北京航空航天大學
BEIHANG UNIVERSITY

<<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
V1	2015.11.27	杨东东	全体	更新第三四六部分
V2	2015.11.28	方凯	全体	更新第一二部分
V3	2015.11.29	李睿霖	全体	更新第五部分
V4	2015.11.29	李睿霖、方凯	全体	审查更改

目 录

1.范围 1

1.1 标识.....1

1.2 系统概述.....1

1.3 文档概述.....1

1.4 术语和缩略词.....2

2.引用文档.....3

3.功能需求.....4

3.1 系统的用例模型，并进行简要的说明.....4

3.2 对系统的用户进行系统的描述.....5

3.3 对系统的功能的系统描述.....5

4.数据需求.....6

4.1 wikipedia 实体及其关系.....6

4.2 微博 ER 图.....7

5.非功能需求.....8

6.运行需求.....9

6.1 硬件接口.....9

6.2 软件接口.....9

6.3 用户界面需求.....9

1.范围

1.1 标识

文档标识号：A2015-11-29-01

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：1.0

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：有提问需求者；

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

a.数据库(Database)：

MySQL: 一种关联数据库管理系统

Neo4j:一个高性能的,NOSQL 图形数据库

Redis:一个 key-value 存储系统，以超高效的查找存储著称，并具有数据持久的特性

b.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

c.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)

2.引用文档

a.书籍包括：

《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

b.本项目的经核准的计划任务书和合同、上级机关的批文:

第九届《大学生创新创业训练计划》

c.引用资料：

<Open Question Answering Over Curated and Extracted Knowledge Bases> from Google Scholar

<syntactic constraints on paraphrases extracted from parallel corpora> from Google Scholar

<基于维基百科的自动词义消歧方法_史天艺> 来自中国知网

<一个中文实体链接语料库的建设_舒佳根> 来自中国知网

3.功能需求

3.1 系统的用例模型，并进行简要的说明

图 1

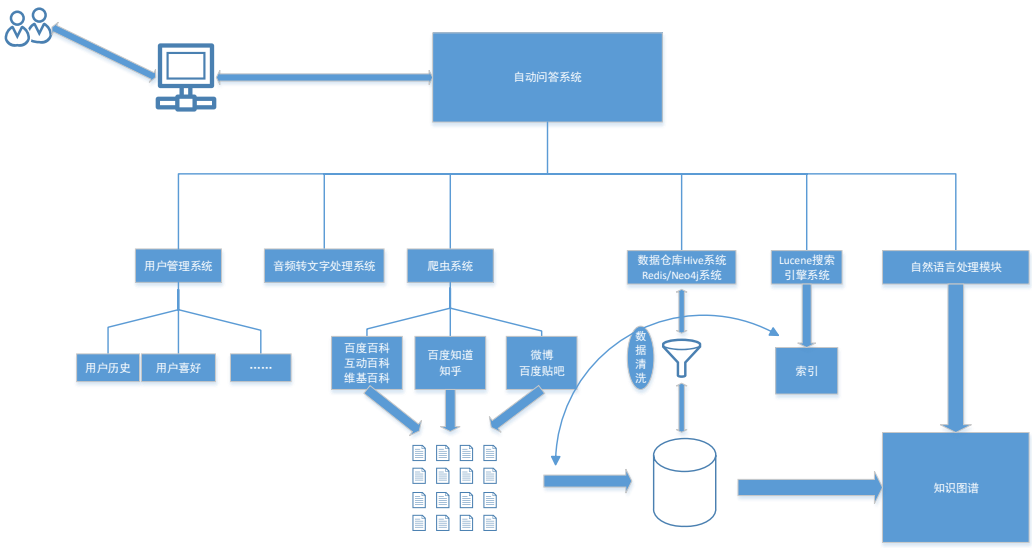
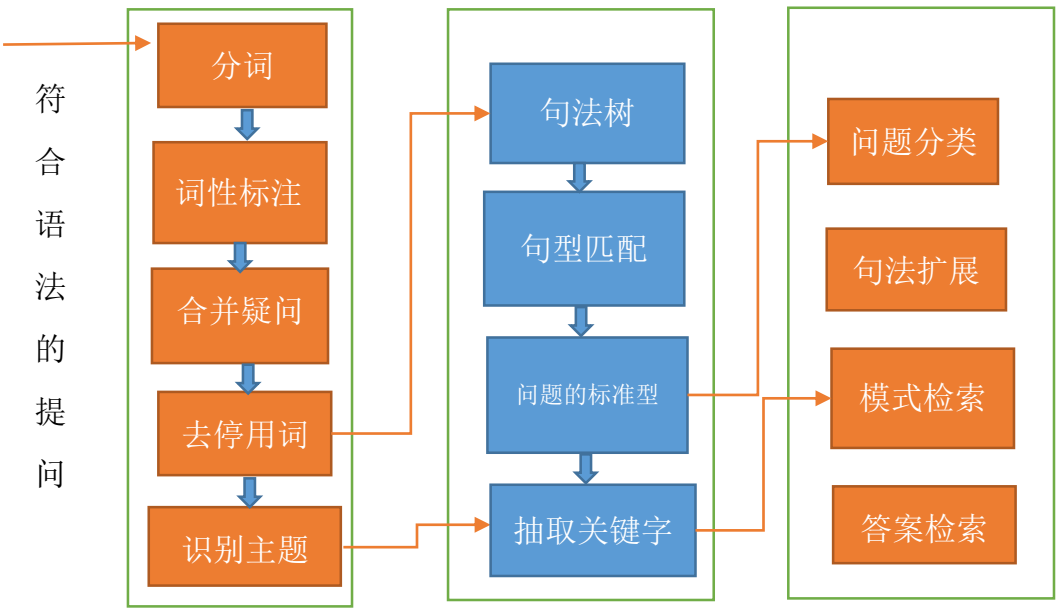


