

**《基于知识图谱的自动问答系统》**

**设计说明书**



北京航空航天大学

2016-01

版本变更历史

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 版本 | 提交日期 | 主要编制人 | 审核人 | 版本说明 |
| 1 | 2015.12.31 | 方凯 |  | 完成1、2(部分)、3(部分)、 |
| 2 | 2015.01.01 | 杨东东 |  | 完成2(部分)、3(部分) 、  4、5、6(部分)、7 |
| 3 | 2015.01.03 | 李睿霖 |  | 完成4（部分）、5（部分）、6（部分），修改整个文档 |
| 4 | 2016.01.18 | 杨东东 |  | 修改 |
|  |  |  |  |  |
|  |  |  |  |  |

目录

[1. 范围 1](#_Toc440903990)

[1.1标识 1](#_Toc440903991)

[1.2系统概述 1](#_Toc440903992)

[1.3文档概述 1](#_Toc440903993)

[1.4术语和缩略词 1](#_Toc440903994)

[1.4.1数据库(Database) 1](#_Toc440903995)

[1.4.2自然语言处理(NLP) 1](#_Toc440903996)

[1.4.3矩阵论(Matrix) 2](#_Toc440903997)

[2. 引用文档 2](#_Toc440903998)

[2.1书籍 2](#_Toc440903999)

[2.2批文 2](#_Toc440904000)

[2.3引用资料 2](#_Toc440904001)

[3. 需求概述 3](#_Toc440904002)

[3.1概述系统特性和需求 3](#_Toc440904003)

[3.2加细节 3](#_Toc440904004)

[3.2.1问题分析 3](#_Toc440904005)

[3.2.2信息检索 3](#_Toc440904006)

[3.2.3答案抽取 3](#_Toc440904007)

[3.3 新需求 4](#_Toc440904008)

[4. 体系结构设计 4](#_Toc440904009)

[4.1总体结构 4](#_Toc440904010)

[4.1.0用例图 5](#_Toc440904011)

[4.1.1活动图 5](#_Toc440904012)

[4.1.2时序图 6](#_Toc440904013)

[4.1.3状态图 6](#_Toc440904014)

[4.1.4类图 6](#_Toc440904015)

[4.2功能分配 7](#_Toc440904016)

[4.3关键问题及解决方案 7](#_Toc440904017)

[4.3.1问题分析模块及解决方案 7](#_Toc440904018)

[4.3.2问题分类模块及其解决方案 8](#_Toc440904019)

[4.3.3答案抽取模块及其解决方案 9](#_Toc440904020)

[5. 接口设计 10](#_Toc440904021)

[5.1用户界面设计 10](#_Toc440904022)

[5.2外部接口设计. 11](#_Toc440904023)

[5.2.1用户登录 11](#_Toc440904024)

[5.2.2用户注册 12](#_Toc440904025)

[5.2.3获得文本问题回答 12](#_Toc440904026)

[5.2.4提交信息 12](#_Toc440904027)

[5.3内部接口设计 13](#_Toc440904028)

[6.数据结构设计 14](#_Toc440904029)

[6.1公共数据结构设计 14](#_Toc440904030)

[6.2数据库设计 15](#_Toc440904031)

[6.2.1文档数据库 15](#_Toc440904032)

[6.2.2通用数据库表的设计 16](#_Toc440904033)

[7.详细设计 17](#_Toc440904034)

[7.1用户模块 17](#_Toc440904035)

[7.1.1概述 17](#_Toc440904036)

[7.1.2接口说明 17](#_Toc440904037)

[7.1.3内部设计（静态） 17](#_Toc440904038)

[7.1.4内部设计（动态） 17](#_Toc440904039)

[7.2语音模块 18](#_Toc440904040)

[7.3.1概述 18](#_Toc440904041)

[7.3.2接口说明 18](#_Toc440904042)

[7.3.3内部设计（静态） 18](#_Toc440904043)

[7.2.4内部设计（动态） 18](#_Toc440904044)

[7.3问题分析 18](#_Toc440904045)

[7.3.1概述 18](#_Toc440904046)

[7.3.2接口说明 19](#_Toc440904047)

[7.3.3内部设计（静态） 19](#_Toc440904048)

[7.3.4内部设计（动态） 20](#_Toc440904049)

[7.4 信息检索 20](#_Toc440904050)

[7.4.1概述 20](#_Toc440904051)

[7.4.2接口说明 21](#_Toc440904052)

[7.4.3内部设计（静态） 21](#_Toc440904053)

[7.4.4内部设计（动态） 21](#_Toc440904054)

[7.5答案抽取 22](#_Toc440904055)

[7.5.1概述 22](#_Toc440904056)

[7.5.2接口说明 22](#_Toc440904057)

[7.5.3内部设计（静态） 23](#_Toc440904058)

[7.5.4内部设计（动态） 23](#_Toc440904059)

# 范围

## 1.1标识

文档标识号：A2015-00-02-00

文档标题：<<基于知识图谱的自动问答系统>>--系统设计说明书

版本号：4.0

## 1.2系统概述

基于知识图谱的自动问答系统，是以中文为载体的系统，其数据库为以百度百科、维基百科、互动百科为主，运用其中的知识性信息进行人机交互以达到自动问答的目的的系统。采用目前发展中的实体分词技术、实体消歧技术、语法分析技术、语义分析技术等作为基础，综合开发而成。

系统尚在开发过程中；

投资方：无；

需方：用户；

用户：问答需求方；

开发方：杨东东，李睿霖，方凯；

运行现场：Android系统。

## 1.3文档概述

本文档用于阐述系统概况以及系统设计说明。该文档使用时保密性一般，由于未涉及核心代码，使用时可以半公开。

## 1.4术语和缩略词

### 1.4.1数据库(Database)

MySQL:关联[数据库管理系统](http://baike.baidu.com/view/68446.htm)

Neo4j:高性能的,NOSQL图形数据库

Redis:内存数据库，以高效的查找存储著称，并具有数据持久的特性

### 1.4.2自然语言处理(NLP)

Entity Linking:实体链接

Page Rank:Google开源的一个搜索算法

Entity Ambiguation:实体歧义

Trie Tree:前缀树

kNN: k近邻算法 (kNN，k-NearestNeighbor)

LSA: 隐式语义分析 (Latent Semantic Analysis)

Markov Model:马尔科夫模型

Lucene: 一个开放源代码的全文检索引擎工具包，是一个全文检索引擎的架构

FudanNLP：一个中国国内做得还算不错的NLP处理开源包

### 1.4.3矩阵论(Matrix)

PCA:主成分分析，用于矩阵维度的降维方法 (Principal Component Analysis)

SVD:矩阵奇异值分解 (singular value decomposition method)

[panion](http://cn.bing.com/dict/search?q=Companion&FORM=BDVSP6&mkt=zh-cn) [Matrix](http://cn.bing.com/dict/search?q=Matrix&FORM=BDVSP6&mkt=zh-cn)：伴随矩阵

[ompanion](http://cn.bing.com/dict/search?q=Companion&FORM=BDVSP6&mkt=zh-cn) [Matrix](http://cn.bing.com/dict/search?q=Matrix&FORM=BDVSP6&mkt=zh-cn)：友矩阵

# 引用文档

## 2.1书籍

《软件项目管理》 朱少民，韩莹 编著，人民邮电出版社

《软件项目管理》 Rajeev T Shandilya编著 科学出版社

## 2.2批文

本项目的经核准的计划任务书和上级机关的批文:第九届《大学生创新创业训练计划》

## 2.3引用资料

《<基于知识图谱的自动问答系统>系统需求规格说明书》

《Open Question Answering Over Curated and Extracted Knowledge Bases》

《syntactic constraints on paraphrases extracted from parallel corpora》

《基于维基百科的自动词义消歧方法》 史天艺

《一个中文实体链接语料库的建设》 舒佳根

《基于角色标注的中国人名自动识别研究》张华平 刘群

《基于层叠隐马尔可夫模型的中文命名实体识别》俞鸿魁 张华平 刘群

《基于角色标注的中文机构名识别》俞鸿魁 张华平 刘群

《基于最大熵的依存句法分析》 辛霄 范士喜 王轩 王晓龙

《社区问答系统中若干关键问题研究》 [廉鑫](http://www.cnki.net/kcms/detail/search.aspx?dbcode=CDFD&sfield=au&skey=%e5%bb%89%e9%91%ab&code=25126167;)

