

学号： 2018091620013

姓名： 廖梓尧

专业方向： 系统与技术

企业名称： 北京字节跳动科技有限公司

毕设课题名称： 基于BERT和BM25的智能问答系统设计与实现

企业导师： 于光华

校内导师： 吴劲

**电子科技大学信息与软件工程学院**

**本科毕业设计（顶岗实习）**

**中期报告**

目 录

[第一章 毕业设计的进展情况 1](#_Toc1377762078)

[1.1课题工作完成情况 1](#_Toc1878562992)

[1.2. 知识技能学习情况 5](#_Toc703628350)

[1.3. 职业素养学习培养 6](#_Toc1836718068)

[第二章 存在问题与解决方案 8](#_Toc1790626898)

[2.1. 存在的主要问题 8](#_Toc230445628)

[2.2. 解决方案与可行性研究 9](#_Toc1186654255)

[第三章 当期任务完成度与后续实施计划 11](#_Toc417434096)

[3.1 当期任务完成度 11](#_Toc2133260370)

[3.2 后续实施计划 11](#_Toc1467551925)

[参考文献 12](#_Toc1295517680)

说明:

1.报告要求3000字以上。

2.本模板仅为基本参考，请各位同学根据个人情况进行目录结构扩展。

3.封面和目录单面打印，从正文第1页开始双面打印。

# 第一章 毕业设计的进展情况

## 1.1课题工作完成情况

根据课题要求，课题工作包括对问答系统进行设计，包括：需求分析、总体设计、详细设计、代码编写、测试等。

目前课题工作已经完成了需求分析、总体设计、详细设计以及部分代码编写，具体完成情况如下：

**①需求分析：**

在需求分析阶段，首先通过需求确定的方法与课题需求描述，来确定完整的系统需求。一个问答系统需要包含如下功能：一个用户可以创建问答聊天窗口；一个用户可以删除一个聊天窗口；一个用户可以在一个聊天窗口内发出多条消息；问答系统可以分别为一条询问消息返回一条回答结果。然后，根据系统需求画出对应的用例图以及活动图，分别如下图1-1和图1-2所示：

