**电 子 科 技 大 学**

UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

**学士学位论文**

**BACHELOR THESIS**



论文题目  **基于BERT和BM25的智能问答**

**系统设计与实现**

学 院 信息与软件工程学院

专 业 **软件工程**

学 号 **2018091620013**

作者姓名 **廖梓尧**

指导教师 吴劲

摘 要

随着社会的高速发展，互联网规模发展的日趋庞大，人们对信息获取需求暴增。特别的，在后疫情时代，人们在医疗健康方面的需求日益增加，通常搜索引擎是人们用来满足自身对医疗问答信息需求的主要渠道。而在当前，国内的搜索引擎针对于用户提出的医疗类问题，缺乏文本的语义理解过程，通常只会展示与问答文本中几个包含关键词的网页链接，导致无法直观且有效地给出精准的回答，搜索引擎的结果难以保证权威性和准确性。

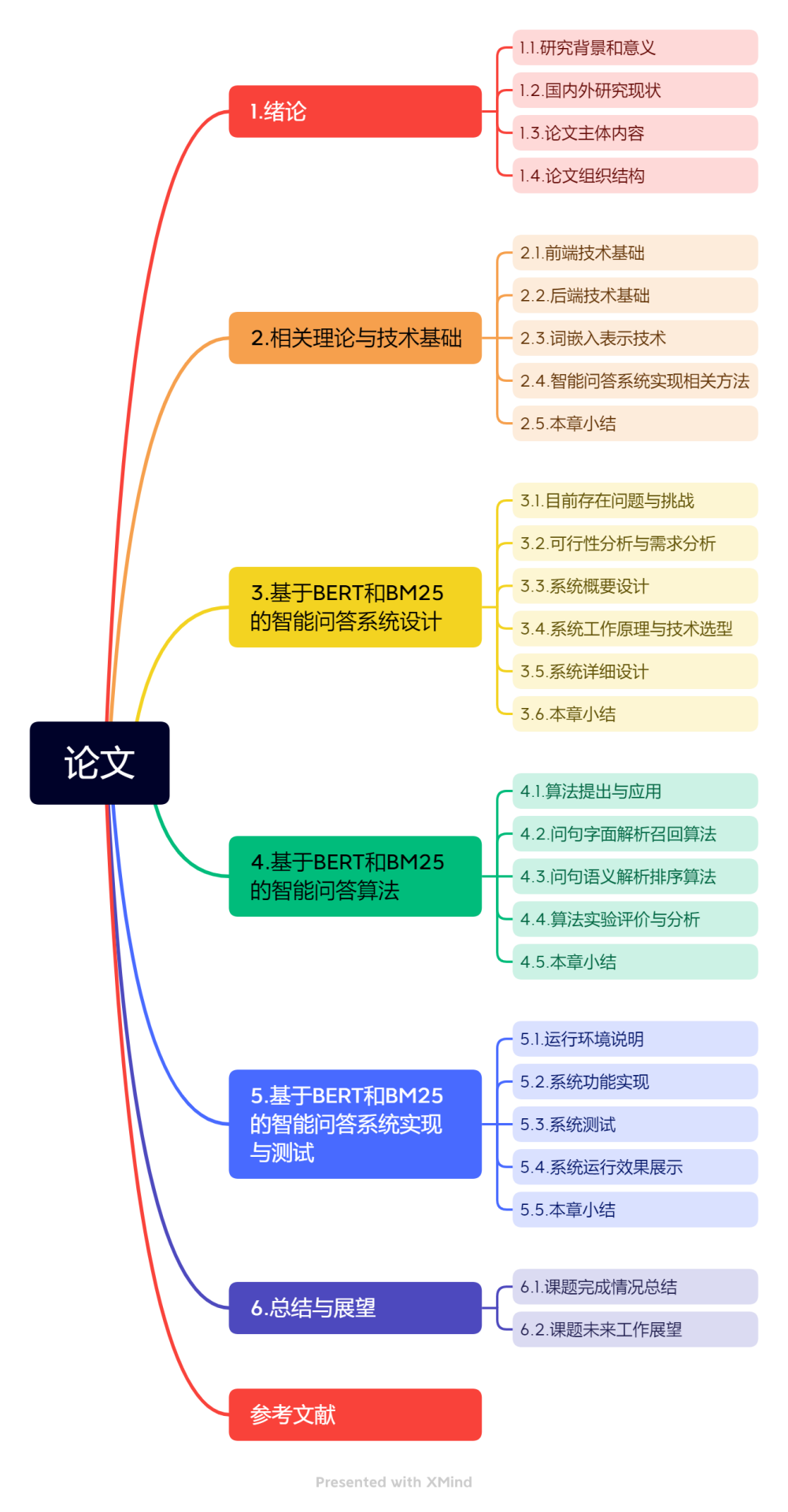
在大数据与人工智能的时代，及国家智慧医疗项目推进的背景下，本文主要的目的是实现医疗垂类下的智能问答系统，旨在为人们提供精准且可靠的医疗健康智能问答服务。 本文提出了一种基于BERT和BM25的智能问答系统设计方案，重点研究了智能问答算法的系统核心支撑部分，并通过实验论证了该套算法方案的可行性，同时在系统实现过程中融入了软件工程相关核心理论，最后以软件的形式来呈现该智能问答系统，本文的研究与实现过程的内容分为三部分：