《汉语大词汇量连续语音识别系统研究进展》倪崇嘉 刘文举 徐波

# 需求概述

## 3.1概述系统特性和需求

用户发送文字信息给自动问答系统，然后通过自动系统的分析然后返回给用户尽可能使他所想要的信息。该系统从用户发送信息给客户，然后通过自动问答系统中的问题分析模块，信息检索模块和答案抽取模块共三个算法模板检索出准确率最高的答案返回给用户。能够与用户进行聊天，能够回答事实型问题、列举型问题、定义型问题和交互型问题等。

根据这些需求，问答系统需由三个模块组成：问题分析模块，信息检索模块和答案抽取模块。如何在问题分析阶段充分理解用户的提问意图，如何在信息检索阶段把相关文档检索出来，如何在答案抽取阶段准确地把答案从相关文档中抽取出来，这三个方面是问答系统要解决的核心，上图给出了问答系统的经典模型。用例中，用户发送文字信息给自动问答系统，然后通过自动系统的分析然后返回给用户尽可能使他所想要的信息。该系统从用户发送信息给客户，然后通过自动问答系统中的问题分析模块，信息检索模块和答案抽取模块共三个算法模板检索出准确率最高的答案返回给用户。

## 3.2加细节

### 3.2.1问题分析

问题分析是问答系统的第一个模块，它负责对用户提出的中文自然语言问题进行深入细致地理解和分析，为信息检索模块和答案抽取模块奠定良好的基础。问题分析阶段的处理结果直接影响着后续阶段的处理。该阶段的分析主要包括词法分析（分词、词性标注）、问题分类、关键词提取及关键词扩展。其中问题分类是此模块的核心，它能有效减少候选答案的空间，提高系统返回答案的准确率，同时问题的类型也决定了答案抽取中一些策略的选择。由于问句与陈述句的句型结构有着明显差别，为方便在答案抽取时进行相似度计算，必要时还需要对问句进行重写。

### 3.2.2信息检索

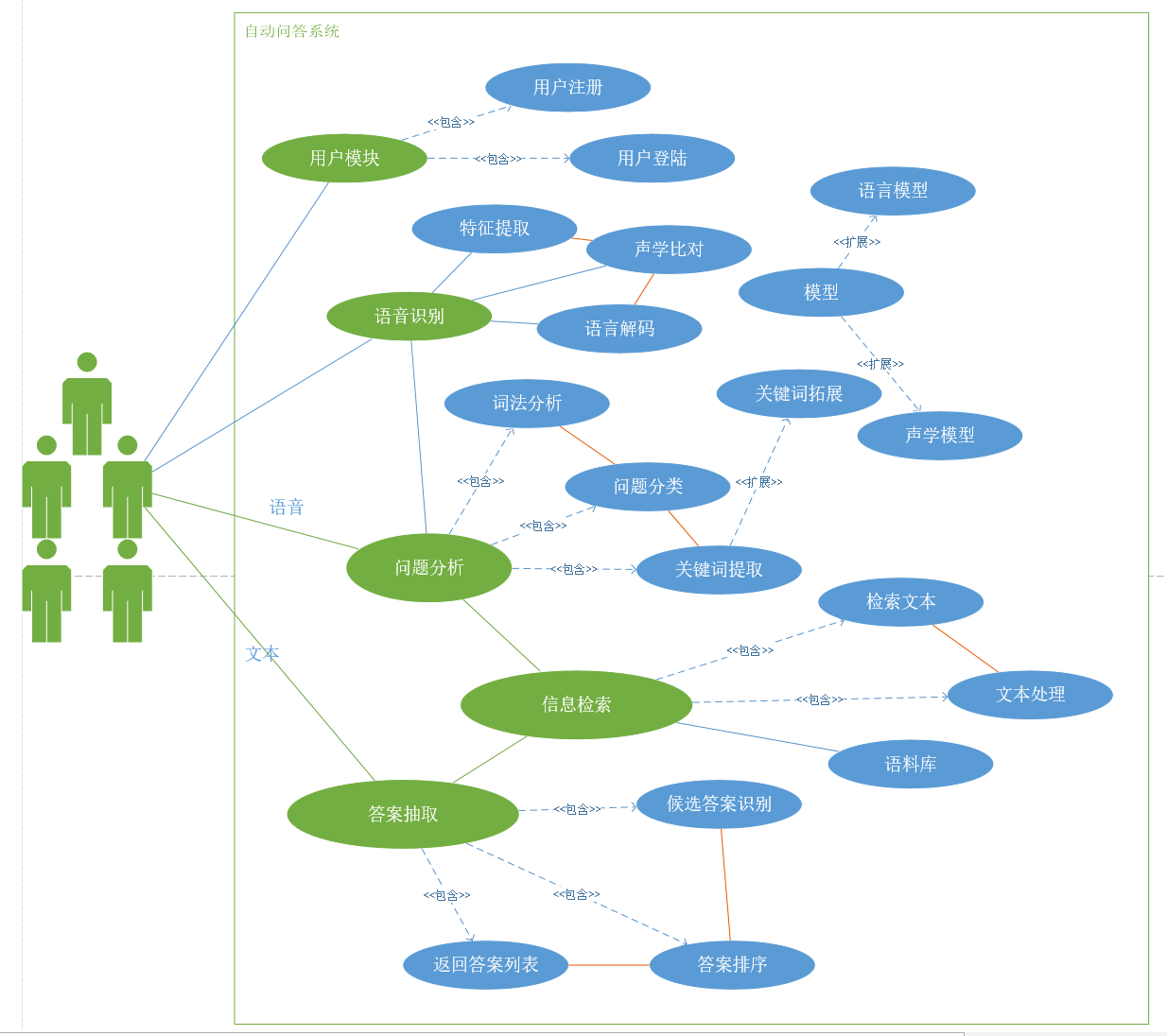
信息检索模块的功能是根据问题分析阶段的关键词提取及扩展生成的查询式在数据库、知识库或互联网[26]上检索所有可能与答案相关的段落、文章或者网页，并做一定的分析处理，将处理结果提交给答案抽取模块进行下一步操作。问答系统中的信息检索同传统的基于关键词匹配的信息检索类似，可以调用现有的比较成熟的各种检索模型。

### 3.2.3答案抽取

答案抽取是问答系统的最终模块，也是最关键的模块。这一模块所应用的答案抽取算法的优劣直接关系着整个问答系统的性能。答案抽取的对象是信息检索模块提交的相关文档，候选答案可以是一句话、一段摘要也可以是一个词，问题类型不同，所得到的答案形式也是不同的，所以答案抽取时还需依据问题类型采取相应的策略。由于信息检索阶段返回的相关文档通常较多，问答系统会抽取出大量的候选答案[27]。因此，候选答案的评估和排序是必不可少的，并最终返回一个最佳答案或按置信度高低排序的答案列表。

## 新需求

1. 之前的部分都对系统如何获取信息，并没有做详细的声明，只是说明了自动问答系统需要一段文本的输入，因此在此，考虑到各种实际情况还有现成的技术，决定将输入自动问答系统的信息媒介允许为语音输入。因此在自动问答系统中需要添加语音转为文字的功能，再进一步接受系统中别的模板的处理。具体的改变的用例模板如图所示。
2. 同时，之前并没有进行用户注册登录等用户信息的模块。此部分功能有成熟的模板可以进行套用，用户方面的信息已经成为了现代的每一个应用所必须会有的，因此我们的系统也具有这样的模块。