图 1 是本系统以最简介的形式解释我们的基于知识图谱的自动问答系统，其中包括了爬虫系统、基于 PageRank 的 Lucene 搜索引擎系统、分词系统、实体消歧系统、图数据库系统等等。由于过于简洁，不适合项目的展开，因此，必须对此进行详细的展开。图 2 针对自然语言处理的各个主要模板进行如下的详细描述。

图 1



3.2 对系统的用户进行系统的描述

用户可以通过自己的设计来设定喜好，主要分成以下几类人：

1、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；

2、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。

3.3 对系统的功能的系统描述

1、能够实现定时的对网络的信息进行爬去，以更新数据库，同时将常用的数据放在系统的内存中进行运行查询删除。

2、能够实现对所得数据的训练，训练包括以下几个方面:

一个是搜索引擎高效索引文件的训练；

一个是分词模型的训练（主要为 CRF 条件随机场的模板数据的训练）；

一个是多层神经网络模型的训练（主要为聊天语料库）；

一个是最大完全子图模型的训练（主要为语法模板的抽取作铺垫）；

3、能够与用户进行聊天；

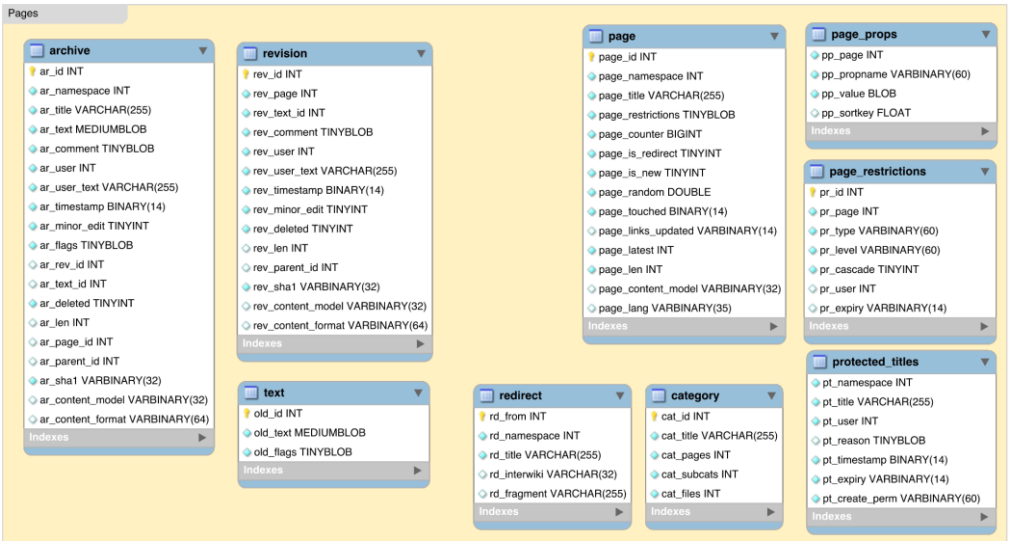
4、能够回答事实型问题、列举型问题、定义型问题和交互型问题等

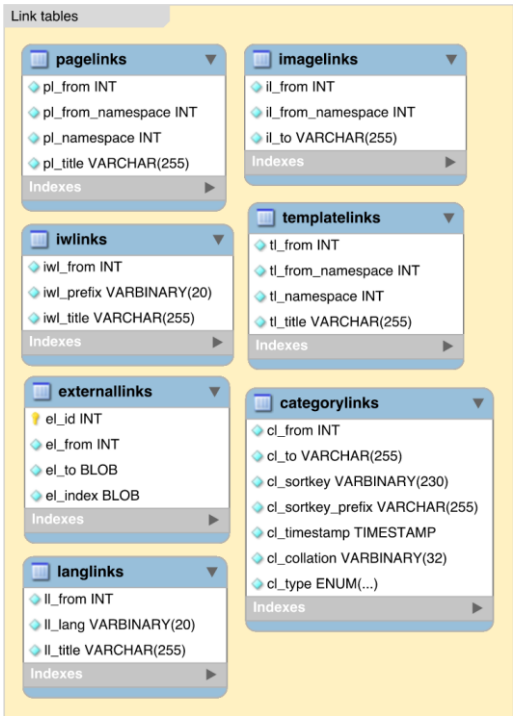
4.数据需求

本系统所利用的原始数据均来源于网上，鉴于不知道怎么抽取出和一般系统一样的关系，因此本系统举两个例子来说明问题。下面将以 wikipedia 为一个主要例子来说明数据需求（其中 wikipedia 中文版数据已导入数据库中），另外举例画出微博数据的 ER 图。

