北京航空航天大學
BEIHANG UNIVERSITY

<<基于知识图谱的自动问答系统>>
软件开发计划书

北京航空航天大学

2015-10

版本变更历史

版本	提交日期	主要编制人	审核人	版本说明
001	2015.10.30	杨东东		完成项目整体的规划
002	2015.10.31	方凯		完成内容补充及修改
003	2015.11.01	李睿霖		完成内容补充及修改

目 录

软件开发计划书

1 引言

1.1 编写目的

为了能按时完成目标，方便项目的管理还有组员了解项目，因此用文件化的形式，把对于在项目生命周期内的工作任务范围、各项工作的任务分解、项目团队组织结构、各团队成员的工作责任、团队内外沟通协作方式、开发进度、经费预算、项目内外环境条件、风险对策等内容以书面的方式叙述出来，作为约定。

本项目开发计划面向项目组全体成员，用于从总体上指导《基于知识图谱的自动问答系统》组员进行有效的工作还有时间安排，并且指明目的和需求。

1.2 背景

a.国际背景：

本项目是应国际会议 ACL、ICAMMP、TREC 的 AI 领域自动问答区而提出的观点，国内外均处于初步研发阶段，但国内的研发迅速更为缓慢，为最新兴的技术系统之一。现在国际背景是已经在 2002 年左右聊天机器人 ALICE(女)和 2005 年左右顶级网络机器人 *Jabberwacky*(男)，但是它们的发展都是基于大量的语料库而发展起来的，技术方法为完全匹配的方法，然而这显然不能满足人们对智能机器人的需求，于是在 2012 左右，IBM 的“沃森”机器人应势而生，这是目前世界上最智能的完整面向用户的机器人。