# 体系结构设计

## 4.1总体结构

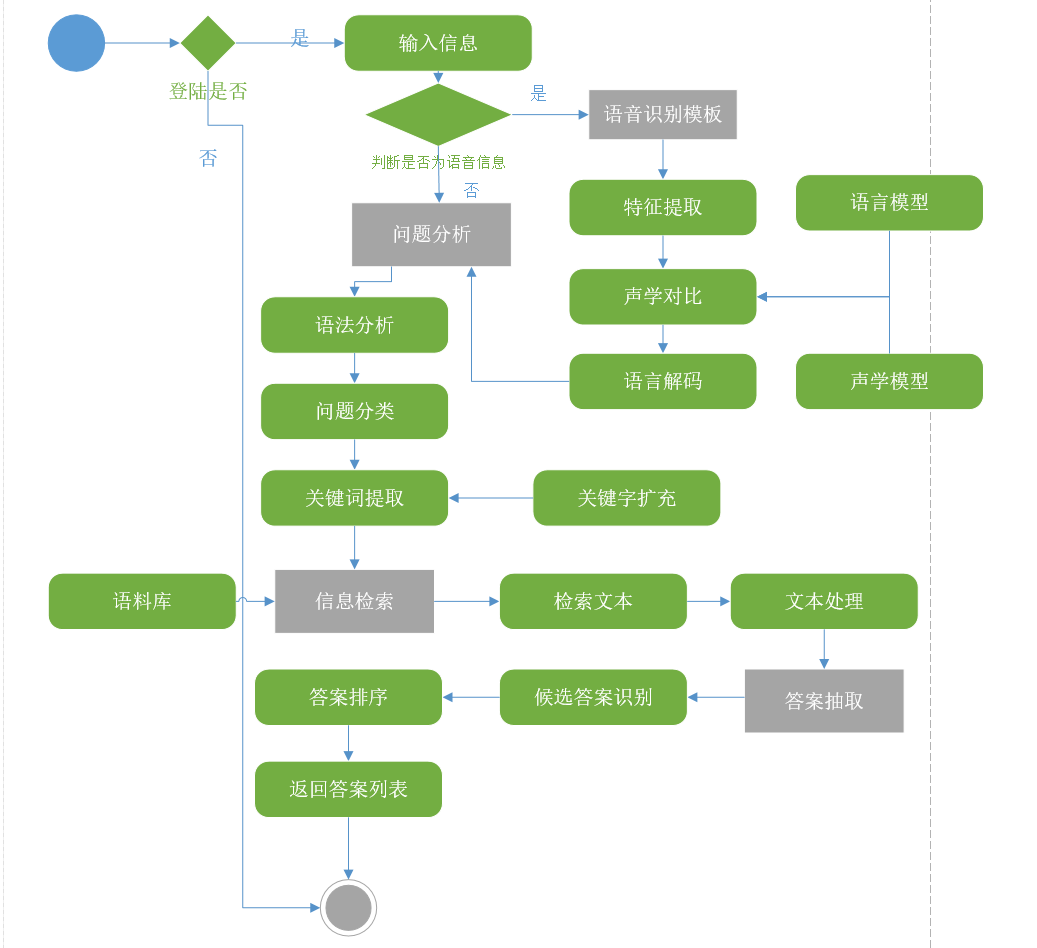
用一系列图表给出系统的总体结构，并对图中的相关内容进行说明。可能包括软件体系结构、硬件体系结构、技术体系结构、支撑体系（部署和实施方案）结构等各个方面，可根据实际情况每个方面分为一小节来写。