1. 本文从系统架构方面进行了技术选型。提出了一个基于BERT和BM25的智能问答系统设计方案，可以精准分析用户在医疗健康领域的提问，满足了用户医疗数字化信息的需求。
2. 本文提出了一个基于BERT和BM25的智能问答算法。该算法引入了BERT（Bidirectional Encoder Representations from Transformers）预训练模型以及BM25（Best Matching 25）词袋模型，分别以语义解析的方式与字面解析的方式构建问答系统的排序层和召回层。本文提出的核心算法方案有效地提升了问句分析的能力。具体的，算法首先结合用户问句的字面分析方法，进行初步相似文本筛选，再通过语义分析方法进行最优排序，得到与用户问句最相似似的问题，最终在关系型数据库（MySQL）通过该问题查找到对应的答案。
3. 本文将基于BERT和BM25的智能问答算法融入到智能问答系统设计方案中，结合HTML、JQuery等前端技术及Flask等后端技术，在Web端进行系统实现。同时为了提高智能问答系统的专业性，系统使用“Chinese medical dialogue data 中文医疗对话数据集”作为问答系统的知识库，经测试，系统性能稳定，交互体验良好，能够达到人们对日常医疗保健相关问题的查询需求满足的标准。

**关键词：**智能问答系统，语义解析，字面解析，BERT，BM25

Abstract

目录



第一章 绪论

1.1研究背景和意义

1.2 国内外研究现状

1.3 论文主体内容

1.4 论文组织结构

第二章 相关理论与技术基础

本章将着重对基于BERT和BM25的智能问答系统设计与实现中涉及到的相关理论与技术基础进行较为详细的介绍。具体的，2.1节将会介绍前端开发技术基础，2.2节将会介绍后端开发技术基础，2.3节将会介绍词嵌入表示技术的相关理论，2.4节将会介绍智能问答系统在实现层面的相关方法，2.5节将会对本章进行一个总结。

