哈 尔 滨 理 工 大 学

# 毕业设计中期检查报告

题　　目：基于机器学习的英语单词智能打卡系统

院　　系：计算机科学与技术学院数据科学与大数据技术系

姓　　名：马超

指导教师：李双翼

系 主 任：姚登举

**2023 年 2 月**

# 一、毕业设计工作的进展情况

## 1. 需求分析

### 1.1 功能需求

该毕业设计是开发一个智能单词打卡系统，该系统主要有四个功能模块，分别是注册登录模块、打卡任务模块、遗忘曲线分析模块、班级权限管理模块。用户通过注册获得账号，然后进行登录，可以进行个人信息修改，系统共划分两类身份，分别是学生、教师。在打卡任务模块中主要功能有教师发布打卡任务，查看学生完成情况，班级成员权限管理。在遗忘曲线分析模块中，主要的功能有根据学生单词记忆情况分析出记忆遗忘曲线，根据曲线遗忘程度着重安排单词出现频率，最大程度的提升背单词效率以及单词的记忆程度，教师根据打卡任务完成情况，单词错误率安排个别单词考察，提升工作效率。班级权限管理模块中，学生可以搜索班级号加入班级，退出班级操作。教师可以将学生加入到自己的班级，也可以将学生移出班级。例如教师注册账户后可以创建多个班级，每个班级有唯一的班级号，学生可以搜索班级号加入班级，也可以通过教师添加的方式进入班级。根据需求，制作系统功能模块划分结构图，如图1-1所示：

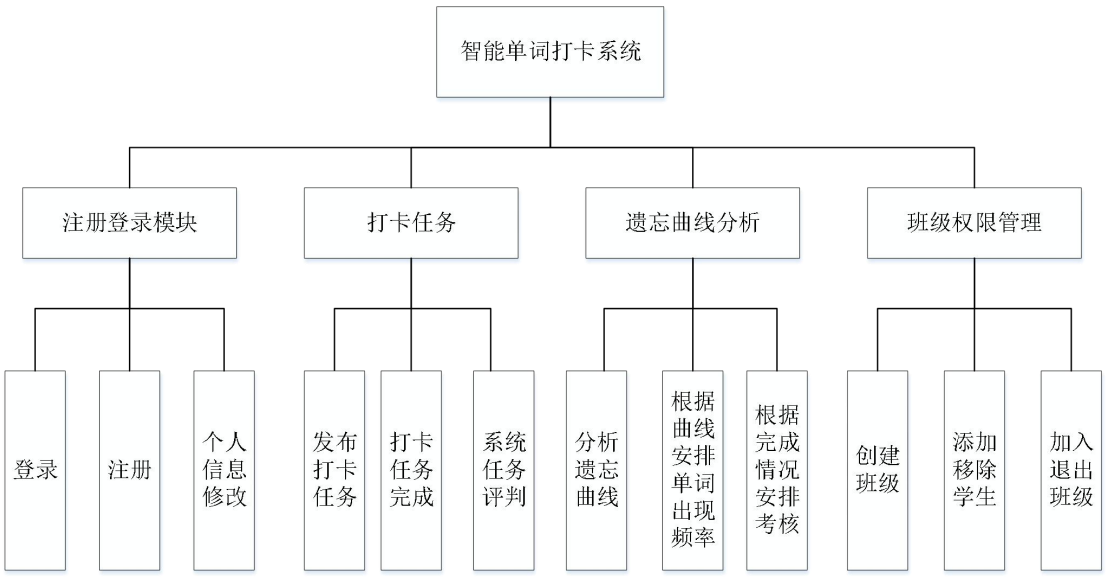


图 1-1 系统功能模块划分结构图

各个模块功能如下：

1) 注册登录模块

用户通过注册获得账号，通过账号进行登录使用系统，用户登录后要进行身份认证，通过账号密码等信息验证学生或老师的身份，不同的身份所展示的效果不同，不同的身份都可以进行个人信息修改。

2) 打卡任务管理

以教师身份登录单词打卡系统后，可以选择班级发放打卡任务，以学生的身份登录单词打卡系统后，可以对老师发放的单词打卡任务进行完成，任务完成后后台自动评判产生成绩。

3) 遗忘曲线分析模块

以教师身份登录单词打卡系统后，可以学生的班级权限管理，根据学生单词完成情况，重点考察错误率较高，易错的单词，以学生的身份登录单词打卡系统后，单词每隔一段时间会再一次出现，根据完成情况系统分析出遗忘曲线，根据遗忘曲线安排后续单词出现的频率，最大程度上保证背单词的效率以及成功率。

