

**<<基于知识图谱的自动问答系统>>**

**需求规格说明书**



北京航空航天大学

2015-11

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 | 2015.11.27 | 杨东东 |  | 更新第三四六部分 |
| 2 | 2015.11.28 | 方凯 |  | 更新第一二部分 |
| 3 | 2015.11.29 | 李睿霖 |  | 更新第五部分 |
| 4 | 2015.11.29 | 李睿霖、方凯 |  | 审查更改 |
| 5 | 2016.01.18 | 全体 |  | 整体修改 |
|  |  |  |  |  |

目 录

[1.范围 1](#_Toc440884215)

[1.1标识 1](#_Toc440884216)

[1.2系统概述 1](#_Toc440884217)

[1.3文档概述 1](#_Toc440884218)

[1.4术语和缩略词 2](#_Toc440884219)

[1.4.1数据库(Database) 2](#_Toc440884220)

[1.4.2自然语言处理(NLP) 2](#_Toc440884221)

[1.4.3矩阵论(Matrix) 2](#_Toc440884222)

[2.引用文档 2](#_Toc440884223)

[2.1书籍 2](#_Toc440884224)

[2.2批文 3](#_Toc440884225)

[2.3引用资料 3](#_Toc440884226)

[3.功能需求 3](#_Toc440884227)

[3.1系统的用例模型并进行简要的说明 4](#_Toc440884228)

[3.2对系统的用户进行系统的描述（即用例图中的参与者） 4](#_Toc440884229)

[3.3对系统的功能的系统描述 5](#_Toc440884230)

[4.数据需求 5](#_Toc440884231)

[4.1 wikipedia实体及其关系 5](#_Toc440884232)

[4.2 微博ER图 6](#_Toc440884233)

[5. 非功能需求 7](#_Toc440884234)

[5.1 性能 7](#_Toc440884235)

[5.2 可靠性 7](#_Toc440884236)

[5.3 易用性 7](#_Toc440884237)

[5.4 可扩展性 7](#_Toc440884238)

[5.5可移植性 8](#_Toc440884239)

[5.6系统安全性 8](#_Toc440884240)

[5.7系统可维护性 8](#_Toc440884241)

[6.运行需求 8](#_Toc440884242)

[6.1硬件接口 8](#_Toc440884243)

[6.2软件接口 8](#_Toc440884244)

[6.3用户界面需求 9](#_Toc440884245)

# 1.范围

## 1.1标识

文档标识号：A 2015-00-01-00

文档标题：<<基于知识图谱的自动问答系统>>--需求规格说明书

版本号：5.0

## 1.2系统概述

本条应简述本文档适用的系统和软件的用途，它应描述系统和软件的一般特性；概述系统开发、运行和维护的历史；标识项目的投资方、需方、用户、开发方和支持机构；标识当前和计划的运行现场。

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：问答需求者；

开发方：杨东东，李睿霖，方凯；

支持机构：提供知识库的单位；

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项目开发过程中的系统需求，做出需求分析。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

# 2.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

《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

# 3.功能需求

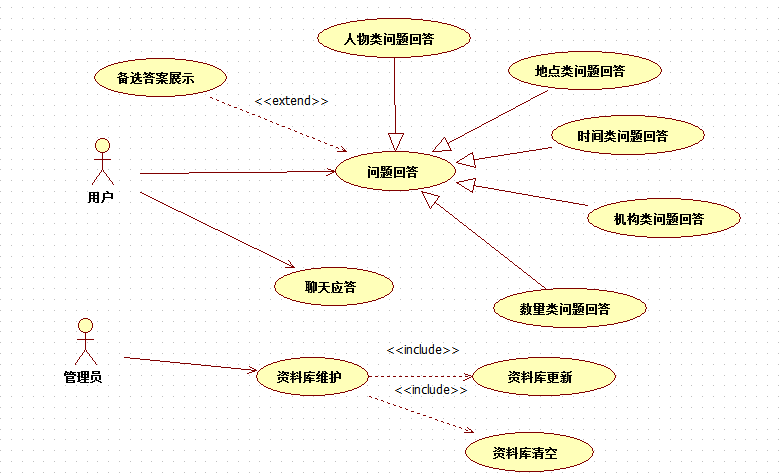
以用例图的形式给出系统功能需求的分解结构，并对用例模型中的参与者和用例进行详细的描述，可参考如下思路将本节划分为几小节（也可按照系统的实际情况进行调整）。