2.1 前端开发技术基础

前端开发（Front-end development）技术涉及到HTML、CSS、JS、JQuery。本节将会对提到的这些技术进行一一介绍。

HTML（HyperText Markup Language），中文全称是“超文本标记语言”。它是一种开发者用于创建Web网页的标记语言，它可以将文本以及文本相关的信息都结合起来，并展现出与该网页结构和数据相关的计算机文字编码。首先，其HTML元素在网站构建中的地位如网站的骨架、基石一般，通常在网络浏览器运行的过程中，会利用读取开发好的HTML文件，同时浏览器会通过渲染的方式来使得网页可视化；其次，HTML语言拥有一个网站的结构语义的功能，也正是这个原因，使之的定位成为一种标记语言，而并非编程语言；另外，HTML也额外提供了允许嵌入图像与对象的功能，并且可以用于创建交互式表单；再者，它被用来结构化如列表、正文、标题等信息，在一定程度上描述了文档的外观和语义；最后介绍一下HTML的语言形式，其为由尖括号包围的元素（如<html>），并且浏览器在运行过程中，是不会将它们在页面上显示的。

CSS(Cascading Style Sheets)，中文全称为“层叠样式表”。CSS目前是一门由W3C定义和维护的计算机语言，它的主要功能是添加字体、间距和颜色等样式至于结构化文档（如HTML文档或XML）——具体而言就是会在HTML开发的网页基础上，使得交互页面更加美观，起到提升用户交互体验的功能。

JS（JavaScript），是一种解释型、即时编译型的Web编程语言。首先，JavaScript支持命令式、面向对象以及声明式（如函数式编程）的编程风格；另外，它的主要作用是用来控制DOM元素组件以及页面上的各种事件触发——具体而言就是会在HTML、CSS开发的网页基础上，控制与用户与界面进行动态交互的后续响应动作，是用户交互体验的核心。

JQuery是一个由JavaScript编程语言编写的跨浏览器的库，主要用途为简化HTML与JavaScript编程语言之间交互的操作流程。首先，JQuery拥有独特的语法设计，使得在前端中许多操作变得容易开发上手——比如处理事件、选择DOM元素、操作文档对象、创建动画效果以及开发Ajax程序等开发常用操作；同时，JQuery库也将创建插件的功能提供了给前端开发人员，这使开发人员可以抽象化动画、高级效果和高级主题化的组件;最后，不得不 提到JQuery函数库模块化的方式，使得它对动态网页的功能灵活性的以及基于web的应用程序提供了强大的支持。

2.2 后端开发技术基础

后端开发（Back-end development）技术涉及到Flask、MySQL。本节将会对提到的这些技术进行一一介绍。

Flask是一款以Python语言为基础编写的轻量级web后端框架，它主要适用于微小Web后端项目构建的领域，通常以精炼简洁，敏捷轻量著称。首先，在本课题研究中，会通过使用Flask框架进行实验研究，尤其在几十行代码编写后，Flask即可快速部署一个web服务，以此来验证课题工作中的一些假设；同时，Flask是一个入门及其友好的web后端框架，详细官方教程以及大量的技术参考文档，对于后端开发者而言，这些是非常好的学习资源，能保证快速上手以及本课题的正常开展；最后，Flask还提供提供轻量级的admin管理后台系统、缓存系统、数据库迁移功能、用户权限管理系统等各种功能，这些额外提供的功能，对新手无疑是非常友好的，也有助于本课题研究的正常推进工作。

MySQL是目前比较流行的关系型数据库管理系统，并且在web端开发这块来讲，MySQL 是最好的关系数据库管理系统软件之一。首先，MySQL会为课题所需要的数据建立不同的表，并将其保存在不同的表中——这样的存储方式与将所有数据放在一个大仓库内相比，提高了访问数据的速度以及查询数据的灵活性；其次，MySQL使用标准的SQL语言来进行增删查改的数据操作，没有额外的学习成本；同时，MySQL可以运行于多个系统上，如Windows、Centos、Debian等，也支持多种编程语言，包括 C、C++、Python、Java等，对于本课题研究来说，各个方面都具有良好的兼容性，因此在本课题中选择了MySQL做数据库管理的应用软件。