4) 班级权限管理模块

在这个模块中，参与者是老师和学生，教师可以创建班级，添加学生，移出学生。学生可以选择加入班级，，退出班级。事件会同步在一个关系内的用户群体中。

## 1.2 非功能需求

1) 性能需求：用户在软件响应速度、结果精度、运行时资源消耗量等方面的要求。

2) 可靠性需求：用户在软件失效的频率、严重程度、易恢复性，以及故障可预测性等方面的要求。

3) 易用性需求：该系统操作简便，人机界面简单明了，能一目了然的清楚需要做什么，该如何操作。不必借助任何操作手册或相关的系统帮助就可顺利进行各种操作。具有很强的易理解性、易学习性和易操作性。

4) 运行环境约束：用户对软件系统运行环境的要求。

5) 外部接口：用户对待开发软件系统与其他软件系统或硬件设备之间的接口的要求。

6) 安全性需求：该系统中涉及用户的重要信息，不同的用户具有不同的使用权限，只有具有最高权限的系统管理员才可以对系统进行修改，具有一般权限的用户只能读取自己相关信息，不能浏览其他用户信息。系统还要提供方便的手段供系统维护人员进行数据备份以及系统意外崩溃时数据的恢复等工作。

7) 可保障性需求：用户在软件可配置性、可扩展性、可维护性、可移植性等方面的要求。

### 1.3 需求模型

1) 学生具备的功能有：登录、注册、个人信息修改、完成打卡任务、加入班级，退出班级。

2) 老师具备的功能有：登录、注册、个人信息修改、发布打卡任务、查看打卡任务完成情况，创建班级，拉取学生，移除学生。

学生老师用例图如图1.2所示：

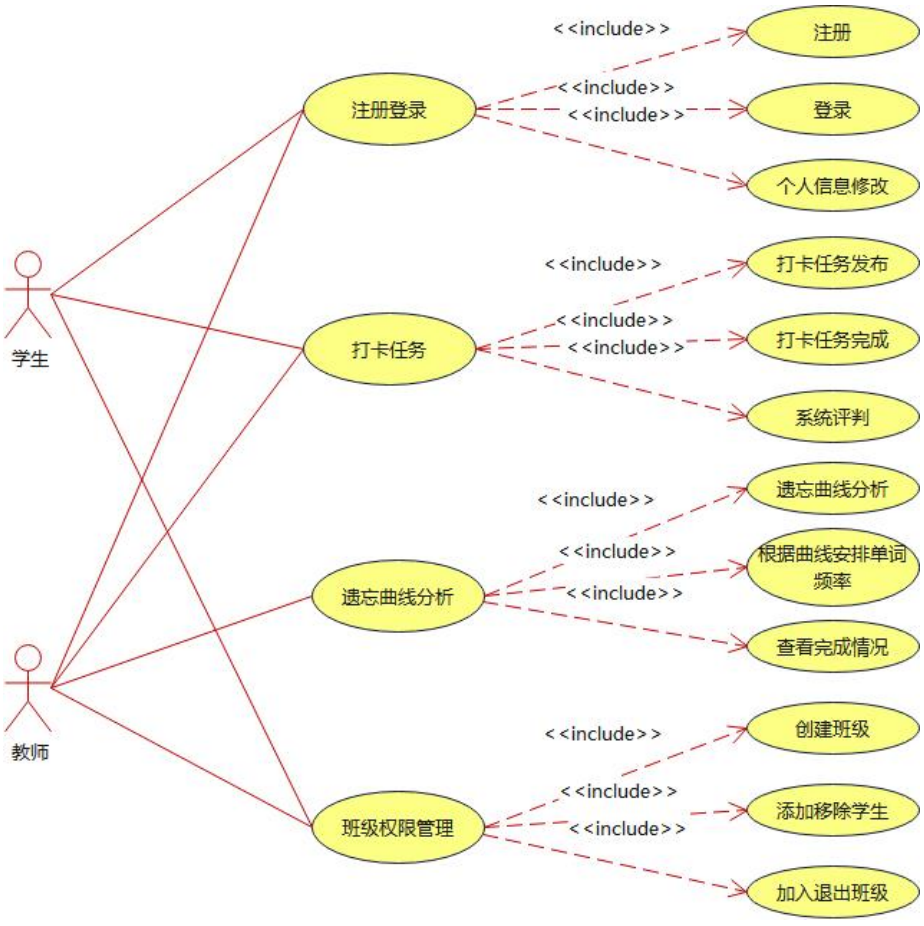


图1-2 学生老师用例图

3）系统用例图如图1.3所示：

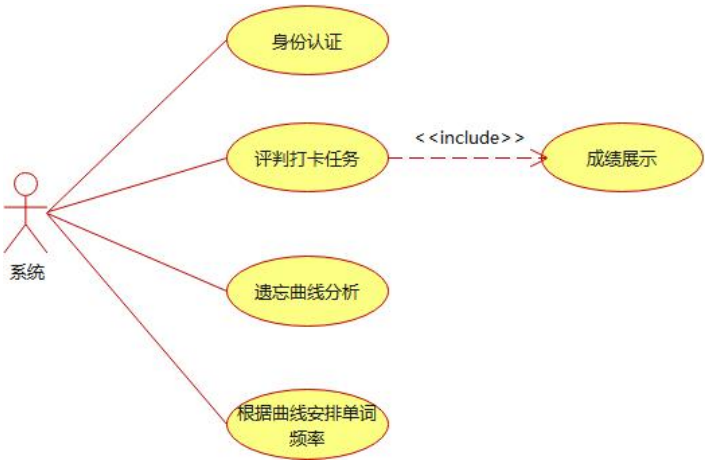


图1-3 系统用例图

### 1.4 可行性分析

1）技术可行性：机器学习算法已经被广泛应用于自然语言处理领域，其在单词分类、预测和翻译方面的表现已经超越了人类。因此，我们可以使用机器学习算法来实现基于英语单词的智能打卡系统。

2）数据可行性：目前，大量的英语单词语料库和单词库已经公开发布，可以使用这些数据来训练和优化机器学习模型，同时也可以通过数据挖掘和爬虫等技术来扩充和更新语料库和单词库。

3）实施可行性：基于机器学习的英语单词智能打卡系统可以为学生、教师和英语学习者提供方便、快捷、高效的学习工具，具有很高的市场需求和潜在用户群体。