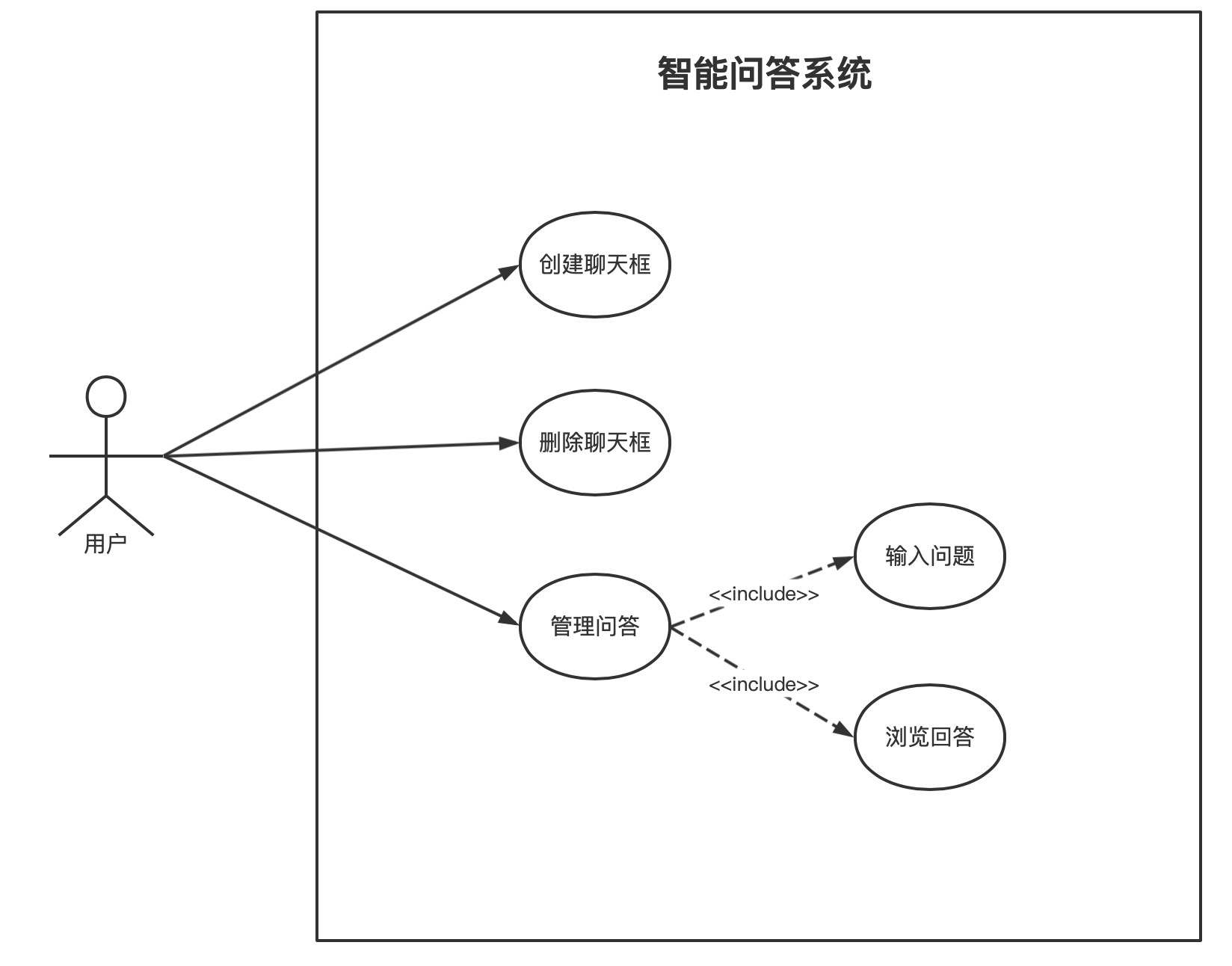


图 1-1 智能问答系统 用例图

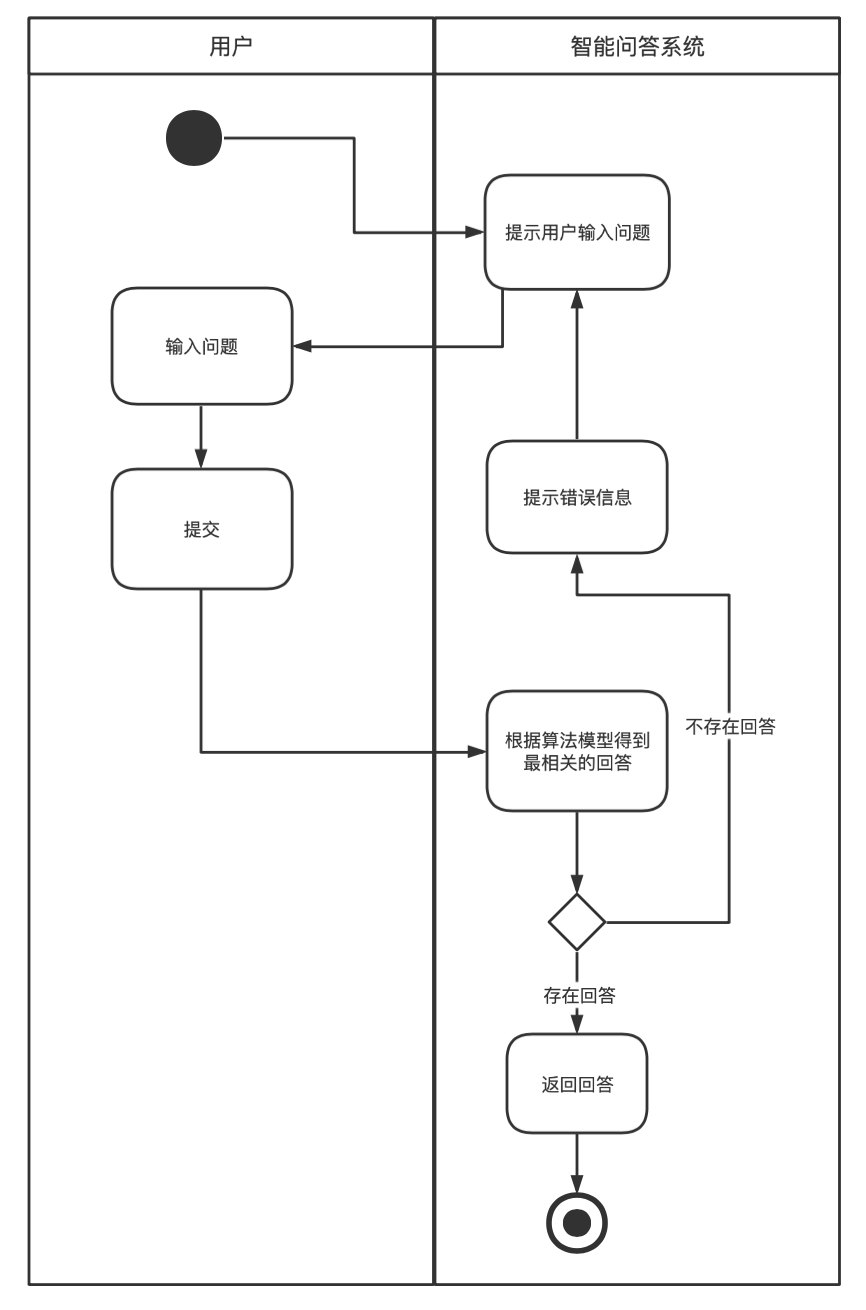
****

图 1-2 智能问答系统 管理问答 活动图

**②系统设计：**

在系统设计阶段，基于课题对智能问答系统的逻辑体系结构和物理体系结构进行建