3.1节给出系统的用例模型，并进行简要的说明。

3.2节对系统的用户进行详细的描述（即用例图中的参与者）

3.3节以后每个小节描述一个用例模型，可采用文字的方式，对于涉及复杂流程的用例可以绘制其活动图。

## 3.1系统的用例模型并进行简要的说明



（UseCase.uml, 可用StarUML打开）

用户对该系统的主要需求为两点，问题回答及及聊天应答，而问题回答又可以按照提问种类分为询问人物类（PERSON）、地点类（LOCATION）、时间类(TIME)、机构类（ORGNIZATION）、数量类（QUANTITY）、聊天类等这几种问题。而管理员只有一个工作，即负责服务器资料库的维护。这个维护包括了资料和清空两部分。

问答系统一般由三个模块组成：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对系统的用户进行系统的描述（即用例图中的参与者）

用户可以通过自己的设计来设定喜好，主要为想用搜索引擎来搜索问题答案的用户均为本系统所收益的对象以及想与智能聊天机器人聊天的对象均为本系统所收益的对象。

从系统所需要实现的功能来看区分用户的类别，于是可以将用户区别为询问人物类（PERSON）、地点类（LOCATION）、时间类(TIME)、机构类（ORGNIZATION）、数量类（QUANTITY）、聊天类等这几种问题的群体。

## 3.3问题回答

用户输入一个问题，系统给出这个问题的答案。具体过程为：用户输入一串文本，系统对其进行词法分析、问题分类、关键词提取，随后在语料库中对关键词及问题进行检索，将检索结果进行包装成自然语言，最后根据相关度将得到的答案进行排序，在图形界面上展示给用户。

输入：一段文本。

反馈 ：返回这一段文本作为问题的结果。

## 3.4聊天应答

用户输入一段话语，系统给出他的应答。具体过程为：用户输入一串文本，系统对其进行词法分析、问题分类、关键词提取，随后在语料库中对关键词及问题进行检索，将检索结果进行包装成自然语言，最后根据前后文及相关度将得到的答案进行排序，在图形界面上展示给用户。该需求的实现方式与3.3相似。