4）安全可行性：在系统的开发过程中，可以使用各种安全技术和措施来保护用户数据的安全和隐私，例如数据加密、用户认证、访问控制等，以确保系统的安全性和可靠性。

综上所述，基于机器学习的英语单词智能打卡系统技术方案具有很高的可行性，可以为用户提供优质的学习体验和商业价值，同时也可以促进机器学习技术在教育领域的应用和发展。

## 2. 概要设计

### 2.1 架构设计

### 2.1.1 设计目标

1) 提高学习效率：通过机器学习算法的应用，实现对学习者的单词掌握程度进行智能分析，从而提高学习效率。
2) 提高系统的智能化程度：通过采用机器学习算法，对学习者的学习情况进行全面分析，从而提高系统的智能化程度。
3) 提供个性化的学习方案：根据学习者的学习情况和单词记忆情况，提供个性化的学习方案和建议，从而更好地帮助学习者掌握英语单词。
4) 提供多种学习方式：除了传统的单词记忆方式外，系统还应提供多种学习方式，例如单词拼写、听力理解等，满足不同学习者的需求。
5) 实现可扩展性：系统架构应该具有可扩展性，能够方便地添加新的学习功能和模块，以适应不断变化的学习需求。
6) 提供良好的用户体验：系统应该提供良好的用户界面和交互体验，方便用户使用，从而提高学习积极性。

### 2.1.2 系统总体架构设计

系统总体架构自顶向下主要包括 4 个层次，包括：

A. 数据层：该层负责管理数据的存储和处理，包括单词库、用户数据等。
B. 处理层：该层负责对用户的操作进行处理，包括用户的登录、注册、单词打卡等功能的处理。
C. 机器学习层：该层负责训练和应用机器学习模型，对用户进行单词学习进度的预测和推荐。
D. 应用层：该层是整个系统的核心层，负责将机器学习层的推荐结果反馈给用户，并提供用户操作的界面。

系统总体架构图如图 2-1 所示：


图 2-1 系统总体架构

### 2.1.3 系统技术架构

系统技术架构设计图通常包括以下组件：

1) 用户界面：提供给用户进行单词打卡、学习进度追踪、单词查询等功能的图形用户界面（GUI）。
2) 数据库：存储单词数据、用户信息、学习记录等数据的数据库。
3) 后端服务器：提供基于机器学习的单词学习算法，包括单词推荐、学习计划生成等功能的服务器端应用。
4) 前端服务器：处理用户请求，与后端服务器进行交互并返回数据给用户的服务器端应用。
5) 机器学习模块：使用机器学习算法进行单词推荐、学习计划生成等功能的模块。