b.系统简介：

而本项目，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。

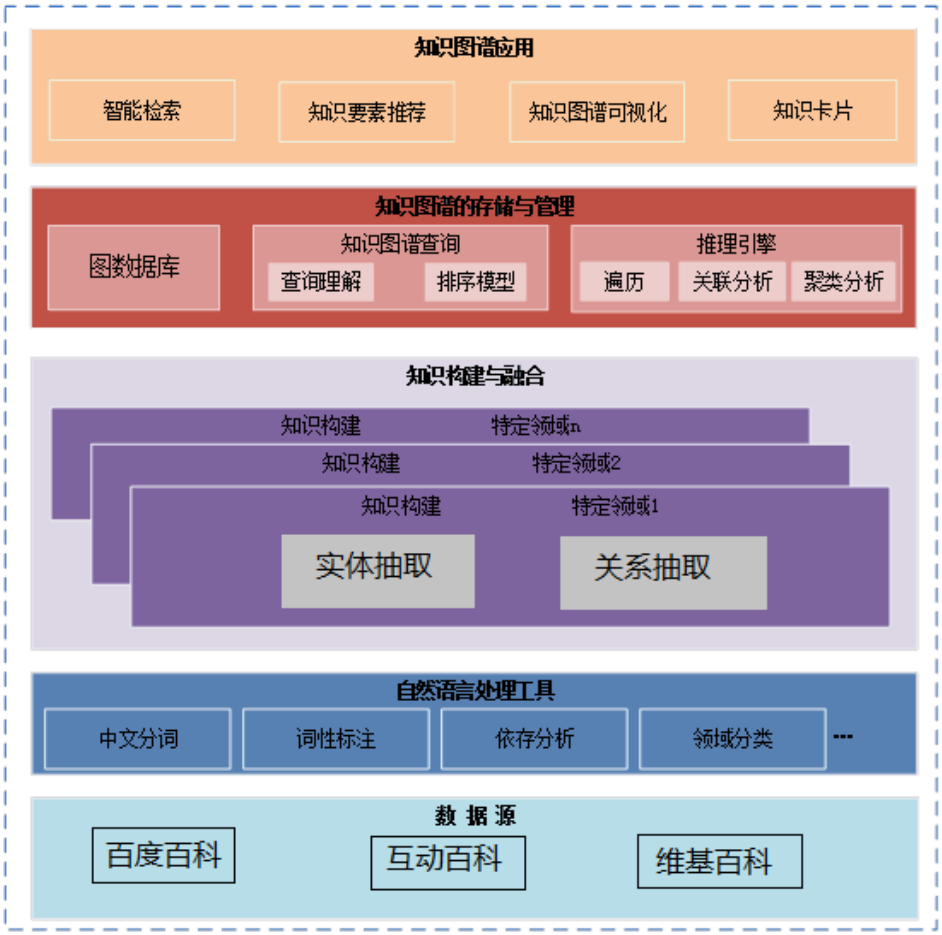
c. 来往关系：

本系统与搜狗-清华实验室有过来往，主要采集搜狗 2008/2012 年开放的网络数据，以用于开放式关系抽取，抽取网页中的关键信息，同时能作为数据源进行词向量、语法模板、实体消歧等的训练，同时依附于北航智能信息处理实验室。

同时爬取了百度百科的半结构化数据、爬取了动百科的分类词条数据、获取从维基开源的数据库中的 dump 数据。

d. 项目提出者为全体成员，委托者以及承担单位均为全体成员。

e. 项目流程如图：



1.3 定义

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

LSA: 隐式语义分析

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

SVD:矩阵奇异值分解

1.4 参考资料

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

2 项目概述

a.项目预期效果和体验：

项目目的就是开发一个能够用于知识类问答的问答系统。主要能够回答以下五种类型问题：

a.factoid（who is the wife of Obama?）

b.definition(whar is operation system)

c.yes-no (is Saddam Hussein alive?)

d.opinion (what do most Americans think of gun control)

e.comparison(what are the differences between Nokia and iPhone)