2.3 词嵌入表示技术

在自然语言处理（NLP）领域中，通常会对中文字符使用词嵌入的技术来进行文本表示，主要做法是将中文字符映射到一个巨大的向量空间中，以一个数学向量为载体，表达该中文字符的文本信息，这也称之为“词嵌入向量”。早期通常会将中文字符进行离散化操作，这样的词嵌入表示方法称之为one-hot向量。之后在信息检索领域发展出了词袋模型，这类模型会将文本视为可以用一个装着这些词的袋子来表示，这种表示方式的核心思想是不考虑词的语法关系以及词之间的顺序关系，BM25模型就是其中之一。随着NLP技术的不断发展，又有学者提出了多种静态词向量的表示方法，比如谷歌公司提出来的Word2Vec方法。近几年，伴随着预训练思想的横空出世，一系列基于预训练方法得到的动态文本词向量方法百花齐放，而BERT模型就是最著名的预训练模型之一 ，促使了词嵌入文本表示技术发展到了新的一个阶段。

1. One-hot

One-hot编码，中文名称为“独热编码”，是一种经典的离散化编码方式，主要做法是将分类变量作为二进制向量的表示——具体而言，首先要求将分类文本字符映射到整数值；然后，将每个整数值表示为二进制向量，其中除了整数所在位置的索引被标记为1之外，它都是零值。以性别特征为例，当一个样本为[“男”]的时候，它的文本独热编码为[1，0]，而为[“女”]的时候，它的文本独热编码为[0，1]。这种编码的好处是将文本字符离散特征的取值扩展到了欧式空间，而离散特征的某个取值就对应欧式空间的某个点，这种做法会让特征之间的距离计算更加合理，更具有可解释性。

在本课题中，尝试过对医疗垂类领域的中文文本语料使用One-hot编码，但由于此类文本的数据特点为词汇复杂且数量巨大，首先需要利用完整的中文汉语字词库来构建词典，最终得到词汇的数量高达6万左右，且随着中文语料库的不断扩张，产生的One-hot向量会造成不可想象的维度灾难，这对于目前有限的空间存储资源来说是不可接受的，同时这种编码会有许多语义信息损失，这对于课题预期的最终效果来讲也是不能接受的。

1. BM25

BM25，英文全称为“Best Matching 25”，其中BM 是Best Matching最佳匹配的缩写，25指的是第25次算法迭代。在信息索引领域，BM25仍然是目前最主流的计算中文文本相似度得分的一个词袋算法模型，其主要做法可以拆解为两个部分——首先，BM25会考虑到中文文本里每一个词在整体文档里的重要性，这部分是由IDF（Inverse Document Frequency，逆文档频率）改进而来的；其次，会考虑到中文文本里每一个词在当前文本的一个重要性，这部分是由TF（Term Frequency，词频）改进而来的，并且还考虑到了文本本身长度对词频的影响。整合这两个部分之后就可以使用离散化的BM25向量来作为中文文本的词嵌入表示，具体计算过程**【如公式x-xx】**所示。另外，这种嵌入向量的可解释性与可用性相对于One-hot独热编码来说是更加合理的。

本课题尝试使用BM25词袋模型对整个中文医疗文本进行了编码处理，这种编码从效果上来讲仍然存在对垂直领域的文本语料适用性一般，且存在一定的OOV（Out of Vocabulary）问题，但是整体上对于课题的最终目标效果比较符合，且考虑到其存储空间资源和运算时间性能也较好，因此可将BM25词袋模型算法作为本课题核心算法之一。

1. Word2Vec

Word2Vec，是由谷歌公司提出的一种基于深度学习的产生静态词嵌入向量的语言模型，是文本表征技术的一大里程碑。Word2Vec提出了两种不同的神经网络模型来训练词向量，一种是CBOW（Continuous Bag-of-Words，连续词袋模型），其核心思想是利用上下文的词来预测中心词，然后利用极大似然法去做模型训练；另一种是Skip-gram（跳字模型），其核心思想与CBOW相反，它是利用中心词去预测上下文的词，同样也是利用极大似然法去做模型训练。Word2Vec的模型训练结构**【如图x-xx】**所示，滑动窗口的大小设置为2，其中w（t）表示中心词的文本表示向量，其余则表示围绕中心词的上下文词的文本表示向量。