系统技术架构设计图如图 2-2 所示：

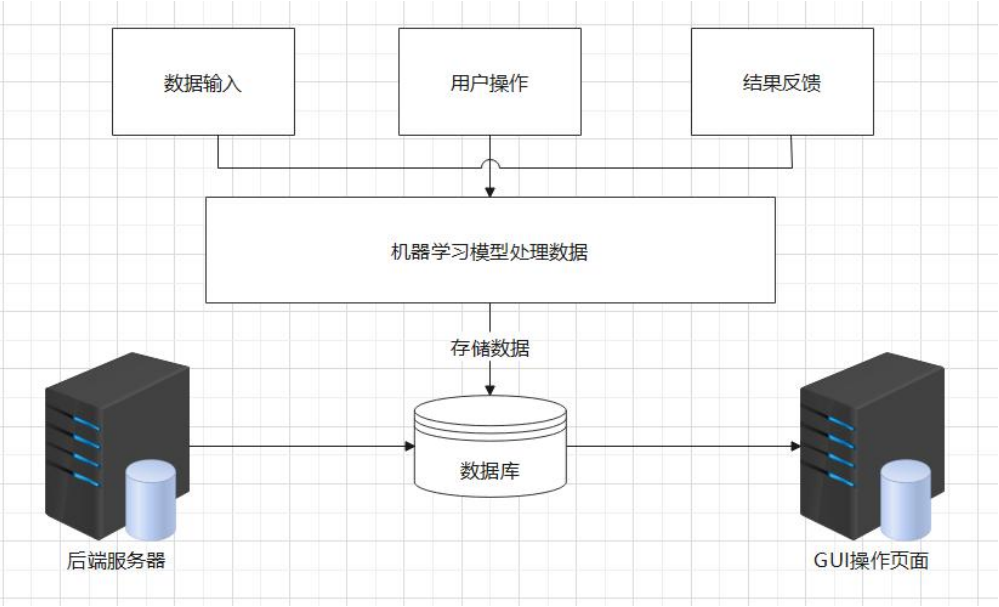


图 2-2 系统技术架构

## 2.2 系统功能

1）用户管理功能：该功能提供用户注册、登录和个人信息管理等功能，以便用户可以使用系统并查看自己的学习记录。

用户注册和登录模块主要实现用户的登录、注册。登录需要输入用户名和密码，如果信息输入正确则跳转到主页，如果信息输入错误则跳转到登录页面；注册需要填写用户信息，如果信息输入正确，则返回登录页面，如果错误则提示注册失败。用户登陆注册流程图如图 2-3 ，2-4 所示：