采用面向对象分析与设计方法，建议从UML的4+1视图描述软件结构：完善用例图、活动图，给出类图、时序图、状态图。

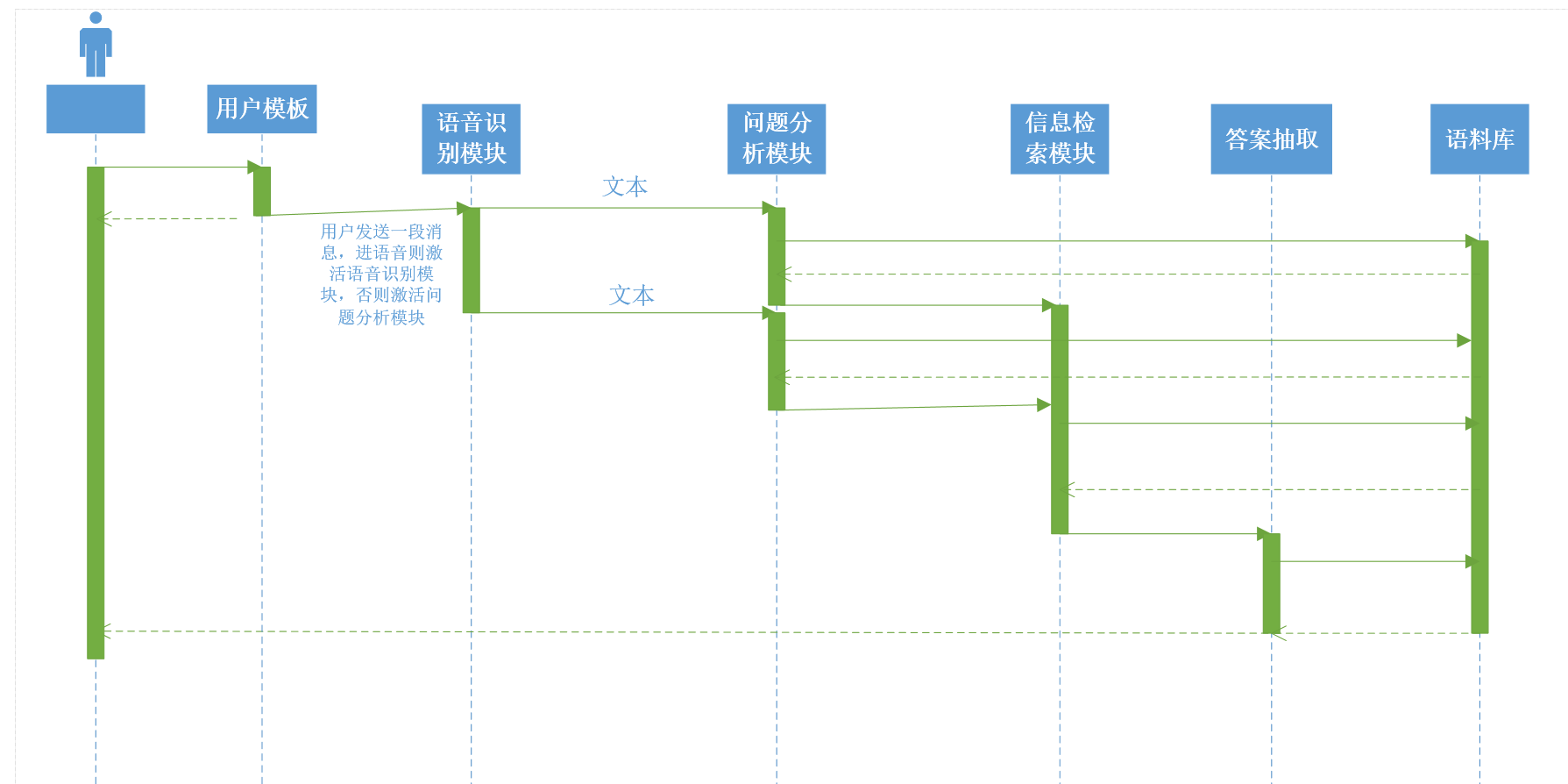
### 4.1.0用例图

如上所示3.3节所示。

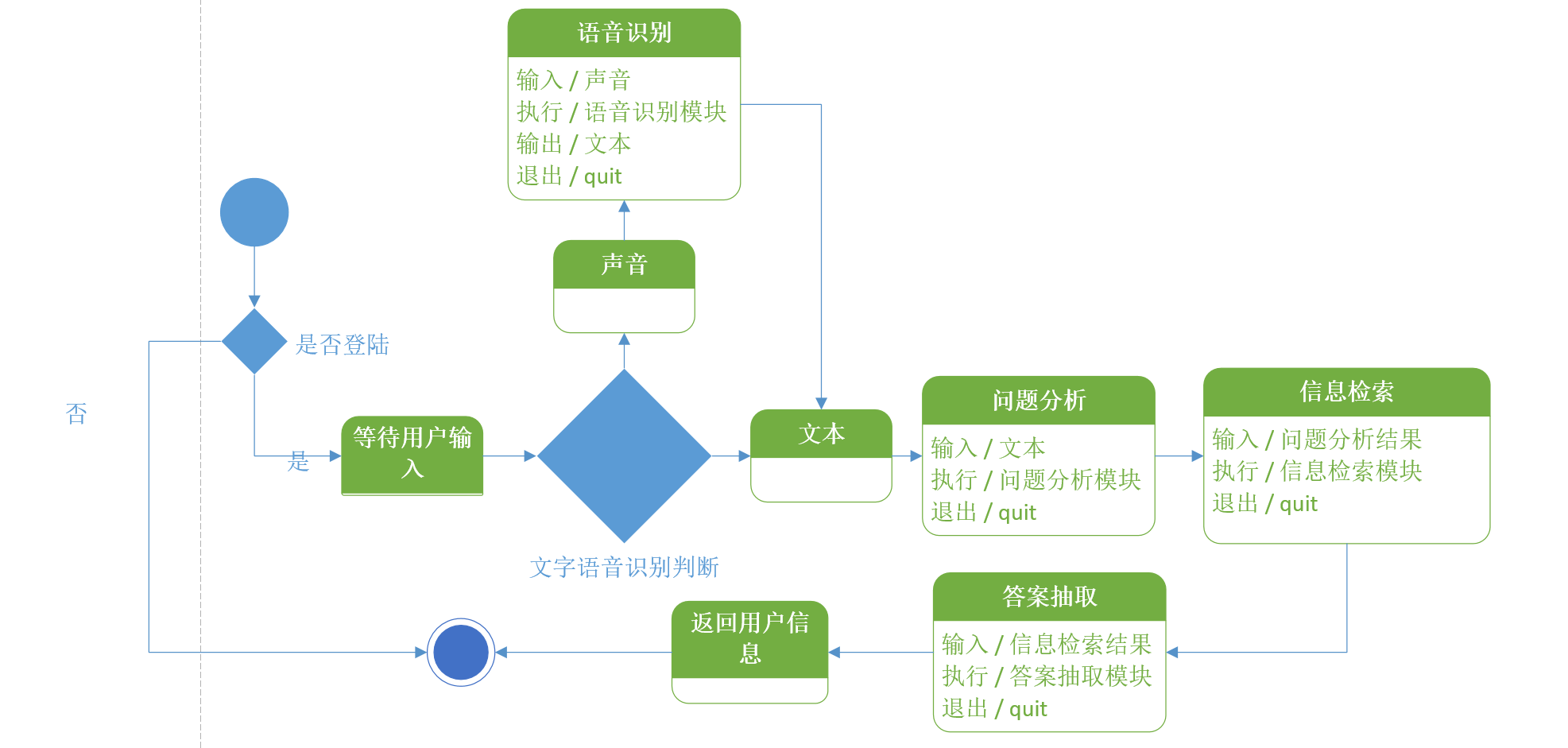
### 4.1.1活动图



### 4.1.2时序图

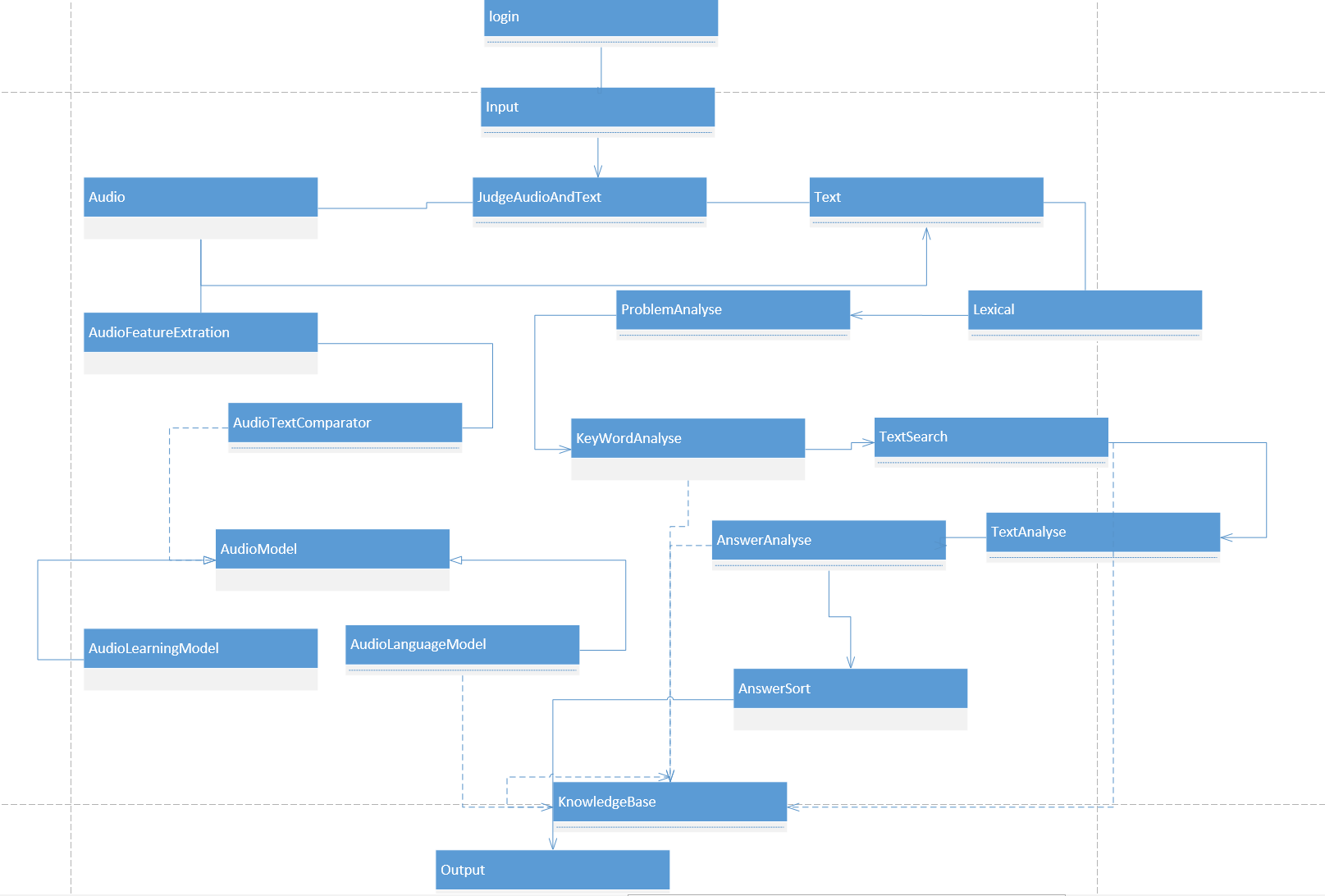


### 4.1.3状态图



### 4.1.4类图

以下类图是从系统层面的类图，是一个具有象征意义的大类。



## 4.2功能分配

说明需求文档当中各项功能要求同总体结构的对应关系。

## 4.3关键问题及解决方案

给出系统总体结构中所涉及的关键问题的设计决策和解决思路。

### 4.3.1问题分析模块及解决方案

由于在用户输入问题中，并非所有的词都能同等的表达相同重要性的含义，因此需要对问题进行预处理，并为后续的问题分类和关键词提取等提供依据。预处理的功能由自然语言处理模块完成，该模块一般包括：中文词法分析模块、命名实体识别模块和句法分析模块。

1. 实现中文词法分析、命名实体识别和句法分析等自然语言处理模块

对用户问句进行分词、词性标注、命名实体识别和句法分析等自然语言处理，可以为问题分类提供依据，同时也使从问句中提取关键词、扩展关键词成为可能。

1. 实现问题分类模块

将用户输入的问题按不同的分类标准归入不同的类别，可以为问题提取关键词提供依据，同时也确定了答案抽取的基本对象。

1. 实现问题关键词提取

根据问题的类别，和自然语言处理的结果，将问题中的关键词提取出来，经过扩展之后构成信息检索的检索词，同时也成为答案排序时的依据。

1. 实现问题关键词的扩展

之所以需要扩展关键词，是因为在正确的答案句子中，很多关键词并不是问题中的关键词，而是问题关键词的同义词、近义词甚至是相关词扩展。

1. 实现查询式的生成

结合一定的启发式规则生成合适的查询，以提高检索模块的召回率。

### 4.3.2问题分类模块及其解决方案

#### 4.3.2.1问题

问题分类是问题分析模块的核心，其目的是根据预期的答案类型把问题划分到相应的语义类别，对后续的答案抽取和选择有重要的导向作用。首先，它有效减少了候选答案的空间，其次，决定了答案抽取的策略。该模块的好坏直接影响着问答系统的性能。但是，在一般的中文问答系统中，问题分类是很复杂也是难度很大的领域之一。这主要归因于它与文本分类的两点不同：一、一个问题往往只有一句话，可获取的上下文信息量很少，从而能用于问题分类的特征很有限；二、许多问题并不能严格划分到明确的领域，限制了问题分类的准确性。

#### 4.3.2.2解决方案：使用基于支持向量机的中文问题分类

本系统选取了4种分类特征：

1）问题疑问词（QW）；

2）疑问词的附属成份（QS）；

3）疑问意向词（IW）；

4）疑问意向词在知网中的义原（IS）。

由于一个问题样本所包含的字数较少，不同于包含上千字的文本，因此选用文本分类中的词频权重如tf-idf等不太合适，相反，用最简单的权重——布尔权重对问题分类的效果更好。即对于要测试的问题，如果它所提取出的特征词在特征空间中出现，则权值为1，否则为0，即所谓的布尔权值。

问题疑问词的选取（QW）

中文疑问句中的疑问词包含着非常重要和明确的问题分类信息，是所有分类特征中最重要的特征，因此正确抽取疑问词非常关键[53]。首先，我们将经常使用的一些疑问词， 例如“什么”、“哪”、“谁”、“多少”等建立一个疑问词表。在问句进行词法分析之后，选取其中标记为“/ry”的词到疑问词表中进行查找，从而确定问题Q的疑问词。