对于本课题而言，Word2Vec相对于传统的词袋模型或者独热编码的文本表征能力更强，在一定程度上可以考虑到中文词之间的上下文关系，但是在医疗垂类领域下，无法针对本课题的具体任务进行单独优化，并且表现情况与课题预期目标效果还有一定的差距，在实际使用上不太符合课题研究要求。

1. BERT

BERT（Bidirectional Encoder Representations from Transformers），中文全称为“基于变换器的双向编码器表示技术”，该预训练模型由谷歌公司在2018年提出，是NLP领域发展的一个重大里程碑，在自然语言理解的多项任务中取得重大突破，成为了目前自然语言处理领域实验中无处不在的基线。BERT预训练模型的结构**【如图x-xx】**所示，若以中文为基本训练语言，只需要将中文文本以“字粒度”输入至BERT模型中，就可以得到整个文本的动态嵌入向量表示。其核心结构是利用Transformer模型的多个Encoder堆叠而成，通过注意力机制对文本进行建模，具体采用的是多头缩放点积注意力结构(muti-head attention & scale dot product)，具体计算过程**【如公式x-xx】**所示。其中Q代表查询向量，K代表键向量，V代表值向量，dk代表查询向量的维度，该公式通过SoftMax函数计算得出一个注意力权重分布矩阵，通过加权求和后得出新的词的动态语义向量表示。

这种训练方法，首先能够有效解决文本中的长距离信息依赖问题；同时，还可以根据上下文不同，表层词语的多语义性——例如“苹果手机”与“苹果树”里的“苹果”，在Word2Vec的静态向量表征中是无法区分的，但在BERT里是可以通过动态向量表征来很好的解决这种一词多义的问题的。

对于本课题而言，首先由于BERT的结构设计使得课题对于文本的处理成本较低，一定程度上提高了课题研究的可行性；另外，BERT还能够根据不同领域的语料进行专门的优化，这与课题目标高度吻合；此外，BERT的初步表现情况比较能满足课题的预期，因此在课题实际使用中，可将BERT预训练模型算法作为本课题核心算法之一。

2.4 智能问答系统相关方法

通常智能问答系统的主流实现均为基于深度学习、神经网络的方法，目前按照“使用技术方向”可以分成三类**【插入参考文献华为的】**，分别是基于检索式的智能问答系统、基于知识图谱的智能问答系统以及基于机器阅读理解的智能问答系统，具体介绍如下：

1. 基于检索式的智能问答系统：

这类问答系统通常称之为FAQ（Frequently Asked Questions），中文直译为“常见问题解答”。它的主要做法是通过在已有的“问题-回答”对集合中，找到与用户提问最为相似或者匹配的问句，并将该问句对应的回答作为最终答案返回给用户。这类问答系统的优点在于回答的覆盖率较高，同时对于未收录的问题也能靠字面语义匹配取得不错的效果；但是也有比较明显的缺点，比如排序后为第一的答案不一定与用户原问题匹配，这会导致返回结果效果受损，而且对于训练语料的要求也较高。

1. 基于知识图谱的智能问答系统：

这类问答系统通常称之为KBQA（Knowledge-Based Question Answering），中文直译为“知识库问答”。它的主要做法是通过构建知识库，并从知识库中搜索与问题相关的实体、关系或属性作为答案返回给用户。这类问答系统的优点在于可以回答知识层面的推理类问题，能够为问题的语义理解提供丰富的背景知识，同时回答准确率较高；但是也有比较明显的缺点，就是构建知识图谱的成本比较高，需要大量人工去标注数据或者高质量的数据集，同时限定知识图谱中必须存在和问题匹配的内容，否则回答不了。