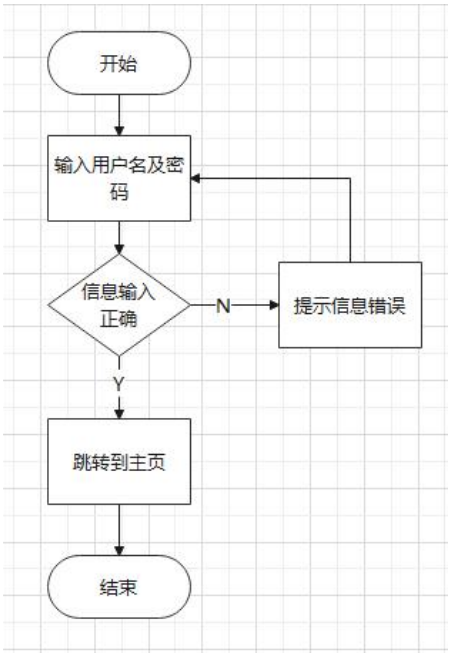


图 2-3 用户登录流程图

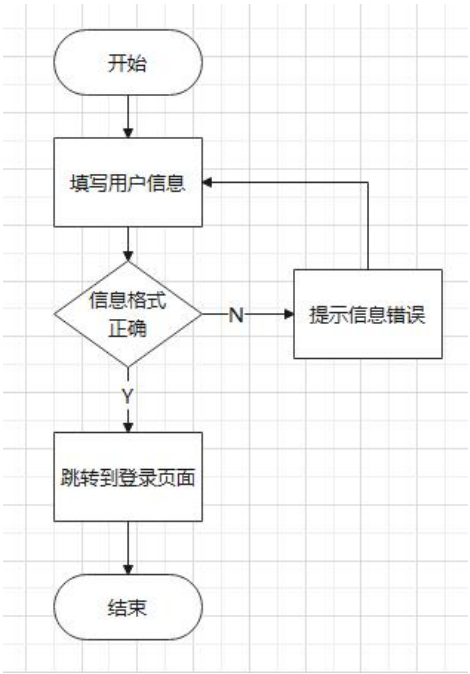


图 2-4 用户注册流程图

2）单词输入功能：该功能允许用户输入要学习或使用的英语单词，并将其传递给单词检测模块进行检测。

3）单词检测功能：该功能使用机器学习模型和规则引擎检测用户输入的英语单词是否被正确使用，并提供错误提示和纠正建议。

4）单词定义功能：该功能允许用户查看英语单词的定义、用法和语境，以便用户更好地理解和使用单词。

5）学习记录和统计功能：该功能记录用户输入的单词和检测结果，并提供单词学习历史记录和统计信息，例如单词学习次数、错误率等。

6）数据收集和预处理功能：该功能负责从不同来源收集英语单词数据，并将其进行预处理，例如清理数据、去除停用词、进行词干提取、词向量化等。

7）模型训练和更新功能：该功能使用机器学习算法对预处理后的数据进行训练，以学习单词的定义、用法和语境，并不断更新和优化模型。

8）数据存储功能：该功能负责存储用户输入的单词、模型训练的数据和模型参数等重要数据。

这个基于机器学习的英语单词智能打卡系统的功能结构设计可以帮助用户学习英语单词，并提供错误提示和纠正建议，从而提高英语写作和口语的准确性和流畅性。同时，它可以为学习者提供更好的学习记录和统计信息，以便他们更好地掌握自己的学习进度和成果。

## 2.3 功能结构设计

系统工作流程为主要如下，系统分为教师端学生端，教师学生进行登陆注册，教师创建班级并发布打卡任务，学生选择加入班级，完成每日教师发布的打卡任务，系统根据学生背单词情况好坏安排后续的单词出现频率，教师端还能看到学生的单词完成情况，以及错误率，以便安排后续的考察。系统工作图如图 2-5 所示：

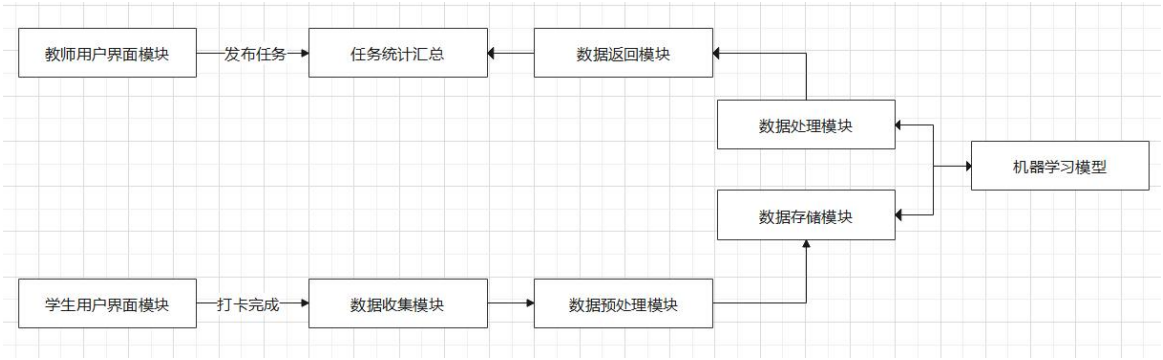


图 2-5 系统工作流程

1）数据收集模块：该模块负责从不同来源收集英语单词数据，并将其存储到数据库中。可以使用公共数据集，或从用户生成的数据集中获取数据。

2）数据预处理模块：该模块负责对收集的英语单词数据进行预处理，例如清理数据、去除停用词、进行词干提取、词向量化等。这些预处理步骤可以提高后续模型

训练的效果。

3）模型训练模块：该模块负责使用机器学习算法，例如神经网络、支持向量机等，对预处理后的数据进行训练，以学习单词的定义、用法和语境。可以使用已有的预训练模型，例如 GPT-2、BERT 等。

4）单词检测模块：该模块负责接收用户输入的英语单词，并使用训练好的模型检测单词是否被正确使用，并提供错误提示和纠正建议。可以使用模型预测或规则引擎进行单词检测。

5）用户界面模块：该模块负责提供用户界面，以便用户输入单词和查看错误提示和纠正建议。用户界面可以是 Web 应用程序、移动应用程序。具体的页面展示流程图如下图 2-6 所示：