----来自《KBQA_keynote_CCIR2015》微软亚研院的周明教授所作的 2015 年度报告

这五类问题也是目前最能体现 KB—QA 作用的问题类型。

b.项目分解：

项目最主要分为如下几个部分：

1、数据爬取(百度百科、互动百科、维基百科)；

2、数据预清理（清除噪声类数据[错误数据、非常规数据]）；

3、数据预处理，全部整理成为三元组 Triple 的形式；

4、对于非及时性需求的数据存入 MySQL、对于及时性数据则存入 Redis，Neo4j 将用于处理数据中实体的逻辑关系而建；

5、Android 界面的搭建；

6、词向量的训练；

7、改进分词效果，使用隐式马尔科夫模型或者条件随机场；

8、优化分词效率，使用双数组 Trie 树的多模式匹配算法；

9、图算 entity-linking（实体消歧）；

10、问题类型识别（构建问题-答案模式）（抽取语法模型）；

11、词法分析；

12、语法树的构建图（节点关系）；

c.项目开发规定

开发环境:Eclipse for j2ee（用于开发 QA 的后台）、Android Studio（用于开发 QA 的前段人机交互界面）

开发语言：Java

附属开发需求：Mysql、Redis、Neo4J、Apache Tomcat、Linux 服务器

项目开发时间：180 天

2.1 工作内容

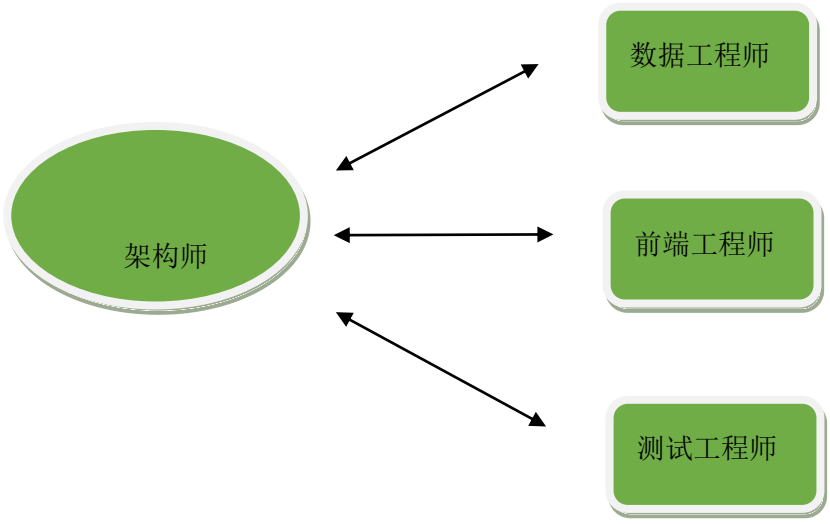
1、制作开发计划；

2、进行核心开发；

3、进行项目测试；

4、项目交付结题。

2.2 主要参加人员



姓名	职位	任务	技术水平
杨东东	架构师	负责计划、分配任务、算法设计	NLP 方向
李睿霖	数据工程师	高效爬取数据、清理数据、存储数据	DM 方向
方凯	前端工程师	前端设计、测试	Android 开发方向

2.3 产品

2.3.1 程序

软件名称：基于知识图谱的自动问答系统

编程语言： JAVA

存储方式: 阿里云服务器、Android 手机

功能和能力：

a.factoid（习近平的老婆是谁？）----Answer: 彭丽媛

b.definition(whar is operation system) ----Answer: 操作系统(operating system)是用户和计算机之间的界面. 一方面操作系统管理着所有计算机系统资源, 另一方面操作系

统为用户提供了一个抽象概念上的计算机

c.yes-no (钓鱼岛是中国的吗？) ----Answer:是

d.opinion (青少年应怎样长高) ----Answer:多吃饭多睡觉多运动

e.comparison(安卓手机和苹果手机的区别)

系统能够回答如上问题。

2.3.2 文件

a.可行性分析报告：说明该软件开发项目的实现在技术上、经济上和社会因素上的可行性，评述为了合理地达到开发目标可供选择的各种可能实施方案，说明并论证所选定实施方案的理由。
b.项目开发计划：为软件项目实施方案制订出具体计划，应该包括各部分工作的负责人员、开发的进度、开发经费的预算、所需的硬件及软件资源等。
c.软件需求说明书：对所开发软件的功能、性能、用户界面及运行环境等做出详细的说明。它是在用户与开发人员双方对软件需求取得共同理解并达成协议的条件下编写的，也是实施开发工作的基础。该说明书应给出数据逻辑和数据采集的各项要求，为生成和维护系统数据文件做好准备。
d.测试分析报告：测试工作完成以后，应提交测试计划执行情况的说明，对测试结果加以分析，并提出测试的结论意见。
e.项目开发总结报告：软件项目开发完成以后，应与项目实施计划对照，总结实际执行的情况，如进度、成果、资源利用、成本和投入的人力，此外，还需对开发工作做出评价，总结出经验和教训。

2.3.3 服务

人员培训、安装、保修、维护均可以由专业人员指导，而对于如下

技术支持：对于某些客户，采取上门指导的方式。

软件维护：获取软件使用中的问题，提供补丁程序。

软件升级：对于注册用户，只需较少的费用即可升级到新的版本。

服务期限：购买产品后的半年内。

服务级别：分为 VIP 用户和普通用户。

数据支付：将数据转交给客户运行管理维护。

2.3.4 非移交的产品

a.可行性分析报告：说明在技术上、经济上的可行性。

b.项目开发计划：为软件项目实施方案制订出具体计划。

c.软件需求说明书：对功能、用户界面及运行环境等做出说明。

d.详细设计说明书：包括实现算法、逻辑流程等。

e.测试计划：包括集成测试和验收测试等。

f.测试分析报告：对测试结果加以分析，和测试的结论意见。

g.源程序：软件开发过程中的全部代码以及注释。

2.4 验收标准

经用户和开发小组负责人双方签字确认的“需求规格说明书”。重点确认软件的可靠性、易使用性和功能完整性

2.4.1 代码的验收

在交付客户之前进行小组内评审，代码编写符合 HB6465 标准，与文档说明保持一致，代码书写风格统一，采用标准规范，没有下列错误：由于软件缺陷造成丢失数据，不符合设计要求，响应时间太长无法接受等问题。