特别地，“多”、“几”这样的疑问词并不能被ICTCLAS分词软件标识为“/ry”，此时若不包含其他标注为“/ry”的词，则将“多”、“几”视为疑问词。另外，为了减少数据稀疏，我们对一些同义疑问词进行了转化处理，例如：“啥时候”、“何地”实际上和“什么时候”、“什么地方”完全等价，经过转化，统一选取“什么”作为疑问词。实验结果证明，这样的转化处理收到了较好的效果，对于句法结构相关特征的正确抽取也有一定帮助。

疑问词附属成份的选取（QS）

在对问题进行句法分析得到依存弧及关系类型并确定问题疑问词之后，即可以抽取疑问词的附属成份（我们把疑问词所依存的词和依存于疑问词的词统称为疑问词的附属成分）。这样的特征提取减少了很多分类噪音。

经大量的统计发现，疑问词附近的量词、数量词、形容词、动词含有重要的信息。因此我们选取与疑问词有弧相连的词中词性标注为“/q”（量词）、“/mq”（数量词）、“/a”（形容词）或“/v”（动词）的词作为疑问词的附属成份。

疑问意向词的选取（IW）

疑问意向词[54]表达“问题问的到底是什么”这样一个含义的概念。关于疑问意向词，目前还没有明确的定义。一般认为，用户的疑问意图就是要得到一个未知信息，也就是问题中最能体现答案类型的词。比如问题“加拿大的官方动物是什么”中的“动物”就是该句的疑问意向词。

根据汉语句子的表达习惯，在问题疑问词附近的词更能表达整个句子所要表达的语义信息，对于问题分类常常具有更加重要的作用，特别是其中具有名词特性的词，也就是标记为“n”的词。同时，经过统计研究发现，疑问词右边标注为“n”的词比左边的“n”更加丰富和有效，但是在选择“n”标记的词的数量上并非越多越好因此，我们采用了如下疑问意向词的选取方法：

1.选取疑问词右边标记为“n”的词作为疑问意向词，并且最多选取两个。如果疑问词的右边没有标记为“n”的词，则转到第2步；

2.在疑问词左边选取标记为“n”的词作为疑问意向词，并且最多选取两个。

实验发现，如果存在有多个标记为“n”的词，并非选择的越多越好，选取的过多反而会增加很多干扰信息，产生噪声。因此，本文对于某一问题最多选取两个疑问意向词作为分类特征。

疑问意向词在知网中首义原的选取（IS）

由于汉语的灵活性，一种语义的问题通常可以有多种表达方式，例如可以选择不同的词或组织成不同的句法结构。但从问题分类的角度来看，相同或相近语义的问题应该被划分为同一类别。进一步分析发现，不同问句之所以属于同一类别，常常是由于不同问句中的疑问意向词具有相同或类似的语义，在知网中它们对应概念的DEF定义中具有相同的首义原。

本文首先对每个小类选取正确的分类义原；然后，对有多个DEF定义的词，直接选取和分类义原一致的DEF首义原作为分类特征。一般情况下，一个词的DEF定义只属于唯一的小类分类。对于其他情况，我们目前只是简单地根据知网中DEF的排列顺序进行选取。虽然以上方法不够精确，但实验结果表明是可行且有效的。

### 4.3.3答案抽取模块及其解决方案

#### 4.3.3.1问题

答案抽取模块是问答系统的最终模块，也是最核心的模块，它的主要工作是根据问题的类型从信息检索模块所得到的结果中找出相关的词或短语作为问题的答案返回给用户。答案抽取模块性能的优劣直接影响着整个问答系统的性能。

#### 4.3.3.2解决方案

在本系统中，答案抽取分以下几个步骤：候选句过滤；候选答案识别；候选答案排序。

**1)候选句过滤**

候选句过滤的目的是从信息检索模块返回的相关摘要中舍弃掉与用户问题不相关的内容。

候选句过滤主要包含两部分工作，一是对相关摘要集进行断句，即将信息检索模块提供的摘要集变为候选句集。断句工作由句子分割器完成，分割的根据主要是摘要中所出现的标点符号，如句号、感叹号、问号、省略号。

另外根据Google和百度搜索结果的研究分析，制定了一些启发式规则来辅助断句的实现。通过对摘要的分割处理后，我们获取了初始候选句集合，这些句子中可能有相当一部分和用户问题并不相关，所以需要进行进一步的候选句过滤。

**2)模式匹配**

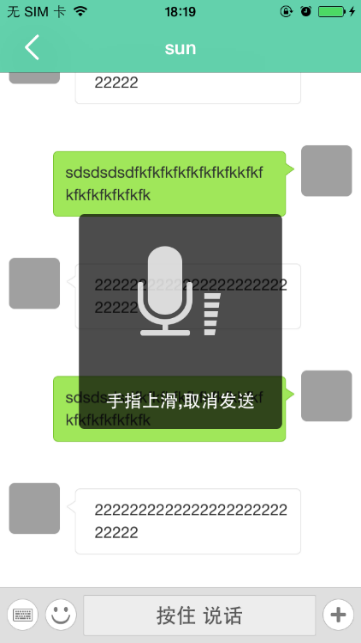
由于目前中文命名实体识别可识别的命名实体类型有限，只有一些人名、地名、机构名等等，不能满足本文问答系统所有问题类型要求的预期答案形式。因此有必要针对某些问题类型设计相应的答案模式，利用模式匹配的方法在候选句中抽取答案。下面列出了本系统的部分答案模式集，分别是具体时间类、星期类、百分比类.

# 5. 接口设计

## 5.1用户界面设计

给出系统用户界面的总体设计决策，和典型的用户界面风格。

说明：如果系统提供了对其他系统的接口，如从其他软件系统导入、导出数据，必须在此说明。

图为用户界面友好的聊天界面，支持用户输入模块(包含输入框、录音、表情选择、拍照、从相册选取照片功能)、录音模块。用户可以通过界面发送文字、语音、图片等，进行使用自动问答系统的提问功能和聊天功能。

## 5.2外部接口设计.

对系统的软件和硬件外部接口进行说明。

## 5.2.1用户登录

|  |  |  |  |
| --- | --- | --- | --- |
| 接口类型 | URL | | |
| 接口名称 | Login | | |
| 接口方向 | 服务器后台-->手机端 | | |
| 接口说明 | 用户登录 | | |
|  | **参数名称** | **参数类型** | **参数说明** |
| 输入值类型 | Key | String | Key密钥保护数据 |
| Type | String | 调用接口的名称 |
| Uid | String | 用户名 |
| Pwd | String | 密码 |
| 返回值类型 | 返回结果具体结构如下 | | |
| status | Int | 200成功300失败 |
| 说明 | 返回JSON格式数据 | | |

## 5.2.2用户注册

|  |  |  |  |
| --- | --- | --- | --- |
| 接口类型 | URL | | |
| 接口名称 | Register | | |
| 接口方向 | 服务器后台-->手机端 | | |
| 接口说明 | 添加用户 | | |
|  | **参数名称** | **参数类型** | **参数说明** |
| 输入值类型 | Key | String | Key密钥保护数据 |
| Type | String | 调用接口的名称 |
| Uid | String | 用户名 |
| Pwd | String | 密码 |
| Phone | String | 手机 |
| QQ | String | QQ |
| Mail | String | 邮箱 |
| 返回值类型 | 返回结果具体结构如下 | | |
| status | Int | 200成功300失败 |
| 说明 | 返回JSON格式数据 | | |