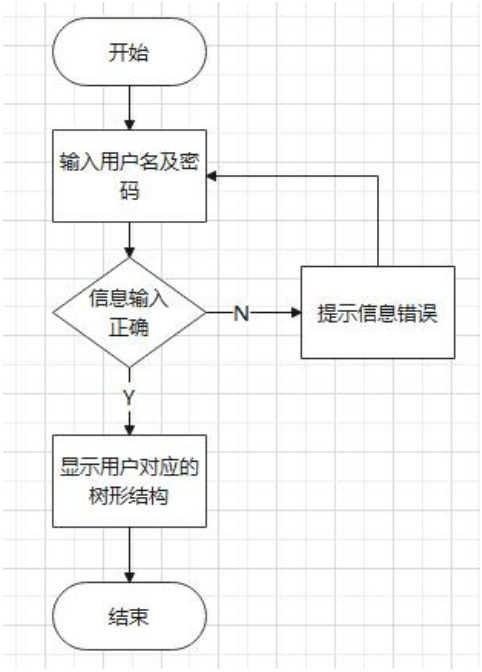


图 2-6 页面展示流程图

用户对于系统的操作就是发布打卡任务与打卡任务的完成，其他的操作也同理，用以下的数据流图 2-7 来表示。

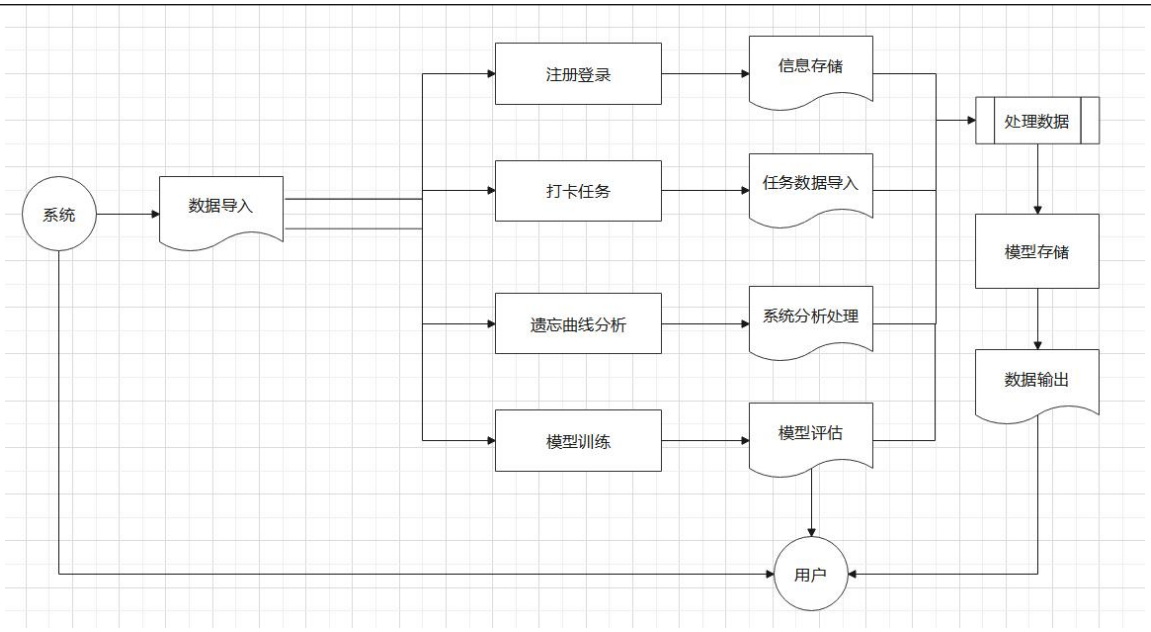


图 2-7 文件操作数据流图

6）数据存储模块：该模块负责存储用户输入的单词、模型训练的数据和模型参数等重要数据。可以使用关系型数据库或 NoSQL 数据库进行数据存储。

总体而言，这个基于机器学习的英语单词智能打卡系统的架构设计可以帮助用户学习英语单词，并提供错误提示和纠正建议，从而提高英语写作和口语的准确性和流畅性。

# 3. 详细设计

## 3.1 数据库设计

该数据库包含以下表格：

1）Users 表: 该表存储系统的用户信息，包括用户 ID、用户名、密码。

表 3-1 用户信息表

| 字段名称 | 数据类型 | 字段说明 | 是否主键 | 是否为空 |
|---|---|---|---|---|
| User_id | Numberic | 用户_id | Yes | No |
| User_name | Varchar2(20) | 用户姓名 | | No |
| password | char(7) | 密码 | | No |
| avatar | Varchar2(30) | 头像 | | |

| | | | | |
|---|---|---|---|---|
| class | numberic | 班级 | | |

2）Words 表：该表存储单词的信息，包括单词 ID、单词名称和单词的定义等。

表 3-2 单词信息表

| 字段名称 | 数据类型 | 字段说明 | 是否主键 | 是否为空 |
|---|---|---|---|---|
| Word_id | Numberic | 单词_id | Yes | No |
| Word_name | Varchar2(20) | 单词名称 | | No |
| Word_defined | char(20) | 定义 | | No |
| Word_phonetic | Varchar2(30) | 音标 | | |