模，并画出对应的构件图与部署图，分别如下图1-3和图1-4所示：

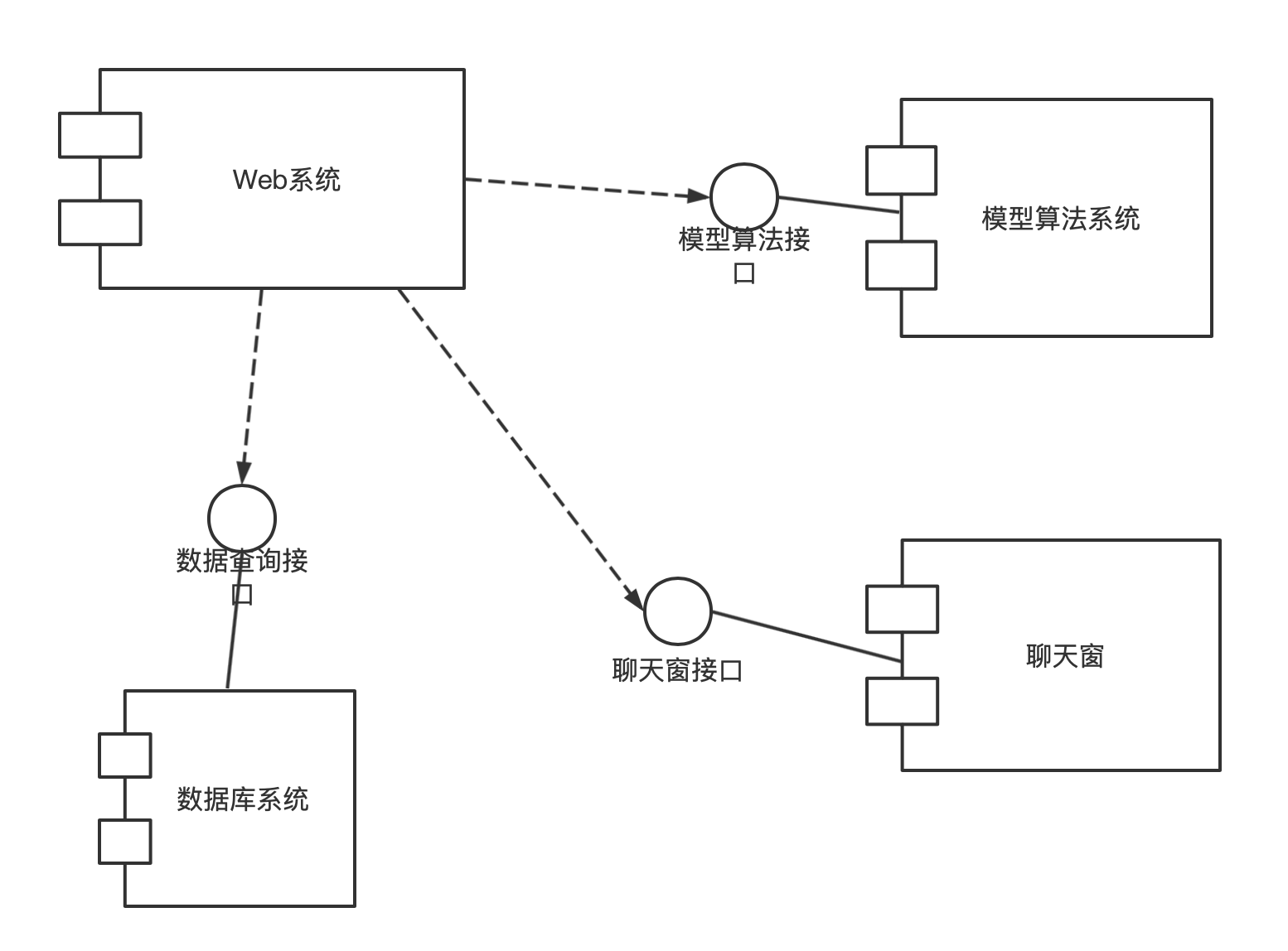
****

图 1-3 智能问答系统 构件图

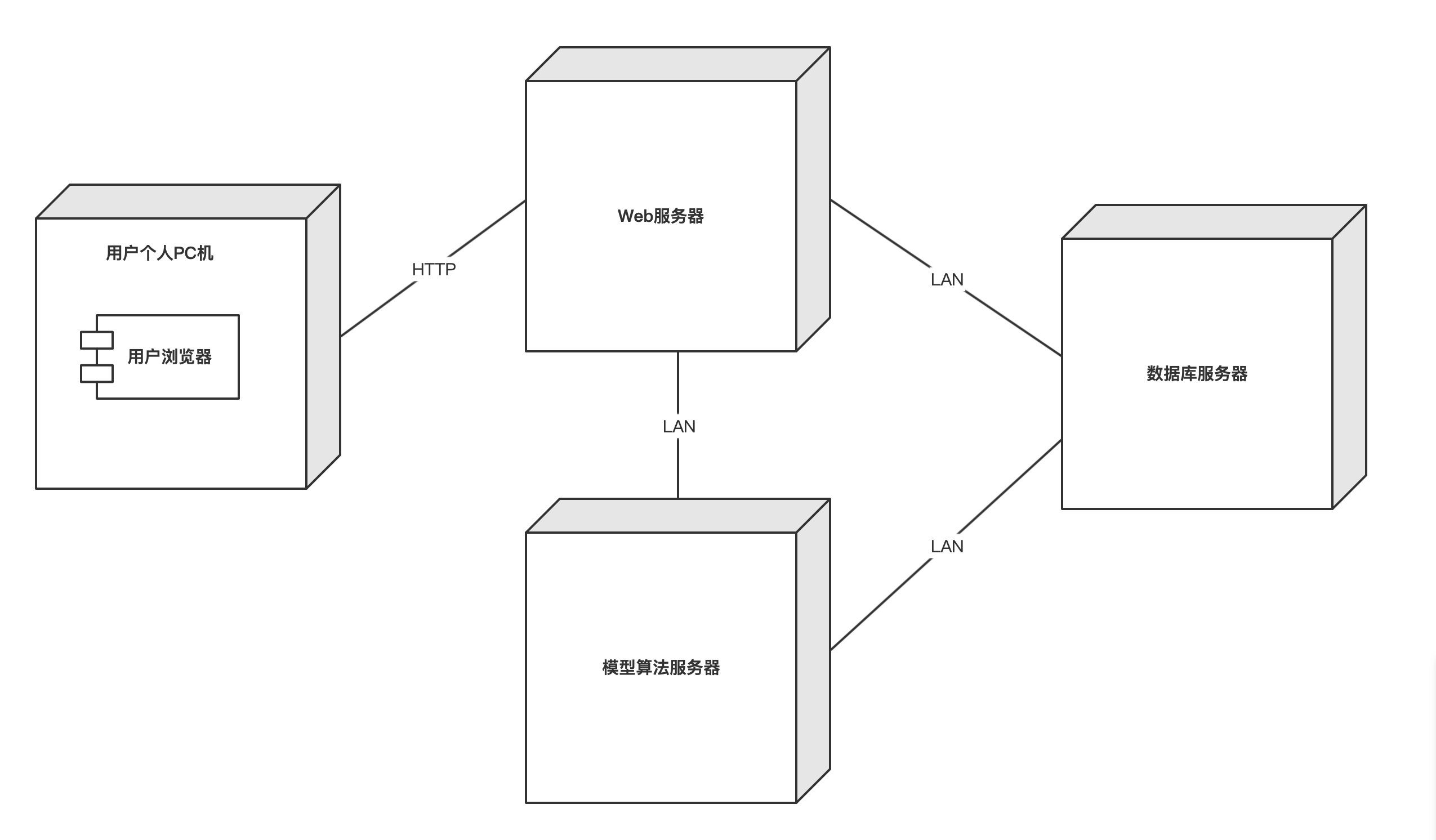


图 1-4 智能问答系统 部署图

**③详细设计：**

在详细设计阶段，基于智能问答系统体系结构和分析模型，对问答系统进行详细设计。从智能问答系统的用户需求和用例场景中发现类，根据需求可得该系统中的主要类中，实体类为：用户、回复系统、消息；辅助类为：召回层、精排层；界面类为：聊天框界面。然后，根据详细设计建立类之间的关系，并画出对应的类图反映系统静态结构设计、顺序图反映系统的动态结构设计，分别如下图1-5和图1-6所示：