2.4.2 文档验收

最后在交付客户之前进行小组内评审，文档格式符合 HB6465 标准，功能符合与客户的合同要求，清晰易读，没有语病与歧义。

2.4.3 服务验收

服务硬件达到文档说明的要求，人员技术考核合格，定期上门维护。

2.5 完成项目的最迟期限

从 2015 年 10 月 31 日开始至 2015 年 11 月 31 日，完成对整个系统的可行性报告分析、需求分析说明书、开发计划说明说、系统设计书、项目测试、项目总结，对概念模型、存储模式、完整性控制、存取权限等进行了定义，对系统功能各模块进行了详细设计，定义了数据库总体结构、编码命名规范，并交付用户。
交付日期为 2016 年 5 月 31 日，延期交付日为 6 月 15 号。

2.6 本计划的批准者和批准日期

组员签字： 日期：

组长签字： 日期：

经理签字： 日期：

3 实施计划

3.1 工作任务的分解与人员分工

工作内容	负责人	参加人员
可行性分析		

开发报告		
需求分析		
系统分析		
数据库建立		
界面设计		
测试计划		
测试报告		
项目开发总结报告		
用户操作手册		
后期维护		

3.2 接口人员

a. 负责本项目同用户的接口人员；

b. 负责本项目同本单位各管理机构，如合同计划管理部门、财务部门、质量管理部门等的接口人员；

c. 负责本项目合同负责的接口人员。

其中，接口人员，由软件发开方派专人，按客户要求，对指定地点进行安装，调试，运行并给客户演示。

3.3 进度

3.3.1 方法

采用迭代式的开发方式

3.3.2 模块开发优先级

里程碑名称	产品名称	提交日期	责任人
需求完成时			
详细设计完成时			
系统编码完成时			
工作完成			

3.4 预算

3.4.1 劳务费如下表

参与人员	时间（月）	预算（元）

3.4.2 经费如下表

办公费	500	差旅费	500
机时费	1000	资料费	500
通讯设备	500	专用设备	5000
总费用支出	8000		

3.5 关键问题

3.5.1 项目风险因素

风险排序	风险项名称	风险描述	风险缓解方案
1	专业基础知识不牢	开发过程中涉及的知识较多	增加学习时间
2	经验欠缺	开发经验不足	通过实战积累
3	缺乏指导	没有有效的指导人士大牛	请相关专家

3.5.2 影响本计划完成的主要问题

a.经费和硬件设施有限
b.用户需求不清，容易产生误解
c.开发人员没有实际经验
d.时间有限

4 支持条件

4.1 计算机系统支持

4.1.1 开发时的支持条件

a.硬件

1. 内存：8GB 以上；
2. 硬盘：至少 1TB 以上；

b.软件

1. 操作系统为 Windows 10
2. 集成开发工具 Eclipse,Android Studio
3. 数据库采用 MySQL,Redis,Neo4j

4.1.2 运行时需要的支持条件

a.服务器的要求

1．服务器内存必须使用服务器专用内存
2．为了保证数据存储的绝对可靠
3．服务器必须不间断电源。

b.服务器上应该配备的软件

1．操作系统：Ubuntu14.04 服务器
2．集成开发环境：LAMP

4.2 需由用户承担的工作

用户仅需配置好程序及知识库后在使用时输入所需要查询的语句，不需要完成其他工作。

4.3 由外单位提供的条件

需要由外单位配置一个大规模云服务器，该服务计划由阿里云提供。

需要由外单位提供一个完整的知识库及相关测试数据，该服务计划由 NLPCC 会议提供。

5 专题计划要点

5.1 测试计划

整个小组在项目开发完毕之后进行集中测试，测试内容包括：单一数据准确性测试，大数据量运算速度测试，服务器负载能力测试。计划测试时间为 2016-01；

5.2 质量保证计划

在开发中避免错误的发生，由小组成员相互监督与检查，严格按照项目开发过程中的各项步骤。保证最终项目运行时不会产生服务器停运等恶性错误，并且对于大部分数据能够输出较为正确且精确的运算结果。