输入：一段文本。

反馈 ：返回一条对话结果。

## 3.5资料库维护

管理员给出资料库更新的命令。系统通过爬虫系统对一系列信息网站进行爬取，从而得到资料库的更新。

输入：资料库更新指令。

反馈：更新成功信息（成功）/更新失败信息及失败原因（失败）

管理员给出资料库清空的命令。系统将资料库全部清空。

输入：资料库清空指令。

反馈：清空成功信息（成功）/清空失败信息及失败原因（失败）。

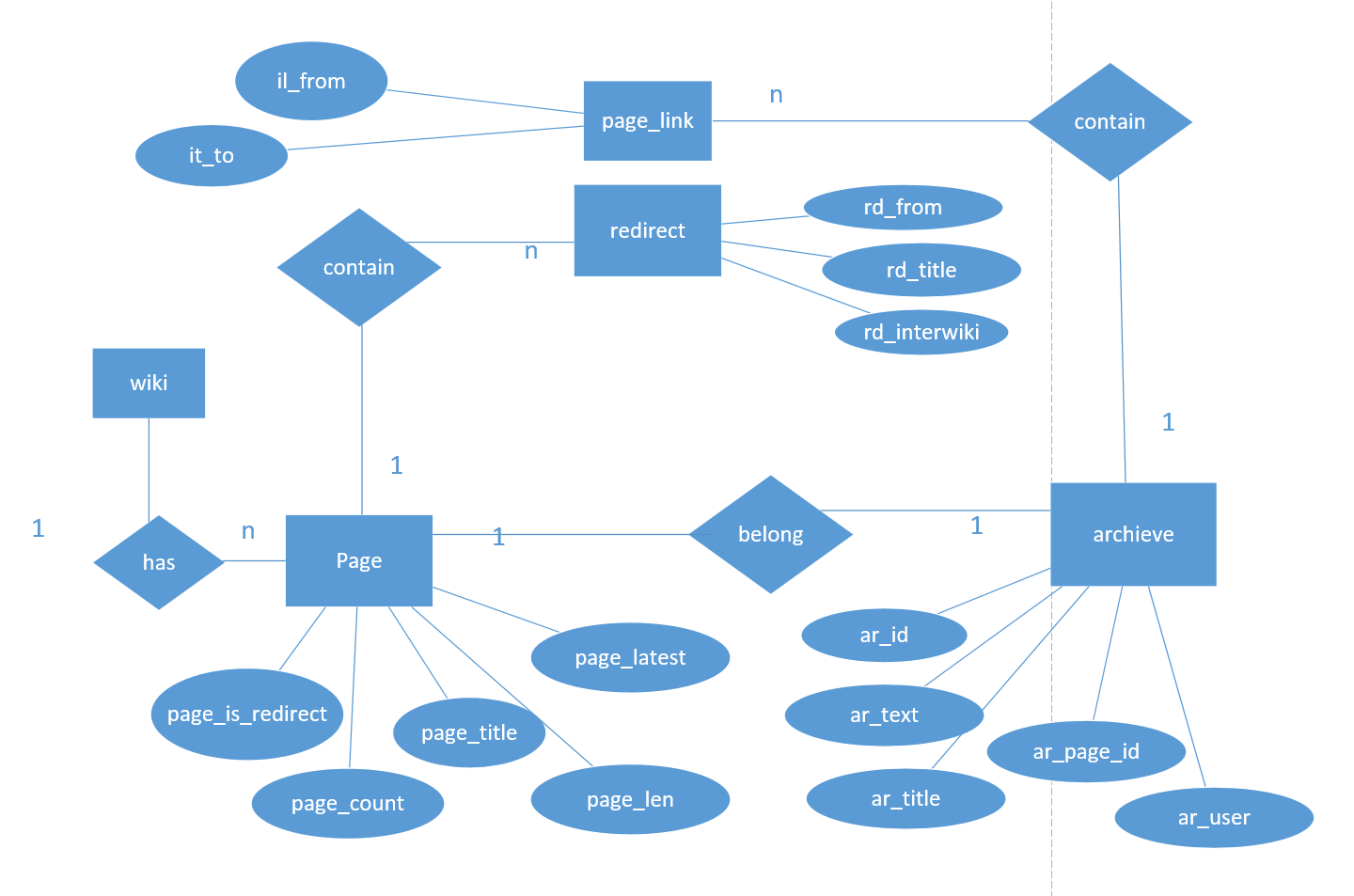
# 4.数据需求

描述该系统所涉及的数据实体。以ER图的方式给出基本的数据实体以及关系，再针对每个数据实体的数据项进行展开介绍。

本系统所利用的原始数据均来源于网上（期中包括百科和微博）。

## 4.1 wikipedia实体及其关系

以下为部分，详细见需求文档原图文件夹的wikipedia的类图。



## 4.2 微博ER图



# 非功能需求

## 5.1 性能

本系统的可靠性要求达到97%以上，能较快响应用户的操作。具有较高的容错性能，发生系统故障时不至于让系统崩溃，且能及时保存故障时的数据，保证后面的恢复工作顺利进行。

系统完成后，应有能力在一般人类对话中可以忍受的时间内给出一次提问的回答。现定为5s，即对于用户的每一次提问，可以在5s内给出回答。同时系统能保证至少50%的准确率。

## 5.2 可靠性

系统完成后，保证对于用户的每一次提问，都能得到较为贴切的分析，并能给出满足用户期望的回答。保证用户对结果的满意度在80%以上。

## 5.3 易用性

用户仅需要使用自然语言进行正常的提问即可，无需任何帮助或教程。要求系统提供良好的用户接口，易用的交互界面。常用功能应出现在界面易见位置，使用通俗易懂的文字描述，提供使用手册和在线帮助。系统要提供方便的手段供系统维护人员进行数据的备份，日常的安全管理，系统意外崩溃时数据的恢复等工作。

## 5.4 可扩展性

系统完成后将给出接口，使得其他应用可通过该系统获得对于一个自然语言写成的问题的答案。本系统应有良好的可移植性，不需要做大量代码修改就可完成软件的跨平台运行。

## 5.5可移植性

要求少用系统特有的特性，如尽量不要使用WINDOWS的API。分离前后端，后端包括数据库、系统内部机制；前端通过后端提供的统一接口访问后端数据；在实现跨平台软件开发时，只用修改前端，减少对后端做出的修改。

## 5.6系统安全性

建立完整的权限机制，以保证本系统的安全性。只有管理员才有权限直接修改后台数据，只有输入正确的口令，才能被授予管理员权限。

## 5.7系统可维护性

系统编写时要求有20%以上的注释，以便当系统出现问题时，软件维护人员容易对软件进行修改。应当有特定的数据库维护人员对数据库进行及时的备份，以保证数据库的安全性。

# 6.运行需求

## 6.1硬件接口

阿里云服务器、

Android手机

## 6.2软件接口

Mysql、

Redis、

Neo4J、

Apache Tomcat、

阿里云服务器、

eclipse集成开发环境

## 6.3用户界面需求

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、表情选择、拍照、从相册选取照片功能)。用户可以通过界面发送文字、图片等，进行使用自动问答系统的提问功能和聊天功能。

