项目编号： 2021PY—537

2021 年度本科生创新创业训练培育项目
中期汇报表

项目名称		书蛙--基于语义理解和知识图谱的学术阅读导引系统
项目负责人	**姓名**	杨天骥
	院系	数据科学与工程学院
	专业	数据科学与大数据技术
	手机	15821694809
	邮箱	10205501401@stu.ecnu.edu.cn
指导教师姓名、职称		兰韵诗 准聘副教授 高明 教授
项目立项时间		2021-10-02

填表日期： 2022 年 04 月 07 日

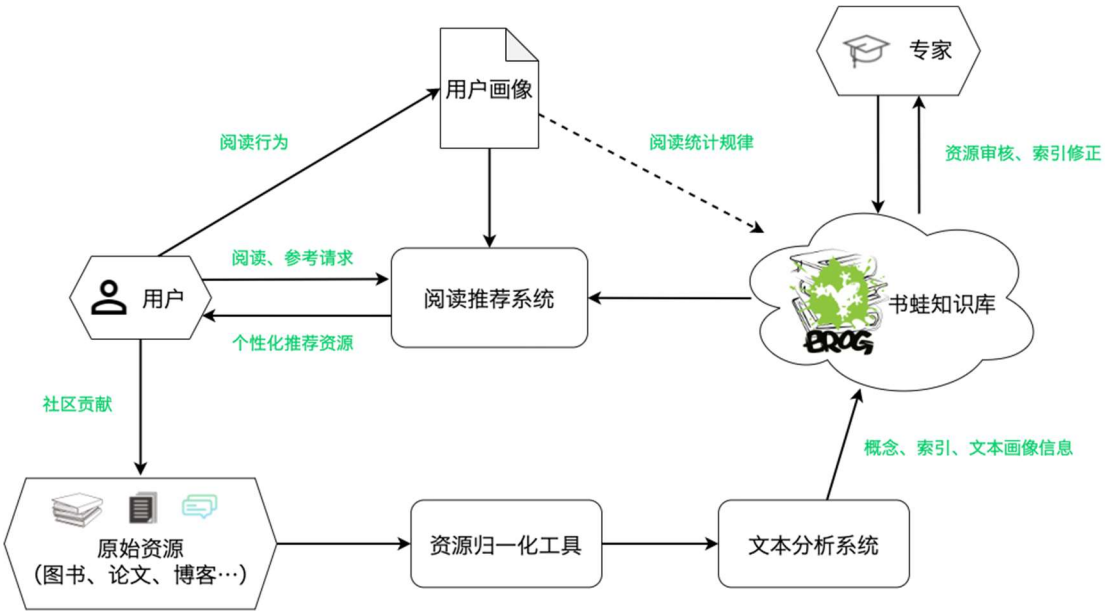
项目基本情况	**项目选题来源** （自立项目或教师科研课题的子项目）	**自立项目**
	依托单位或合作单位 （实验室、研究中心、中小学机构等）	**无**

一、项目计划达到的目标和内容

本项目旨在搭建一个智能的学术阅读云平台，解决学术阅读者的信息过载困境，填补软件市场上学术阅读服务应用的空白。通过自动化手段为学术阅读者提供陌生概念索引，阅读路径规划，阅读历史可视化，个人知识管理和阅读资源社区结合的一站式服务，最小化用户在整合和组织阅读资源上耗费的时间，提供提高学术阅读效率的根本性解决方案。

二、研究进展和当前成果（2000 字以内）

项目的前中期工作主要聚焦在阅读系统的搭建、知识库的搭建以及概念索引功能的实现上。整个阅读系统主要由知识库、社区、文本分析系统和阅读推荐系统四部分组成，其运行逻辑如下图所示：