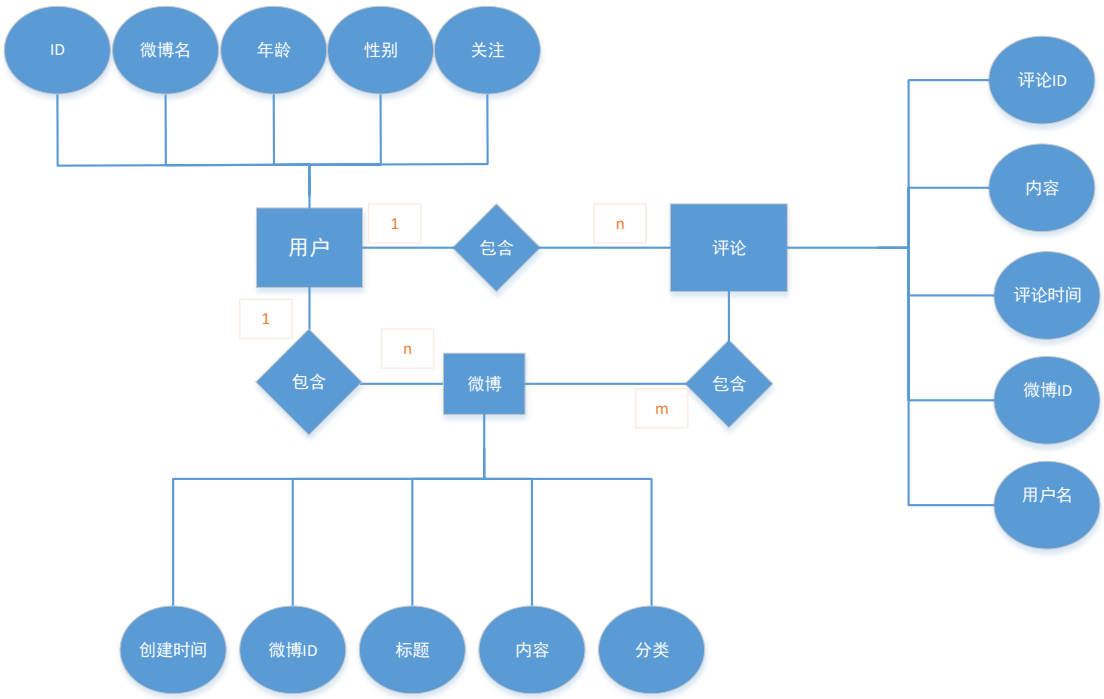
4.1 wikipedia 实体及其关系

数据的关系参见表格内的具体内容，图为本系统所需要的实体及其关系。下部分的图来自 https://www.mediawiki.org/wiki/Manual:Database_layout




Link tables
pagelinks
pl_from INT
pl_from_namespace INT
pl_namespace INT
pl_title VARCHAR(255)
Indexes
imagelinks
il_from INT
il_from_namespace INT
il_to VARCHAR(255)
Indexes
iwlinks
iwl_from INT
iwl_prefix VARBINARY(20)
iwl_title VARCHAR(255)
Indexes
templatelinks
tl_from INT
tl_from_namespace INT
tl_namespace INT
tl_title VARCHAR(255)
Indexes
externallinks
el_id INT
el_from INT
el_to BLOB
el_index BLOB
Indexes
categorylinks
cl_from INT
cl_to VARCHAR(255)
cl_sortkey VARBINARY(230)
cl_sortkey_prefix VARCHAR(255)
cl_timestamp TIMESTAMP
cl_collation VARBINARY(32)
cl_type ENUM(...)
Indexes
langlinks
ll_from INT
ll_lang VARBINARY(20)
ll_title VARCHAR(255)
Indexes

4.2 微博 ER 图


ID
微博名
年龄
性别
关注
用户
1
包含
n
评论
1
包含
n
微博
包含
m
评论ID
内容
评论时间
微博ID
用户名
创建时间
微博ID
标题
内容
分类

5.非功能需求

5.1 性能

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。限定为 1s 之内，即对于用户的每一次提问，可以在 1s 内给出回答。

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。
经检测，达到国内人工智能自然语言处理峰会 NLPCC-OPEN—QA 测试集 50%以上的准确率。

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。

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。
为提高本系统的应答的高效性，后期将采用 Hadoop 分布式运算来提高运算速率。

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、Linux 服务器

6.3 用户界面需求

（描述对该系统用户界面的基本要求，可以给出用户界面原型方案。）

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。

Demo 具体的地址为：http://pan.baidu.com/s/1mg927o8