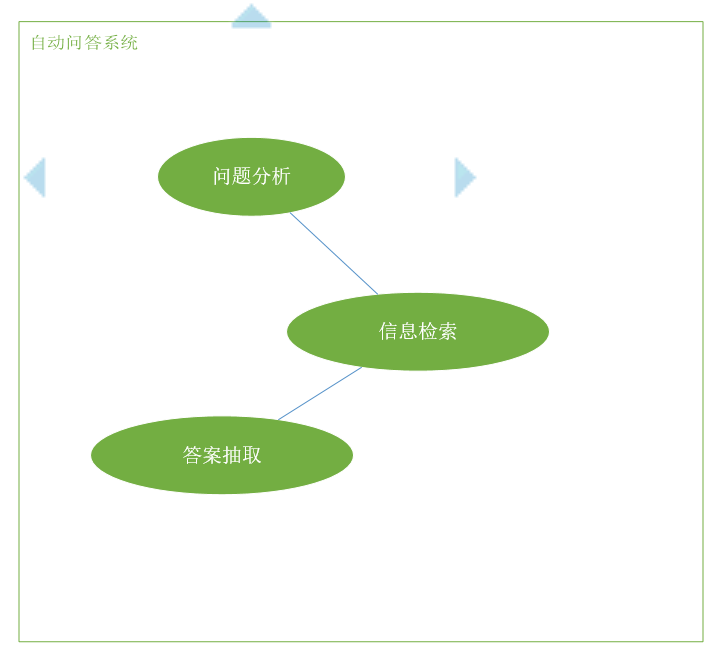
## 5.2.3获得文本问题回答

|  |  |  |  |
| --- | --- | --- | --- |
| 接口类型 | URL | | |
| 接口名称 | GetResByText | | |
| 接口方向 | 服务器后台-->手机端 | | |
| 接口说明 | 给出一段文本，得到这段文本的回答 | | |
|  | **参数名称** | **参数类型** | **参数说明** |
| 输入值类型 | Key | String | Key密钥保护数据 |
| Type | String | 调用接口的名称 |
| Text | String | 用户的提问 |
| 返回值类型 | 返回结果具体结构如下 | | |
| status | Int | 200成功300失败 |
| Result | String | 问题的回答 |
| 说明 | 返回JSON格式数据 | | |

## 5.2.4提交信息

|  |  |  |  |
| --- | --- | --- | --- |
| 接口类型 | URL | | |
| 接口名称 | GetResBySound | | |
| 接口方向 | 服务器后台-->手机端 | | |
| 接口说明 | 提交订单 | | |
|  | **参数名称** | **参数类型** | **参数说明** |
| 输入值类型 | Key | String | Key密钥保护数据 |
| Type | String | 调用接口的名称 |
| Sound | File | 提交的音频信息 |
| 返回值类型 | 返回结果具体结构如下 | | |
| status | Int | 200成功300失败 |
| Result | String | 问题的回答 |
| 说明 | 返回JSON格式数据 | | |

## 5.3内部接口设计

对系统各构件（模块）之间的接口进行说明，如下图所示。

|  |  |  |  |
| --- | --- | --- | --- |
| 接口类型 | Set<String> | | |
| 接口名称 | LexicalAnalyse | | |
| 接口方向 | TextInput->LexicalAnalyse | | |
| 接口说明 | 实现中文词法分析、命名实体识别和句法分析,问题分类模块,问题关键词提取,问题关键词的扩展,查询式的生成 | | |
|  | **参数名称** | **参数类型** |
| 输入值类型 | Text | String |
| isAudio | Boolean |
| 返回值类型 | 返回结果具体结构如下 | | |
| result | Set<String> |
| 说明 | 返回下一个模块需要处理的数据集 | | |

问题分析接口：

|  |  |  |  |
| --- | --- | --- | --- |
| 接口类型 | Set<String> | | |
| 接口名称 | InfoSearch | | |
| 接口方向 | LexicalAnalyse -> InfoSearch | | |
| 接口说明 | 实现爬虫实现,相关性评估, 网页覆盖率共三个部分 | | |
|  | **参数名称** | **参数类型** | **参数说明** |
| 输入值类型 | words | Set<String> | 问题分析结果 |
| 返回值类型 | 返回结果具体结构如下 | | |
| result | Set<String> | |
| 说明 | 返回下一个模块需要处理的数据集 | | |

信息检索接口：

答案抽取接口：

|  |  |  |  |
| --- | --- | --- | --- |
| 接口类型 | Set<String> | | |
| 接口名称 | AnswerAnalyse | | |
| 接口方向 | 信息检索->答案抽取 | | |
| 接口说明 | 实现候选句过滤；候选答案识别；候选答案排序 | | |
| 输入值类型 | **参数名称** | **参数类型** | **参数说明** |
| words | Set<String> | 检索分析得到的字符集 |
| 返回值类型 | 返回结果具体结构如下 | | |
| result | String | |
| 说明 | 返回用户自动问答系统所计算得到的答案 | | |

# 6.数据结构设计

## 6.1公共数据结构设计

对程序当中所定义的全局变量、常量、全局数据结构或类当中Public数据结构进行定义和说明。

|  |  |
| --- | --- |
| 数据成员 | 说明 |
| Analyzer \_analyzer | 分析器指针 |
| Merger \_merger | 归并器指针 |
| Fields \_fields | 系统中涉及到的文本信息 |
| vector<Document> \_docs | 问答对资源 |
| vector< Posting> \_postings | 答对资源分词后的结果 |
| string \_directory | 问答对路径 |
| int64\_t \_bufferStart | 缓冲区的起始位置 |
| size\_t \_bufferPosition | 缓冲区的当前位置 |
| float \_fWeight | 领域权重信息 |
| DocFrqInfo\_docFrqInfo | 关键词的大规模统计权重信息 |
| set<string>\_stopWords | 停用词词表 |
| set<string>\_quesWords | 疑问词词表 |
| NormBytes \_norms | 问句的长度信息 |
| int32\_t \_docs | 文档号数组 |
| TermDocs \_pTermDocs | 文档信息 |

## 6.2数据库设计

对程序当中所采用的数据库相关的名称和标识符、在数据库当中的位置、定义、度量单位、格式和值域、敏感程度、数据项名、缩写词和代码，包括用于数据库的规格说明等进行定义。建议给出数据库表结构。

说明：数据库表结构可以参考如下，字段值域不够可以自己增加。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据库的表结构 | 表名称 | 设计者 | 审核者 | 完成日期 |
|  |  |  |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 字段代号 |  | 类型 | 值域 | 数据项名 | 索引或键 | 备注 | 缩写词 |
|  |  |  |  |  |  |  |  |

### 6.2.1文档数据库

主要使用文档进行初步存储。其中词典有两个形态：文本文件(filename.txt)和缓存文件(filename.txt.bin或filename.txt.trie.dat和filename.txt.trie.value)。

#### 6.2.1.1文本文件

采用明文储存，UTF-8编码，CRLF换行符。

#### 6.2.1.2缓存文件

就是一些二进制文件，通常在文本文件的文件名后面加上.bin表示。有时候是.trie.dat和.trie.value。后者是历史遗留产物，分别代表trie树的数组和值。修改了任何词典，只有删除缓存才能生效。词典说明：

词典分为词频词性词典和词频词典。

词频词性词典

每一行代表一个单词，格式遵从[单词] [词性A] [A的频次] [词性B] [B的频次] 。

词频词典

每一行代表一个单词，格式遵从[单词] [单词的频次]。

### 6.2.2通用数据库表的设计

例举如下，详见设计文档文件夹下wikipedia-database-schema文件。

界面表：

|  |  |
| --- | --- |
| 数据库名：Page | |
| 名称 | 类型 |
| page\_id | int |
| page\_namespace | int |
| page\_title | varchar(255) |
| page\_counter | bigint |
| page\_is\_redect | tinyint |
| page\_latest | int |
| page\_len | int |

重定向表：

|  |  |
| --- | --- |
| 数据库名：redirect | |
| 名称 | 类型 |
| rd\_from | int |
| rd\_namespace | int |
| rd\_title | varchar(255) |
| rd\_interwiki | varchar(255) |

文档存储表：

|  |  |
| --- | --- |
| 数据库名：archive | |
| 名称 | 类型 |
| ar\_id | int |
| ar\_title | varchar(255) |
| ar\_text | mediumblob |
| ar\_comment | tinyblob |
| ar\_user | int |
| ar\_text\_id | int |
| ar\_len | int |

类别表：

|  |  |
| --- | --- |
| 数据库名：category | |
| 名称 | 类型 |
| cat\_id | int |
| cat\_title | varchar(255) |
| cat\_pages | int |
| cat\_files | int |

搜索表：

|  |  |
| --- | --- |
| 数据库名：searindex | |
| 名称 | 类型 |
| si\_page | int |
| si\_title | varchar(255) |
| si\_text | mediumtext |

# 7.详细设计

每一小节给出一个模块（构件）的详细设计方案。包括模块概述、模块的接口说明（即输入、输出）、以及内部结构设计。其中内部结构又可以考虑从静态、动态结构两个方面阐述；静态结构应给出该模块（构件）的类结构（类图），动态结构应给出该模块关键业务流程的交互模型（顺序图），还可根据实际情况给出状态图（某个构件或对象的状态迁移）和活动图（某个算法的实现流程）等内容。