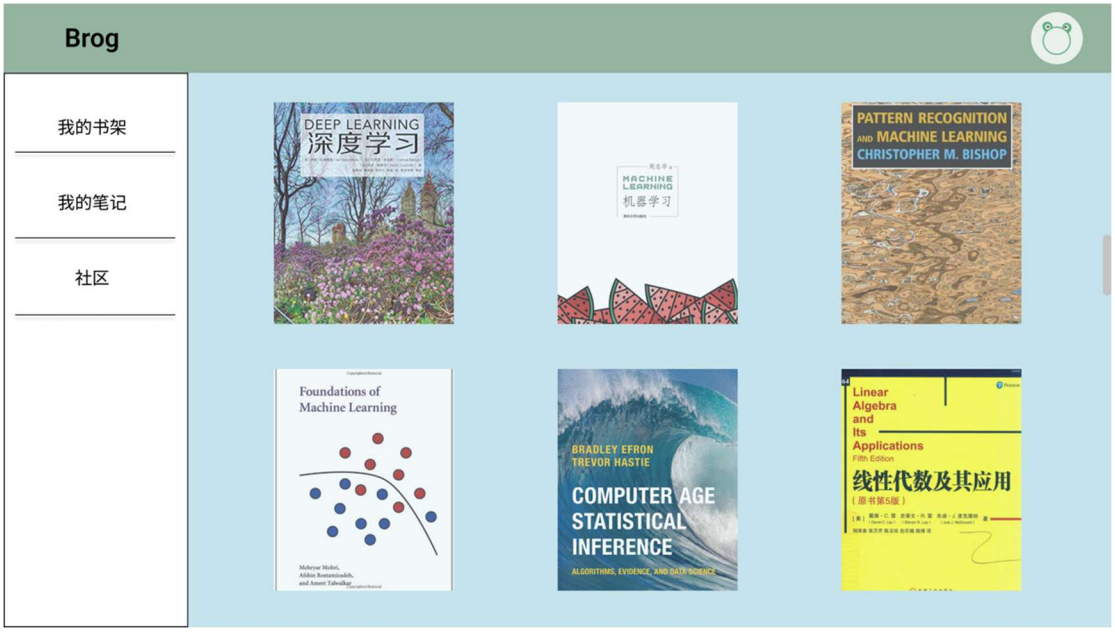
由于阅读资源种类繁多，其呈现形式也大不相同，对于一个新加入系统的阅读资源，系统首先会对其内容进行归一化。这一过程涉及诸如 PDF 结构解析、OCR、段落标记等操作，系统会自动根据输入的资源形式作相应的处理。随后，一个由多个机器学习模型组成的文本分析系统会对阅读资源的内容进行分析，抽取每一段落的核心主题和依赖主题，分析文本的阅读复杂度和背景偏向，并以此为依据

对文本建立索引和文本画像。分析完成后，这些内容会统一进入系统的结构化知识库中，并推送给相应的专家进行资源的审核和合理性的修正（用户也可辅助对资源进行标注）。另一方面，阅读系统中的推荐系统会对每一个新加入系统的用户建立用户画像，并通过概念抽样的方式对画像进行初始化。当用户对某个概念发起索引请求时，推荐系统会根据用户画像和阅读资源的文本画像匹配知识库中最适合用户阅读的参考资源，并将其推送给用户。随着用户阅读和查询的记录增多，系统会自动更新用户画像来确保推荐的资源符合用户当前的阅读需求。

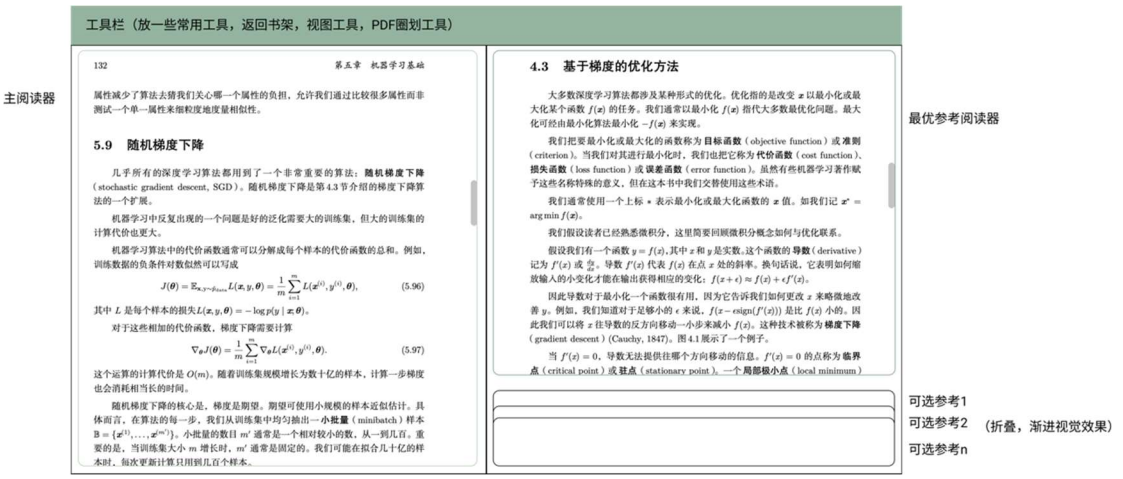
在这一系统构架和运行逻辑之上，我们在项目前中期主要完成了以下一些工作：

(1) 实现了基于 Google 公开的预训练语言模型 BERT，实现了一个准确率较高的概念标注模型。该模型能通过注意力机制理解句内上下文信息，端到端地自动标注一句句子当中的被定义项，被定义项所依赖的关键概念，与非关键概念的普通文本。