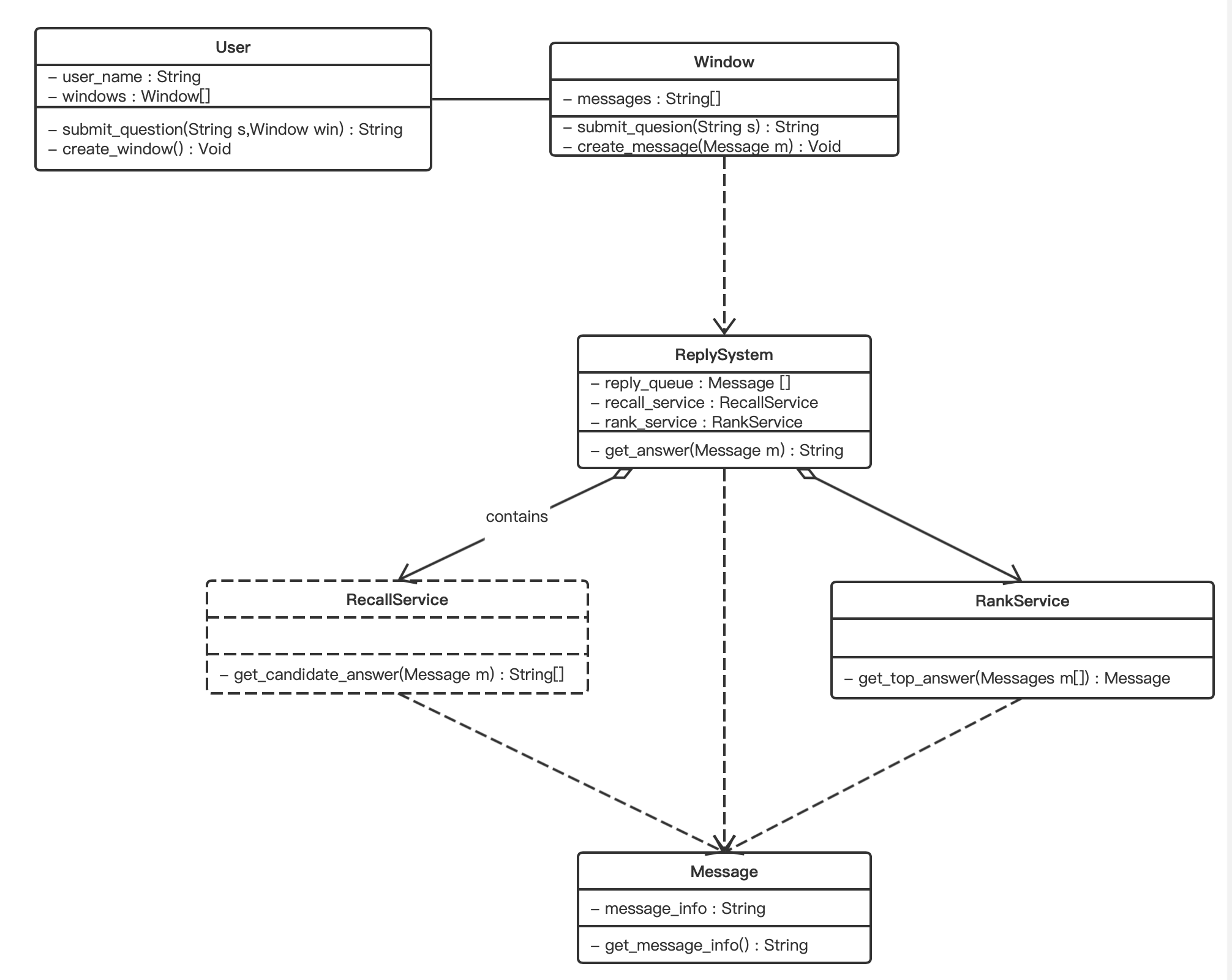


图 1-5 智能问答系统 类图

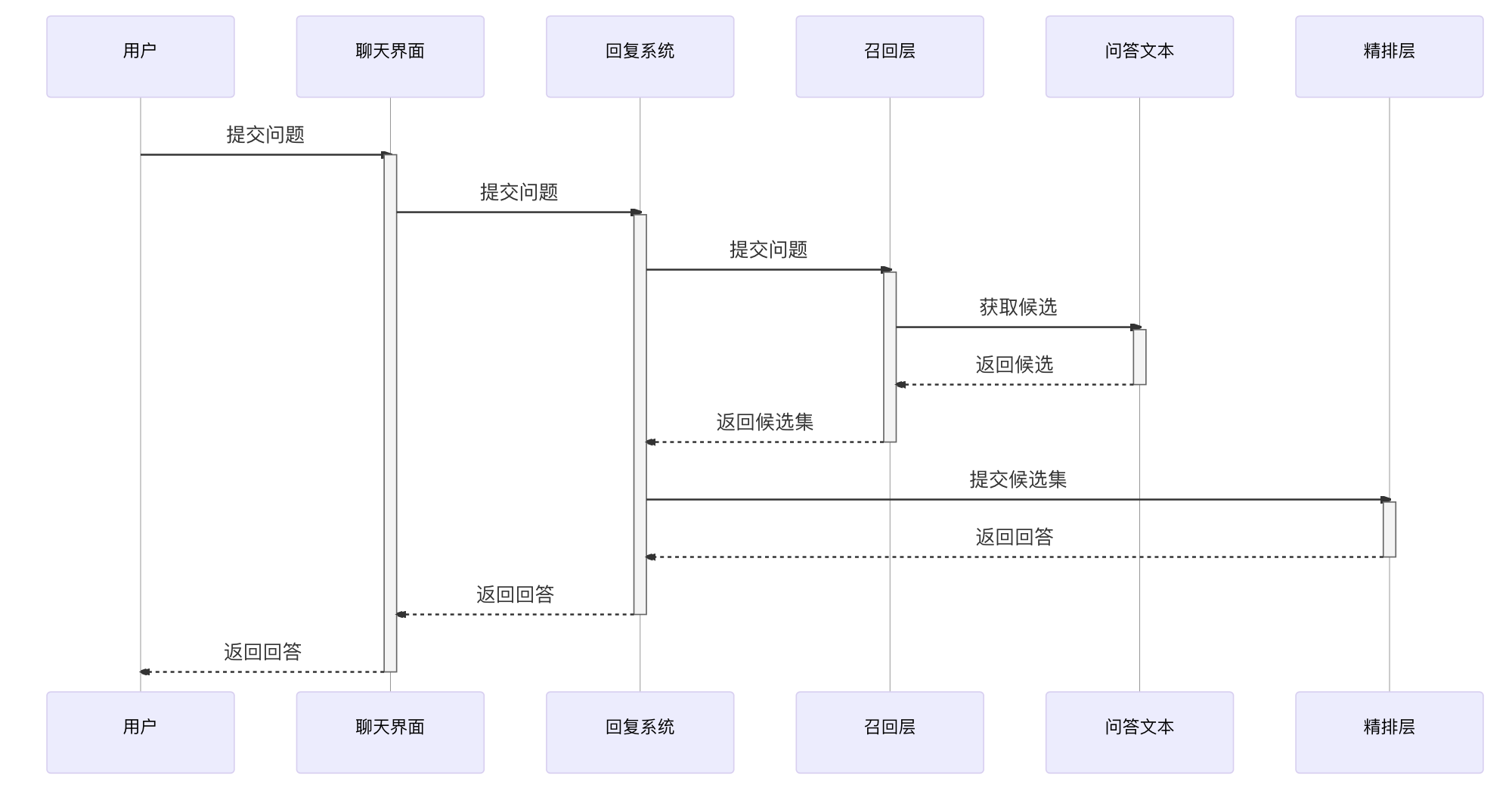


图 1-6 智能问答系统 时序图

**④代码编写：**

在代码编写阶段，根据详细设计的系统静态结构设计与动态结构设计的要求，目前主要完成了部分问答系统界面前后端交互模块、以及算法部分的召回层模块，保证了目前课题所涉及的问答系统有一个基本可交互的功能。