## 7.1用户模块

### 7.1.1概述

### 7.1.2接口说明

### 7.1.3内部设计（静态）

该模块（构件）的类结构（类图）

### 7.1.4内部设计（动态）

关键业务流程的交互模型（顺序图），还可根据实际情况给出状态图（某个构件或对象的状态迁移）和活动图（某个算法的实现流程）等内容。

## 7.2语音模块

### 7.3.1概述

### 7.3.2接口说明

### 7.3.3内部设计（静态）

该模块（构件）的类结构（类图）

### 7.2.4内部设计（动态）

关键业务流程的交互模型（顺序图），还可根据实际情况给出状态图（某个构件或对象的状态迁移）和活动图（某个算法的实现流程）等内容。

## 7.3问题分析

### 7.3.1概述

问题分析是问答系统的第一步，负责对用户输入的问题进行充分地理解。对于中文问答系统来说，这个任务是相对复杂而又困难的。问题分析模块主要实现以下几个功能：

实现中文词法分析、命名实体识别和句法分析等自然语言处理模块：对用户问句进行分词、词性标注、命名实体识别和句法分析等自然语言处理，可以为问题分类提供依据，同时也使从问句中提取关键词、扩展关键词成为可能。

实现问题分类模块：将用户输入的问题按不同的分类标准归入不同的类别，可以为问题提取关键词提供依据，同时也确定了答案抽取的基本对象。

实现问题关键词提取：根据问题的类别，和自然语言处理的结果，将问题中的关键词提取出来，经过扩展之后构成信息检索的检索词，同时也成为答案排序时的依据。

实现问题关键词的扩展：之所以需要扩展关键词，是因为在正确的答案句子中，很多关键词并不是问题中的关键词，而是问题关键词的同义词、近义词甚至是相关词扩展。

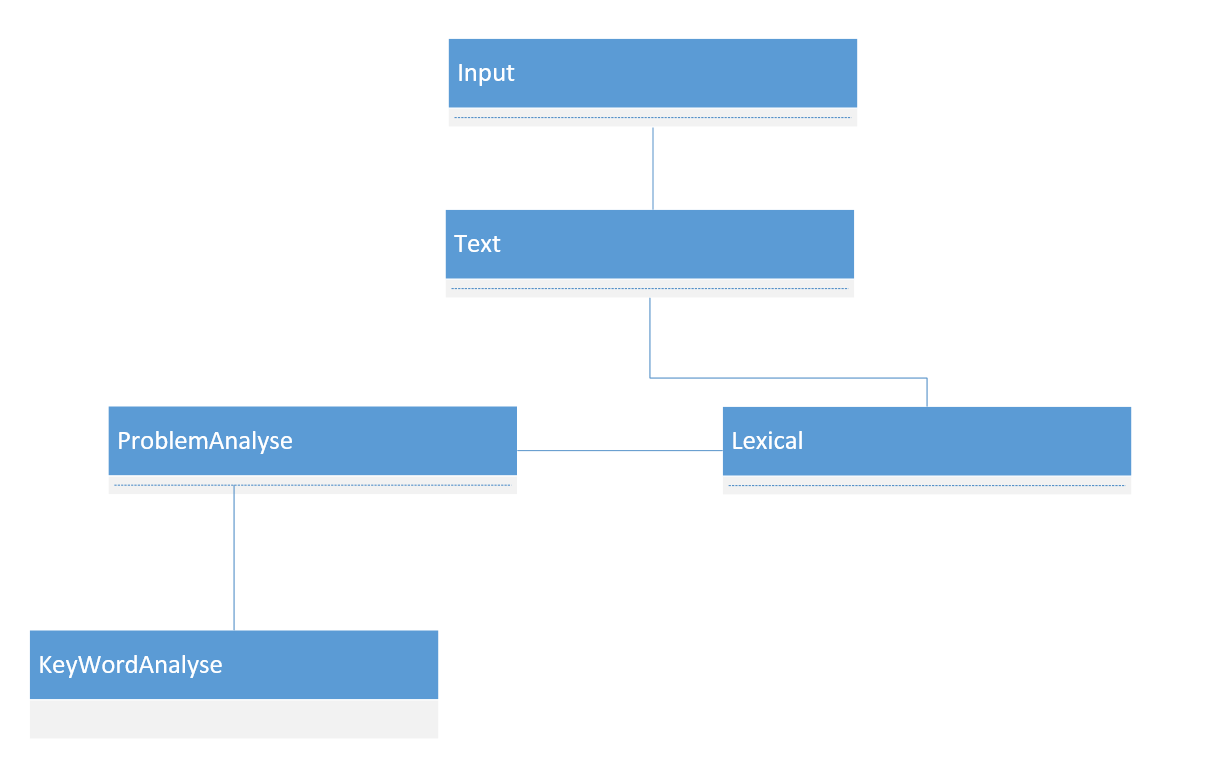
实现查询式的生成：结合一定的启发式规则生成合适的查询，以提高检索模块的召回率。

### 7.3.2接口说明

|  |  |  |  |
| --- | --- | --- | --- |
| 接口类型 | Set<String> | | |
| 接口名称 | LexicalAnalyse | | |
| 接口方向 | TextInput->LexicalAnalyse | | |
| 接口说明 | 实现中文词法分析、命名实体识别和句法分析,问题分类模块,问题关键词提取,问题关键词的扩展,查询式的生成 | | |
|  | **参数名称** | **参数类型** | **参数说明** |
| 输入值类型 | Text | String | 输入的文本信息 |
| isAudio | Boolean | 是否为语音信息的记录 |
| 返回值类型 | 返回结果具体结构如下 | | |
| result | Set<String> | 返回 |
| 说明 | 返回下一个模块需要处理的数据集 | | |

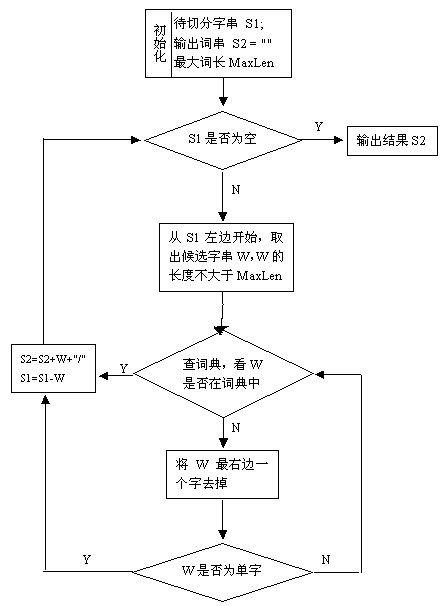
### 7.3.3内部设计（静态）

该模块（构件）的类结构（类图）



### 7.3.4内部设计（动态）

关键业务流程的交互模型（顺序图），还可根据实际情况给出状态图（某个构件或对象的状态迁移）和活动图（某个算法的实现流程）等内容。

下面是例举的分析问题模块中重要的算法流程图：

## 7.4 信息检索

### 7.4.1概述

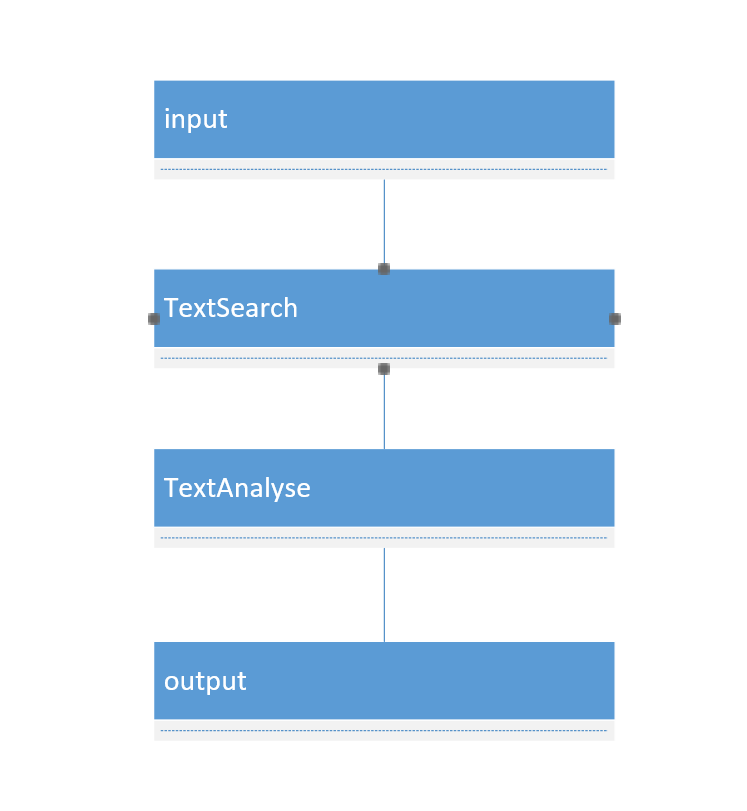
对于开放域问答系统，比较常用的方法是将整个网络看作一个半结构化文档的知识库，利用搜索引擎处理和过滤信息。在本系统的实现中，信息检索模块也是直接调用Internet上已有的搜索引擎：Google和百度，并将返回结果中的网页片断(Snippets)作为答案抽取的资源。