(2) 实现了一个使用体验较好的 web 应用平台，用户能在平台上自主上传文件并进行阅读，平台能识别用户上传文本中的概念并提供搜索功能。网站的主体样式如下图所示：



(3) 实现了浏览器阅读和解析 PDF 的前端组件，使得能够捕捉到用户在阅读过程中划取的单词，用户在阅读参考文献的过程中能保持无注意力瞬脱的流畅阅读。阅读器的主体样式如下图所示：



(4) 构建了关于概念间关联的知识图谱。通过对限定范围内主题的论文的爬取，完成了概念和主题词知识图谱的构建，其中大约拥有 10 万个概念节点和 100 万条概念,文本,与作者之间的关系。

(5) 初步实现了基于元学习和文本理解的冷启动序列推荐，模型能够结合既往用户的阅读情况和知识图谱的信息，在该用户并未积累阅读历史记录的情况下给出大致符合认知规律的，用于支持今后阅读的推荐信息。

(6) 阅读参考文献，初步构想了基于元学习和知识图谱的冷启动序列规划，基于用户输入的阅读目标（这与上一条并不相同），由模型生成符合用户认知习惯的阅读路径，这将作为项目中后期的主要目标。

三、**项目创新点**（项目实施过程中使用的新方法或者项目特色，600 字以内）

从用户的角度来看，本项目有以下一些创新点：

(1) 实现了一个可分屏参考的智能阅读器，解决电子书翻页造成的注意力瞬脱问题，使得用户的阅读体验更为流畅。

(2) 系统可根据用户当前的阅读内容和用户知识背景个性化地提供参考资源和学习路径推荐，帮助用户建立知识关联，加强知识的理解程度，提高阅读效率。

(3) 为学生和领域学者提供一个动态的社区平台，聚集优秀的阅读和参考资源，使画像相似的用户能够共享阅读资源和路径，提高阅读效率。

在算法设计和实现的过程（尤其是文本分析系统和推荐系统的设计与实现）中，本项目有以下一些创新点：

(1) 采用机器学习辅助标注+专家审核修正的混合资源标注方式，既保证了系统中阅读资源索引和画像信息的准确性和可用性，又能够使得大量非结构化的资源文本能够快速进行索引建立和结构化，有效降低人工建立同等系统的工作量。

(2) 在概念标注模型中，实现了新的注意力机制模块 KL-Former，能利用对称的 MLP 层编码不对称的注意力，减少了正向传播的计算开销，提高了注意力层的训练效率。

(3) 经典推荐系统以及序列推荐系统中，总是假设用户并不确定自己的浏览目标。

本项目通过考察具体应用场景，构想了用户确定浏览目标的序列推荐任务。

四、研究心得

（项目实施过程中的得失成败，如在创新思维和成长方面有何收获，有何值得借鉴的成功经验和失败教训等，2000 字以内）

（1） 团队成长

本项目团队成员都是具有较强的个人能力的本科生，都有独立进行应用开发和机器学习实现的经验。但从团队写作方面来看，大家在协作经验上仍然有所不足。例如在开会交流时，大家的思维并不活跃，因此很少产生有用的想法，却往往在进行代码实现时产生一些新的想法（实践驱动型的思维方法），这样导致项目结构的设计和排期上经常出现障碍，最终实现的质量有时不如预期。然而在临近寒假末尾的时候，团队逐渐找到了合作的最佳实践，以小周期、少人数、短时间的交流方式，将刚产生的新想法随时分享出来，使得项目取得了很多积极的进展，团队成员的之间也变得更加默契。在可预期的未来中，算法和应用实现上的重要变更也将来源于这种交流方式，并且我们的协作也会保持这样的高效率。

（2） 模型迭代

本项目当中有很多典型的必须通过机器学习算法解决的问题。由于项目团队成员都是本科生，对该领域发展的脉络并不像有经验的科学工作者那样了解，我们在遇到待解决的问题时常并不知道该检索哪个方向的文献，或者该领域内有什么经典方法能够提供可转换的相似问题。这导致了很多效率不高，但回头看却非常必要的试错过程。例如在概念识别这个问题上，我们尝试的模型包括字符串核支持向量机，叠加在 GloVe 上的 Fast-RCNN（参考 kaggle 某期比赛的基线模型），BiLSTM-CRF 等等。到最终确定要使用 BERT 预训练模型之后，我们又尝试了单纯叠加线性层和感知机层，句中单个词语间的对比学习等等，最终才醒悟到我们的任务必须要捕捉额外的句法结构信息，而 BERT 的预训练目标只关注上下文中的语义，并不能有效表征句法信息，因此下游模型仍然要添加注意力机制。在这个过程中，负责算法的同学们掌握了阅读文献和在其他资料帮助下复现论文中模型的能力，并对一些机器学习领域内的话题有了深入的了解。相信无论在项目的中后期，还是在将来的科研中，这些能力和知识都能让同学们的步履更加迅捷。