## 1.2. 知识技能学习情况

目前个人职业定位为算法工程师，同时针对个人职业发展需求，在NLP方向进行了一定程度上的知识研究。另外，根据课题应用场景，对毕设课题执行过程中以及在个人职业发展中，所运用的知识与工具进行了系统性地总结归纳，具体学习情况如下：

**①课题算法理论部分学习：**

排序层所用算法为BERT[1]，BERT的全称为Bidirectional Encoder Representation from Transformers，是一个预训练的语言表征模型。它强调了不再像以往一样采用传统的单向语言模型或者把两个单向语言模型进行浅层拼接的方法进行预训练，而是采用新的masked language model（MLM），以致能生成深度的双向语言表征。BERT论文发表时提及在11个NLP（Natural Language Processing，自然语言处理）任务中获得了巨大突破。考虑到该模型有两种优点：其一是采用MLM对双向的Transformer[2]进行预训练，以生成深层的双向语言表征；其二是在预训练后，只需要添加一个额外的输出层进行fine-tune，就可以在各种各样的下游任务中取得较好的表现。在这过程中并不需要对BERT进行任务特定的结构修改。基于模型的如上两点优势与课题任务目标的契合，同时又考虑到模型的训练和推理速度，决定使用BERT来充当智能问答系统排序层的任务。模型结构如图1-7所示：

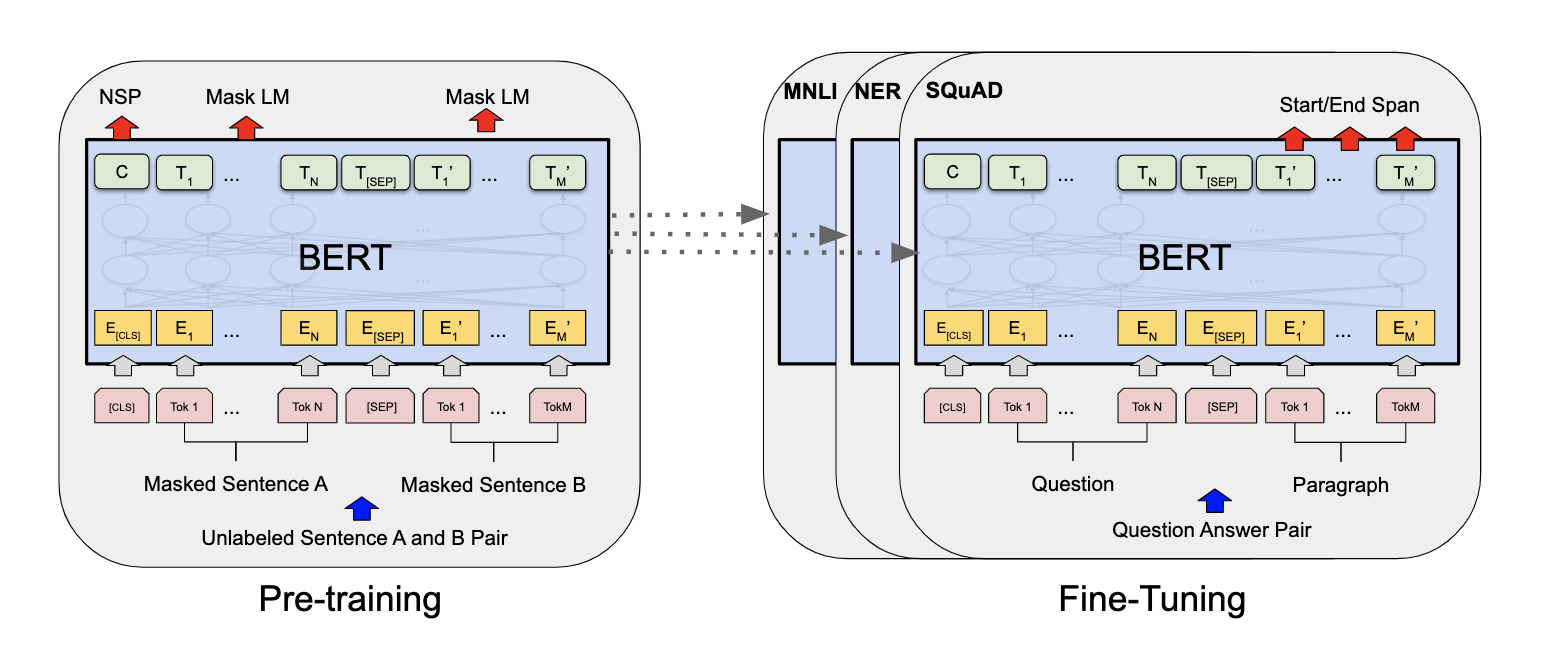


图 1-7 BERT模型结构

召回层所用算法为BM25[3]，BM25的全称为Best Matching 25，是一个基于词袋的检索模型，同时也是信息索引领域用来计算query与文档相似度得分的经典算法，它提出了词频和相关性之间的关系是非线性的，也就是说，每个词对于文档的相关性分数不会超过一个特定的阈值，当词出现的次数达到一个阈值后，其影响就不在线性增加了，而这个阈值会跟文档本身有关。因此，考虑到该算法有两种优点：其一是词袋检索模型相对于深度学习模型计算量小；其二是该算法是基于统计学推导而来，模型可解释性高。基于模型的如上两点优势与课题任务目标的契合，决定使用BM25算法来作为智能问答系统召回层的任务。算法公式如图1-8所示：