网络搜素引擎的运行机制是，首先由网页获取器Spider根据地址管理器中的地址在互联网上进行搜索网页，然后将获得的不同格式的网页进行格式统一化处理，接着对处理后的网页进行索引化管理，并将其放入索引数据库中。当用户进行检索信息时，检索器根据用户的检索请求对网页进行相关度的计算，并对结果进行排序，最后将排序结果返回给用户。

### 7.4.2接口说明

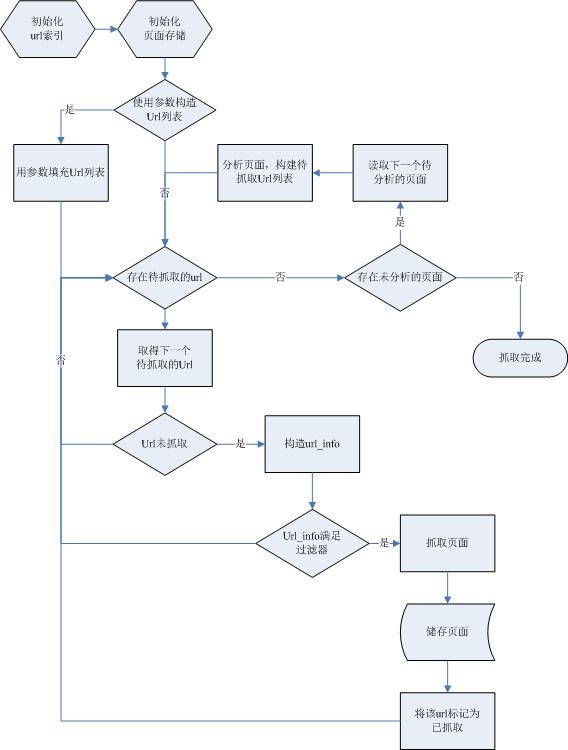
|  |  |  |  |
| --- | --- | --- | --- |
| 接口类型 | Set<String> | | |
| 接口名称 | InfoSearch | | |
| 接口方向 | LexicalAnalyse -> InfoSearch | | |
| 接口说明 | 实现爬虫实现,相关性评估, 网页覆盖率共三个部分 | | |
|  | **参数名称** | **参数类型** | **参数说明** |
| 输入值类型 | words | Set<String> | 问题分析结果 |
| 返回值类型 | 返回结果具体结构如下 | | |
| result | Set<String> | |
| 说明 | 返回下一个模块需要处理的数据集 | | |

### 7.4.3内部设计（静态）

该模块（构件）的类结构（类图）

### 7.4.4内部设计（动态）

关键业务流程的交互模型（顺序图），还可根据实际情况给出状态图（某个构件或对象的状态迁移）和活动图（某个算法的实现流程）等内容。



## 7.5答案抽取

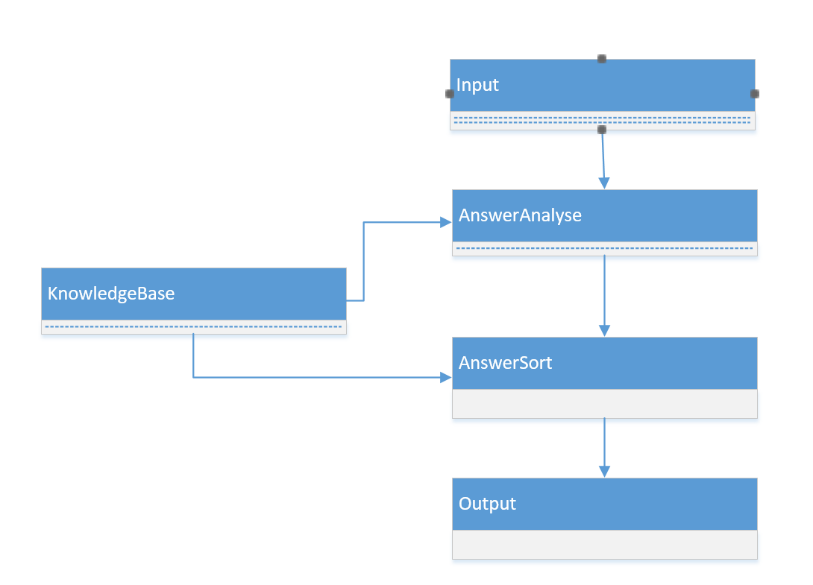
### 7.5.1概述

答案抽取模块是问答系统的最终模块，也是最核心的模块，它的主要工作是根据问题的类型从信息检索模块所得到的结果中找出相关的词或短语作为问题的答案返回给用户。答案抽取模块性能的优劣直接影响着整个问答系统的性能。在本系统中，答案抽取分以下几个步骤：候选句过滤；候选答案识别；候选答案排序。

### 7.5.2接口说明

|  |  |  |  |
| --- | --- | --- | --- |
| 接口类型 | Set<String> | | |
| 接口名称 | AnswerAnalyse | | |
| 接口方向 | 信息检索->答案抽取 | | |
| 接口说明 | 实现候选句过滤；候选答案识别；候选答案排序 | | |
| 输入值类型 | **参数名称** | **参数类型** | **参数说明** |
| words | Set<String> | 检索分析得到的字符集 |
| 返回值类型 | 返回结果具体结构如下 | | |
| result | String | |
| 说明 | 返回用户自动问答系统所计算得到的答案 | | |

### 7.5.3内部设计（静态）

该模块（构件）的类结构（类图）

### 7.5.4内部设计（动态）

关键业务流程的交互模型（顺序图），还可根据实际情况给出状态图（某个构件或对象的状态迁移）和活动图（某个算法的实现流程）等内容。

**模式匹配**

由于目前中文命名实体识别可识别的命名实体类型有限，只有一些人名、地名、机构名等等，不能满足本文问答系统所有问题类型要求的预期答案形式。因此有必要针对某些问题类型设计相应的答案模式，利用模式匹配的方法在候选句中抽取答案。下面列出了本系统的部分答案模式集，分别是具体时间（Time）类、星期（Week）类、百分比（Percent）类：

Time：\*点｜\*时｜\*点\*分｜\*时\*分

Week：周\*｜星期\*｜礼拜\*

Percent：\*分之\* | 百分之\* | \*%

答案模式集主要是根据问题类型的常见答案句结构而建立的[ ]，其中“\*”是通配符，但是对于不同类型的问题，通配符常常有一定的词性要求。如果候选句能与相应问题类型的某个答案模式匹配，则匹配位置代表一个候选答案。

具体算法步骤如下：

|  |
| --- |
| 1. 根据问题类型取出所有相应的可选答案模式，形成队列，并在所有可选答案模式及当前候选句末尾各加一个结束标志； |
| 1. 取出第一个可选答案模式； |
| 1. 设置答案模式字符串指针AP和候选句字符串指针CP分别指向当前答案模式的第一个字符和候选句的第一个字符，初始化候选答案匹配位置MP=-1； |
| 1. 若AP指向结束标志，转7）；否则，从答案模式的AP位置起逐个向右读取字符，直至遇到第一个“\*”标志或结束标志（记录此位置为EP），将答案模式从AP到EP位置所构成的子串赋给一个临时字符串变量ss，并置AP=EP+1，若EP位置对应结束标志，则AP=AP-1； |
| 1. 若CP指向结束标志，转9）；否则从候选句的CP位置起逐个向右读取字符，并将临时字符串ss与候选句进行匹配，若匹配（当ss以“\*”结尾时，要求候选句中与“\*”对应位置内容不为空且词性与答案模式在此位置要求的词性相符），转6）；否则，若MP==-1转9），否则转8）； |
| 1. 令CP为候选句最后一个匹配字符的下一个位置；若MP==-1，则令MP=匹配的第一个字符位置，转4）； |
| 1. 保存候选答案的匹配位置及匹配的答案模式序号，重置MP =-1； |
| 1. 若CP未指向候选句结束标志，则重置AP指向当前答案模式的第一个字符，重置MP=-1，转4）； |
| 1. 取下一个可选候选答案模式，转3）；若无可选答案模式，输出所识别的该候选句中全部候选答案及匹配位置； |