1. 基于机器阅读理解的智能问答系统：

这列问答系统通常称之为MRC（Machine Reading Comprehension），中文直译为“机器阅读理解”。它的主要做法是根据给定的上下文回答问题，例如中高考中的英文阅读理解题目。这类问答系统的优点在于它的回答文本获取容易，不需要进行额外的文本结构化，因此大大减少了对文本数据做结构化处理的时间和人力成本；但是也有比较明显的缺点，就是这类系统需要大量人工标注的数据，以及一个强力的文本理解模型，同时限定了答案必须在文本中出现的条件，并强调必须是文本中的连续片段。

2.5 本章小结

本章主要讲解了基于BERT和BM25的智能问答系统设计与实现中涉及到的相关理论与技术基础。首先介绍了前端开发技术基础，其次是后端开发技术基础，另外也对词嵌入表示技术的相关理论进行了详细的阐述，最后在智能问答系统在实现层面的相关方法进行了展开。综上所述，以上所提及的内容奠定了基于BERT和BM25的智能问答系统设计与实现的技术基础与理论基石。

第三章 基于BERT和BM25的智能问答系统设计

本章将着重对基于BERT和BM25的智能问答系统的具体设计进行较为详细的介绍。具体的，3.1节将会介绍目前在系统功能层面和系统算法层面存在的问题与挑战，3.2节将会进行详细的系统需求分析，主要从可行性分析、功能性需求分析以及非功能性需求分析三个方面进行叙述，3.3节将会进行系统的概要设计，分别从系统的架构设计与系统的模块设计进行详细阐述，3.4节将会介绍本系统的工作原理以及最终的技术选型，3.5节将会介绍系统的详细设计，3.6节对本章进行一个总结。

3.1存在问题与挑战

3.2 可行性分析与需求分析

3.3 系统概要设计

3.4 系统工作原理与技术选型

3.5 系统详细设计

3.6 本章小结

第四章 基于BERT和BM25的智能问答算法

为了保证智能问答系统能够达到人们对日常医疗保健相关问题的查询需求满足的标准，本章引入了字面解析和语义解析的思想，提出了一种基于BERT和BM25的智能问答算法。具体的，4.1节将介绍算法的提出与应用，会整体介绍算法方法中的各个模块内容，4.2节主要针对字面解析召回模块，讲述字面解析算法的流程，4.3节从模型训练以及模型预测两个方面，来参数语义解析排序模块的算法流程，4.4节将会对本章提出的算法方法进行验证并完成结果分析，4.5节对本章内容进行总结。

4.1算法提出与应用

4.2 字面解析召回算法

4.3 语义解析排序算法

4.4 算法实验评价与分析

4.5 本章小结

第五章 基于BERT和BM25的智能问答系统实现与测试

基于BERT和BM25的智能问答系统由数据处理模块、基于BERT和BM25的智能问答算法模块以及用户交互模块构成，本章将详细介绍三大模块的功能实现与功能测试过程。具体的，5.1节将会介绍系统所在的运行环境，5.2节将会介绍系统主要的模块功能实现过程，5.3节将从智能问答系统的功能性测试和非功能性测试两方面进行测试流程的介绍，5.4节将会进行系统运行效果的展示，5.5节对本章内容进行总结。

5.1 运行环境说明

5.2 系统功能实现

5.3 系统测试

5.4 系统运行效果展示

5.5本章小结

1. 总结与展望

6.1 课题完成情况总结

随着互联网行业的高速发展与规模的日趋庞大，网络完全融入了人们生活的各个方面，人们对信息获取需求暴增，产业转型数字化的进程也在不断推进。在后疫情时代，人们在医疗健康方面的需求日益增加，而国内的搜索引擎针对于用户提出的医疗类问题，无法直观且有效地给出精准的回答。基于国家智慧医疗项目推进的背景下，本文针对医疗垂类下的智能问答进行了系统研究，并最终完成了智能问答系统的设计与实现。当前本文已完成的主要工作如下：

6.2 课题未来工作展望

参考文献