五、项目组成员

姓名	学号	专业	项目研究中承担的主要任务
杨天骥	10205501401	数据科学与大数据	机器学习模型开发，平台后端开

		技术	发
龚敬洋	10195501436	数据科学与大数据技术	构架设计，机器学习模型开发
俞致远	10195501403	数据科学与大数据技术	算法设计，机器学习模型开发
龚知遥	10194810401	数据科学与大数据技术	平台后端开发
王淼	10205501422	数据科学与大数据技术	平台前端开发

六、经费使用情况及下一步研究计划

（培育阶段经费的使用情况及下一步研究计划，2000 字以内）

经费使用情况：

目前本项目共收到培育经费 1500¥，其中 500¥已用于云服务器的租赁（至 2022/4/7），剩余 1000¥。

下一步研究计划：

完善概念标注模型和索引系统

在项目的中前期过程中，我们完成了概念标注模型，但是使用的训练语料仅限于 Wikipedia 的部分话题，因此对 prompt 的捕捉和特殊语义的捕捉上（例如：Once we have learned ..., we can now enter the field of ...）并不能达到接近人类的水准。另外我们还将添加中文支持。

在项目的中后期，我们将找到更多合适的训练数据，包括更多话题下的 Wikipedia 页面，教材或者参考书类型的材料，类似 Stack Overflow 的专业论坛中的解释性文字，与一些专业标准词典。另外，我们将尝试改变模型的训练任务，使得我们能判断并未直接出现在句中的概念。

从索引系统的角度来说，概念标注模型主要通过句法信息猜测哪些是定义，这导致同一概念的实体对齐成为了另一任务（最初的构想中同一语义的表征应该本来就是对齐的）。这也是我们之后构建索引系统时的主要目标。

构建阅读路径规划系统

目前我们完成了阅读路径规划系统的初步构想和简单实现，然而仍然缺乏训练数据和足够的实验来展现这个系统的有效性。在之后的平台测试中，我们将追踪用户的阅读路径并将这些数据持久化保存在磁盘上；

我们将迭代改进阅读路径规划系统，并反复使用这些保存的数据进行训练和测试。

模型部署及模型优化

当前推荐系统模型并未搭载到平台上，并且是用效率较低的 pytorch 框架实现的（该框架广泛用于科研实验和原型实现）。我们希望通过 tensorflow 框架的 C++API 或者 Mind Spore 框架来重写一份运行效率更高的相同模型并搭载到平台后端，这样可以减少服务器的运行时负担。

界面和用户体验优化

当前的用户界面并不足够友好，由于团队缺少工业设计方面的能力，考虑使用经费有偿聘请外部人员进行用户界面设计。

在前端实现上，目前第三个版本实现了动态分段加载长文件的组件，但这些改动并未合并到仓库主分支，使得网页端浏览长文件前会出现较长的文件准备时间，从用户角度来讲这种体验是不流畅的，这可能导致平台测试时的用户数量减少，因此在 4 月到 6 月我们将分出部分精力来快速解决这个问题。

开展平台用户测试

由于我们租赁的腾讯云服务器没有 GPU，我们实现的推荐系统在一期测试（前后端对接和实验性部署）中并未搭载到平台上，因此无法看到用户数据积累提高推荐的准确率和个性化程度的直观效果。

在二期测试（7 月到 8 月）中，我们将提高服务器的租赁费用，并把推荐系统部署到云端，使用搭载推荐系统的后端进行测试，并在真实用户测试中积累一部分用户数据来改进我们的推荐系统。

七、指导教师意见（请对项目进展情况作出评价，并对下一步研究计划提出建议）

该项目目前构建了一个可用性较高的综合的学术阅读导引系统，并且能够灵活地扩展和更换搭载的智能算法。于此同时，项目利用注意力机制标注文本中的关键概念并构建索引，使用了前沿的机器学习技术，从科学研究的角度也有一定创新性。建议在之后的研究当中更注重推荐任务，尤其是由用户确定浏览目标的序列推荐任务的研究，在推荐系统方面做出一些创新。

指导教师（签字）： 兰翠玲

2022 年 04 月 07 日