3）Word_Usages:表：该表存储单词的用法信息，包括用法 ID、单词 ID、用法的语境和示例句子等。

表 3-3 单词用法表

| 字段名称 | 数据类型 | 字段说明 | 是否主键 | 是否为空 |
|---|---|---|---|---|
| Usages_id | Numberic | 用法_id | Yes | No |
| Word_id | Numberic | 用户姓名 | | No |
| Usages | Varchar2(30) | 用法 | | No |
| example | Varchar2(30) | 例句 | | |

4）User_Words 表：该表存储用户学习单词的记录，包括用户 ID、单词 ID、学习时间和学习次数等。

表 3-4 学习记录表

| 字段名称 | 数据类型 | 字段说明 | 是否主键 | 是否为空 |
|---|---|---|---|---|
| User_id | Numberic | 用户_id | Yes | No |
| User_id | Numberic | 单词_id | | No |

| | | | | |
|---|---|---|---|---|
| Time | Numberic | 学习时间 | | No |
| frequency | Numberic | 学习次数 | | |

5）User_Word_Usages 表：该表存储用户学习单词用法的记录，包括用户 ID、用法 ID、学习时间和学习次数等。

表 3-5 学习用法表

| 字段名称 | 数据类型 | 字段说明 | 是否主键 | 是否为空 |
|---|---|---|---|---|
| User_id | Numberic | 用户_id | Yes | No |
| Usages_id | Numberic | 单词_id | | No |
| Time | Numberic | 学习时间 | | No |
| frequency | Numberic | 学习次数 | | |

6）User_Word_Checkins 表：该表存储用户打卡记录，包括用户 ID、单词 ID、打卡时间和打卡结果等。

表 3-6 打卡记录表

| 字段名称 | 数据类型 | 字段说明 | 是否主键 | 是否为空 |
|---|---|---|---|---|
| User_id | Numberic | 用户_id | Yes | No |
| Word_id | Numberic | 单词_id | | No |
| Create_time | Numberic | 打卡时间 | | No |
| result | Varchar2(30) | 打卡结果 | | |
| Correct_rate | numberic | 正确率 | | |

7）Word_Vectors 表：该表存储单词的向量表示，用于机器学习模型训练和推断。

表 3-7 模型表

| 字段名称 | 数据类型 | 字段说明 | 是否主键 | 是否为空 |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| Model | Varchar2(30) | 模型 | Yes | No |
| Judge | numberic | 模型推断 | | No |
| time | char(7) | 时间 | | No |
| frequency | numberic | 次数 | | |
| architecture | numberic | 模型结构 | | |

8）Model_Weights 表: 该表存储机器学习模型的参数，用于模型训练和更新。

这些表共同支持基于机器学习的英语单词智能打卡系统的数据管理和分析功能，从而帮助用户更好地学习英语单词。

表 3-8 参数表

| 字段名称 | 数据类型 | 字段说明 | 是否主键 | 是否为空 |
|---|---|---|---|---|
| Parameter | Numberic | 参数系数 | Yes | No |
| Parameter_x | numberic | 参数_x | | No |
| Parameter_y | numberic | 参数_y | | No |
| Parameter_z | numberic | 参数_z | | |