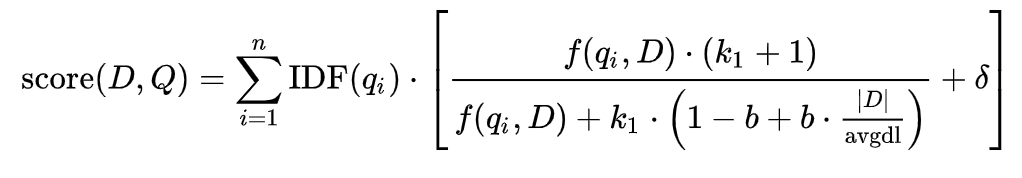


图 1-8 BM25算法公式

**②课题工程实现部分学习：**

前端语言部分使用的为HTML、CSS、JavaScript。HTML元素是构建网站的基石。HTML允许嵌入图像与对象，并且可以用于创建交互式表单，它被用来结构化信息——例如标题、段落和列表等等，也可用来在一定程度上描述文档的外观和语义。浏览器使用HTML标签和脚本来诠释网页内容；CSS最重要的作用是将文件的内容与它的显示分隔开来，CSS使开发者可以将这些信息中的大部分隔离出来，从而简化HTML文件；JavaScript作为客户端脚本语言，拥有可以直接在用户的浏览器上运行，不需要服务器的支持的优点。同时，可以减少对服务器的负担,同时JavaScript对于开发者来说有跨平台、容易上手等优势。

后端使用的是由Python编写的轻量级服务端框架Flask。它的优点有：使用简单，可扩展性强，能增加其他功能；同时Flask保留了扩增的弹性，可以用Flask插件加入所需功能，如ORM、窗体验证工具、文件上传、各种开放式身份验证技术等。

## 1.3. 职业素养学习培养

在课题任务实施过程中，也对个人职业素养进行了一定的学习培养，保证在个人职业发展中始终保持行业人才的专业性。作为一名软件工程师，个人在职业素养学习做了一些总结，总结如下：

1. 在职业道德与规范方面：首先，不可故意破坏软件功能，专业的程序员有责任让它无线接近零，而且发送明知有缺陷的代码，这种做法是极其不专业的；其次，不要随意破坏软件代码结构，要坚持代码设计原则；再者，需要对自己的专业能力充满自信，并因此勇于承担有把握的风险，也要清楚自己的自负；同时，信守承诺是一名严谨负责的专业软件开发人员应该需要具备的素质；最后，作为一名软件工程师，需要始终维护国家法律法规，不越过红线滥用用户信息，必须保证所开发的软件不会对社会造成不良影响甚至危害。
2. 在软件编程与职业健康方面：首先，需要保证代码要能正常工作、解决问题、跟现有系统结合、具有高的可读性；其次，确保将睡眠、健康和生活方式调整到最佳状况，这样才能在工作中全力以赴，产生优秀的代码；同时，作为专业的软件工程师，需要广泛地阅读，学习各种各样的知识，才能能激活人的创造力和增加很多灵光一闪的机会；最后，要适当分配精力，不要一直困在解决问题的状态中，可以暂时从问题中脱离出来，有助于在职业发展中以不同且更具创造性的方式搜寻解决方案。

# 第二章 存在问题与解决方案

## 2.1. 存在的主要问题

首先，课题任务是要求在垂类业务下，用户并在可交互的页面，输入一个Query/Question，系统能够提取Query、Answer的语义&字面信息，给出最满足Query需求的Answer。根据课题任务，总结出了在毕设执行中存在的一些问题，具体问题描述如下：

1. 基础数据清洗问题：

根据课题任务，目前是选用“Chinese medical dialogue data”数据集[4]，来做医疗垂类下的智能问答系统。但是该数据集中由于是用户真实问答文本，会含有诸多特殊中文符号、无意义字词、乱码、文本缺失等，不可直接供于问答系统的召回算法和排序模型使用。如何清洗与处理数据，直至召回算法与排序模型可直接使用，同时保证预期输出功能正常，是一个课题中的问题点。

1. 基础数据存储问题：

目前面临候选文本集合大，无法全部加载入内存的现状，那么如何将候选文本数据集合，通过使用高效的索引结构，进行合理的存储，来保证在召回阶段和排序阶段的正常读写功能，是课题的一个问题点。

1. 核心模型训练问题：

对BERT模型进行fine-tuning时，所占的内存大小资源比较大，同时在工程上如何协调模型训练时长和效果是一个问题；另外，如何定义两条不同文本的语义相似度，才能更好地适配课题任务目标，也是一个问题点。

1. 核心模型性能问题：

在召回阶段，需要将所有候选文本集合与用户在问答系统交互页面上真实输入的中文文本进行BM25 Ranking，如何高效的对可能是目标文本的集合进行Ranking 是课题的一个问题点。

