试题。

Step5：根据用户的组卷意图，将对应参数替换到模板中并输入试卷大模型，大模型自动根据要求生成相应试题。

Step6：最后，将举一反三生成的类似题、遗传算法的组卷和大模型生成的试题相结合，按用户需求进行自动排版。此外，用户还可手动调整题目位置、选择同配置题目进行替换，从而生成全新的试卷。试卷的格式可以包括Word、PDF、Markdown等各种文本形式，用户可通过系统提供的链接进行下载和使用。其中，大模型组卷的流程如图6所示。

**七、发明人认为要保护的发明内容的技术要点以及相应的有益效果**

1、一种基于大语言模型的三层智能组卷方法及系统，其特征在于，包括：

步骤1：对用户上传的试卷文件进行试题分解入库，基于OSS存储、正则表达式匹配和OCR识别技术对试题进行解析，形成标准试题库；

步骤2：完成知识点和难度标注，利用大模型训练微调技术获取题目对应的知识点和难度，并分析出用户组卷意图；

步骤3：完成组卷，结合类似题推荐、遗传算法和组卷大模型，综合考虑题型、知识点覆盖、难度等因素，实时生成符合教学需求、满足用户定制化需求的试卷。

2、根据权利要求1所述的一种基于大语言模型的三层智能组卷方法及系统，其特征在于，所述步骤1对用户上传的文件进行试题分解入库，具体包括：

对已有试卷进行数据预处理，自动将其中的图片上传到对象存储服务OSS中；

使用正则表达式匹配技术得到科目和题型字段，并对各个题型进行分割；

针对每个题型分别使用正则和OCR技术解析题目中的题干、答案和分析，并将数据存储到数据库中。

3、根据权利要求1所述的一种基于大语言模型的三层智能组卷方法及系统，其特征在于，所述步骤2进行知识点和难度标注，具体包括：

通过调整合适参数，利用大量优质试卷数据训练微调一个专用于解决试卷问题的大语言模型，通过设置提示词prompt模板，得到对应的知识点和难度并将其存储到数据库中。大模型的训练微调主要分为四个步骤，分别是数据预处理，模型微调，验证和测试以及输出结果规范化。

系统根据上传的试卷数据分析出用户的组卷意图，包括希望练习的题型和知识点、各个题型对应数量、试卷整体难度。

4、根据权利要求1所述的一种基于大语言模型的三层智能组卷方法及系统，其特征在于，所述步骤3完成组卷，具体包括：

通过设置提示词prompt模板，调整数字、改变公式、重新设定假设条件等方式对原题进行变化，以生成与母体类似但不同难度和类型的举一反三题目；

使用解析入库的大量题目数据对大模型进行训练微调，然后根据分析出的组卷意图为用户量身定制相同知识点、难度契合的试题。最后，将举一反三生成的类似题、遗传算法的组卷和大模型生成的试题相结合，按用户需求进行自动排版。

**有益效果**

随着人工智能技术的不断进步，个性化教育正迅速崛起，成为教育界的热门话题和发展方向，这对组卷领域提出了新的挑战和更高的要求。传统的组卷方法效率低下、存在主观性和局限性问题、不够灵活等问题。个性化教育不仅可以满足学生多样化的学习需求，还能够激发他们的学习动力和兴趣，提高学习效率和成绩。

本发明的提出具有以下意义：

1、传统架构的组卷方法及系统在生成试卷时通常依赖于教师的主观判断和教学经验，这种方式容易受到个人偏好和局限性的影响，导致试卷的设计缺乏客观性和多样性。相比之下，本发明采用先进的人工智能技术，通过自动分析用户上传的文件，深度理解用户需求和教学内容，从而能够精准地定制一套符合要求的试卷。不仅如此，本发明还能确保试卷的知识点和难度与用户上传的文件相匹配，有效解决了传统组卷方法中存在的主观性和局限性问题，为教学评估提供更加客观、高效的解决方案。

2、传统架构的组卷方法及系统依赖于固定的题库，生成的试卷题目内容不变，容易导致学生机械记忆答案而忽略真正理解知识点的重要性。本发明能够实时生成相应题目举一反三的类似题，不断扩展题目范围，激发学生思维，促进深层次理解和应用能力的培养。通过多样性和灵活性的题目设计，鼓励学生思考和探索，避免了对特定答案的依赖。这种方法培养了学生的创新意识和问题解决能力，从而提高了教学质量和学习效果。

3、传统的组卷方法一方面需要耗费大量的时间和人力资源，另一方面生成的试卷质量良莠不齐。本发明利用三层架构的大模型技术生成试题可以大大缩短试题编写的时间，减少人力成本。同时，结合题库组卷和无题库模型生成的方式，充分利用其训练的海量数据和语言理解能力，自动标注试题的知识点和难度，生成的试题更加准确和有效。先进的人工智能组卷技术促进教育教学方式的创新和改革，推动教育信息化进程，为教育教学提供更多可能性和机遇。

**八、附图**

图1为三层智能组卷系统架构示意图；

图2为现存问题分析及解决方案示意图；

图3为模型微调模块示意图；

图4为三层智能组卷方法示意图；

图5为遗传组卷模块流程图；

图6为大模型组卷流程图。



图1 三层智能组卷系统架构示意图



图2 现存问题分析及解决方案示意图



图3 模型微调模块示意图



图4 三层智能组卷方法示意图



图5 遗传组卷模块流程图



图6 大模型组卷流程图