## 3.2 算法设计

1）数据准备: 从单词表格中读取单词及其定义信息，并使用朴素贝叶斯等算法将单词转换为向量表示。

2）模型训练: 使用训练数据，如单词及其用法信息，训练机器学习模型。常见的机器学习算法包括朴素贝叶斯、支持向量机（SVM）和随机森林等。

3）模型更新: 在用户学习新单词或单词用法时，将新数据添加到数据表格中，并使用更新算法更新机器学习模型的参数。常见的更新算法增量学习。

这些步骤共同支持基于机器学习的英语单词智能打卡系统的学习和管理功能，帮助用户更好地学习英语单词和提高语言水平。

## 二、毕业设计工作存在的问题及解决方案

### 2.1 现有背单词 App 功能较完善

互联网技术的不断发展与普及以及教育教学的不断改革与深化，移动互联网技术在教育领域也得到了广泛的实践与应用。随着技术的不断进步，新的教育与学习方式也随之应运而生，以满足现代化人们的生活与学习需求。英语作为重要的交流工具与教育事业中学习的重点学科，其教育与学习的方式同样受到技术发展的影响，无论是科学研究还是工作与生活对于英语方面的人才需求很大，其高水平的英语人才在社会的各个方面需求更是炙手可热。移动语言学习应用于相关学科的教学活动的教育改革正是在这样的社会背景下应运而生的。手机上下载这些App之后，我们便可以利用碎片化的时间进行学习，而不必时刻抱着一本厚厚的单词书，不必每次打开单词书总是abandon，这些软件会根据我们的水平和需求提供相应的单词的练习。市面上现有的背单词软件发展较完善，发展空间比较小。但是现有软件绝大部分都是付费机制，相对于付费，免费使用更能吸引人。而且本系统应用单词打卡模式，教师学生两端相配合。教师负责发布打卡任务，学生可以自行完成，家长也可以起到监督的作用，背单词的任务会完成的更好，更有效果。而且我认为这种打卡鼓励机制可以使我们更动力背单词，在一个人背单词的情况下，坚持的时间也是一个问题，这种鼓励机制和家长老师的双重监督，不会存在中途放弃的情况。而且本系统主要针对于二年及以上大学以下的小中学生，受众群体不同。市面上的软件绝大多是针对于高等教育人群，以及应试人群，少有软件主要针对中小学提出背单词的解决策略。本系统所用鼓励机制以及打卡任务形式双重模式，针对中小学生，会更好的形成约束力，久而久之，背单词形成一种习惯，可以脱离监督。我相信这种打卡鼓励机制会有一定的前景以及市场空间。

### 2.2 产生的问题

1）数据采集问题：为了训练机器学习模型和构建单词库，系统需要大量的语料库和相关数据。这些数据可能很难获取，特别是对于一些不常见的单词和语言结构。

解决方案：可以使用一些已有的英语单词数据集，并将其与网上公开的语料库结

合使用。此外，还可以通过对用户学习数据进行分析来识别有用的单词和语言结构。

2）人机交互问题：由于系统需要和用户进行交互，可能存在用户体验不佳的问题，例如操作复杂、页面设计不合理等。

解决方案：可以进行用户体验测试和调查，以确定用户的需求和偏好，并优化系统的界面设计和功能交互。

3）系统性能问题：由于系统需要处理大量的单词和学习数据，可能会导致性能问题，例如响应时间过长或资源占用过多。

解决方案：可以优化系统的代码和算法，增加硬件资源以提高系统性能，并使用缓存等技术来减少对后端数据库的访问次数。

综上所述，这些问题并不是无法解决的，只要在系统设计和开发过程中充分考虑这些问题，并采取合适的解决方案，就可以开发出高效、准确和用户友好的基于机器学习的英语单词智能打卡系统。

## 三、下一步工作预测及可能存在的问题

### 3.1 工作预测：

1)模型优化：可以使用更多的数据和更复杂的模型来进一步提高系统的性能和准确度。此外，可以使用迁移学习、增量学习等技术来快速适应新的数据集。

2)用户体验改进：可以根据用户反馈和使用数据来改进系统的用户体验。例如，改善系统的界面设计和操作流程，增加用户互动和反馈机制等。

3)新功能开发：可以开发新的功能来增强系统的实用性和吸引力，例如单词分类、单词互动游戏、多语言支持等。

### 3.2 可能存在的问题：

1)数据采集问题：随着系统的发展，需要不断扩充和更新语料库和单词库，而这些数据的收集和处理可能会变得更加困难和复杂。

2)模型迁移问题：随着系统规模的扩大和业务场景的变化，可能需要将机器学习

<table>
<tr><td colspan="3">模型从一个环境或场景迁移到另一个环境或场景，这可能会导致模型性能下降或不适用。<br>3)个性化预测问题：在个性化预测方面，可能会存在一些无法捕捉的特征和因素，例如用户的情感状态、记忆能力等，这可能会影响系统的预测准确度。<br>4)数据隐私问题：随着用户数量的增加，系统需要更加严格地管理用户的隐私和数据安全问题，以避免可能的数据泄漏和滥用。<br>综上所述，基于机器学习的英语单词智能打卡系统在未来的发展中需要解决这些问题，并不断更新和优化系统的功能和性能，以满足用户的需求和提高用户的体验。</td></tr>
<tr><td>检查结论</td><td>指导教师意见</td><td>指导教师签字：<br>年　月　日</td></tr>
</table>