在排序阶段，考虑到BERT模型参数量巨大，在预测阶段的耗时是所有模块中最高的，如何能够在正常的用户容忍时间内，对召回服务返回的所有候选进行排序，并保证问答系统的可用性，是课题的一个问题点。

1. 系统交互体验问题：

在前后端交互阶段，用传统的HTML表单方式进行页面的更新提交会产生一个问题——每次都要将请求提交到服务器，待服务器返回后再重新渲染界面，这样界面就会经历︰提交→变白→重新显示这样一个过程，会导致用户体验非常差，这是课题的一个问题点。

## 2.2. 解决方案与可行性研究

针对前文提到的在课题任务执行中所存在的问题，在安全、法律的约束条件下，根据自身所学以及一些资料，分别对其进行了解决方案的拆解以及对应方案的可行性研究，整体保证了项目课题执行的合理性与可行性。

针对具体问题的解决方案与可行性研究如下：

1. 基础数据清洗方案：

针对前文提到的数据清洗问题，分析的结论为两点：首先，去掉停用词和特殊符号，保留“你、我”等一些在此场景中有实际意义、有区分度的代词等；其次，对于有缺失以及乱码的文本进行丢弃，保证训练数据的正确性与可用性。

综上所述，该方案可以解决如上的数据清洗问题。

1. 基础数据存储方案：

针对前文提到的数据存储问题，分析的结论为使用MySQL、SQLite等关系型数据库存储技术对候选文本集合进行数据持久化操作。该方法的优点在于：首先，程序代码重用性强，即使更换数据库，只需要更改配置文件，不必重写程序代码；另外，持久化技术可以自动优化，以减少对数据库的访问量，提高程序运行效率。

综上所述，该方案可以解决如上的数据存储问题。

1. 核心模型训练方案：

针对模型训练时长和效果的平衡问题，采用Early Stopping方法，即当模型的损失函数更新幅度低于某阈值之后，此时可认为模型已经收敛。该方法的优点在于：首先，可以大幅度减少无用的训练时长，提高模型交付速度；另外，该方法在一定程度上可以防止过拟合，提高模型的泛化能力。

综上所述，该方案可以解决如上的模型训练问题。

1. 核心模型性能方案：

针对召回阶段，离线设计一个倒排索引与一个正排索引，两者均用k-v存储的形式。其中，倒排的key设计为使用中文jieba[5]分词后的token，value为包含该token的候选文本集合，这样设计可以使得摒弃大量与用户真实问句无token交集的计算，提高在线计算效率；正排的key设计为候选文本，value为其token：token frequency的映射，这样设计可以避免在线进行中文分词，提高在线计算效率。

针对排序阶段，发现该计算场景为I/O密集型的场景，那么可以利用微服务架构的思想，将fine-tuning 后的BERT模型包装成一个服务，服务对外多个计算接口，同时对该服务进行并发请求，保证在排序阶段不会让I/O阻塞过长时间。

综上所述，该方案可以解决如上的模型性能问题。

1. 系统交互体验方案：

针对前文提到的交互体验问题，可以通过AJAX方法来请求后台服务，本质上它的实现是在JavaScript中使用XMLHTTPRequest进行HTTP的请求，在实际开发中课题选择使用JQuery 提供的AJAX 功能。但同时，AJAX 也带来了无法跨域请求的问题——也就是无法在页面中向和当前域名不同的页面发送请求，针对新引入的问题，可以使用在当前页面所在的域的服务端做代理页面的方式来解决。

综上所述，该方案可以解决如上的交互体验问题。

# 第三章 当期任务完成度与后续实施计划

## 3.1 当期任务完成度

目前课题工作已经完成了需求分析、总体设计、详细设计以及部分算法核心代码、前后端交互代码编写，总完成度约为60%。后续还需进行部分算法核心代码编写、完整的系统测试和调试以及最终论文算法的撰写。

同时目前个人职业定位为算法工程师，因此针对个人职业发展需求，还需在毕业设计完成的过程中，总结与反思在课题设计中所遇到的算法或工程问题，并学习更多的有关NLP方向的算法模型、计算机网络通信、大数据方向等知识，促进个人健康的职业发展生态，扩充个人的职业选择空间。

## 3.2 后续实施计划

针对课题未完成工作内容，已制定了对应的后期的实施计划以及后期任务计划表，具体计划内容参见如下表1-1：

表 1-1 后期任务计划表

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 工作内容 | 工作开始时间 | 工作结束时间 |
| 1 | 完成算法核心代码编写 | 4.5 | 4.15 |
| 2 | 完成完整的系统测试和调试 | 4.16 | 4.22 |
| 3 | 最终论文的撰写 | 4.23 | 5.5之前 |
| 4 |  |  |  |

# 参考文献

1. Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova.“BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”.[C].North American Chapter of the Association for Computational Linguistics .2018
2. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin.“Attention is All you Need”.[C].Neural Information Processing Systems .2017
3. Stephen Robertson and Hugo Zaragoza.“The Probabilistic Relevance Framework”.[M].2009, 347-370.
4. Toyhom.“Chinese medical dialogue data 中文医疗对话数据集”.https://github.com/Toyhom/Chinese-medical-dialogue-data.[CP]
5. Neutrino.“‘结巴‘中文分词：做最好的 Python 中文分词组件：.https://github.com/fxsjy/jieba.[CP]