**<<**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 |
| 1 | 2015.10.30 | 杨东东 |  | 完成项目整体的规划 |
| 2 | 2015.10.31 | 方凯 |  | 完成内容补充及修改 |
| 3 | 2015.11.01 | 李睿霖 |  | 完成内容补充及修改 |
| 4 | 2015.01.18 | 杨东东 |  | 完善补充第3部分表格空格 |
|  |  |  |  |  |
|  |  |  |  |  |

目录

[1 引言 1](#_Toc440906362)

[**1.1** 编写目的 1](#_Toc440906363)

[**1.2** 背景 1](#_Toc440906364)

[1.2.1国际背景 1](#_Toc440906365)

[1.2.2系统简介 1](#_Toc440906366)

[1.2.3来往关系 1](#_Toc440906367)

[1.2.4项目简介 1](#_Toc440906368)

[1.3术语和缩略词 2](#_Toc440906369)

[1.3.1数据库(Database) 2](#_Toc440906370)

[1.3.2自然语言处理(NLP) 2](#_Toc440906371)

[1.3.3矩阵论(Matrix) 3](#_Toc440906372)

[1.4引用文档 3](#_Toc440906373)

[1.4.1书籍 3](#_Toc440906374)

[1.4.2批文 3](#_Toc440906375)

[1.4.3引用资料 3](#_Toc440906376)

[2项目概述 4](#_Toc440906377)

[**2.1** 工作内容 4](#_Toc440906378)

[**2.2** 主要参加人员 5](#_Toc440906379)

[2.3 产品 5](#_Toc440906380)

[2.3.1 程序 5](#_Toc440906381)

[2.3.2 文件 6](#_Toc440906382)

[2.3.3 服务 6](#_Toc440906383)

[2.3.4 非移交的产品 6](#_Toc440906384)

[**2.4** 验收标准 7](#_Toc440906385)

[2.4.1 代码的验收 7](#_Toc440906386)

[2.4.2 文档验收 7](#_Toc440906387)

[2.4.3 服务验收 7](#_Toc440906388)

[**2.5** 完成项目的最迟期限 7](#_Toc440906389)

[**2.6** 本计划的批准者和批准日期 7](#_Toc440906390)

[3 实施计划 8](#_Toc440906391)

[**3.1** 工作任务的分解与人员分工 8](#_Toc440906392)

[**3.2** 接口人员 8](#_Toc440906393)

[**3.3** 进度 9](#_Toc440906394)

[3.3.1 方法 9](#_Toc440906395)

[3.3.2 模块开发优先级 9](#_Toc440906396)

[**3.4** 预算 9](#_Toc440906397)

[3.4.1 劳务费如下表 9](#_Toc440906398)

[3.4.2 经费如下表 9](#_Toc440906399)

[**3.5** 关键问题 10](#_Toc440906400)

[3.5.1 项目风险因素 10](#_Toc440906401)

[3.5.2 影响本计划完成的主要问题 10](#_Toc440906402)

[4 支持条件 10](#_Toc440906403)

[**4.1** 计算机系统支持 10](#_Toc440906404)

[4.1.1开发时的支持条件 10](#_Toc440906405)

[4.1.2 运行时需要的支持条件 11](#_Toc440906406)

[**4.2** 需由用户承担的工作 11](#_Toc440906407)

[**4.3** 由外单位提供的条件 11](#_Toc440906408)

[5 专题计划要点 11](#_Toc440906409)

[5.1 测试计划 11](#_Toc440906410)

[5.2 质量保证计划 11](#_Toc440906411)

[5.3安全保密计划 11](#_Toc440906412)

[5.4开发人员培训计划 11](#_Toc440906413)

# **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

### 1.2.1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。

### 1.2.2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。

### 1.2.3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。

### 1.2.4项目简介

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；

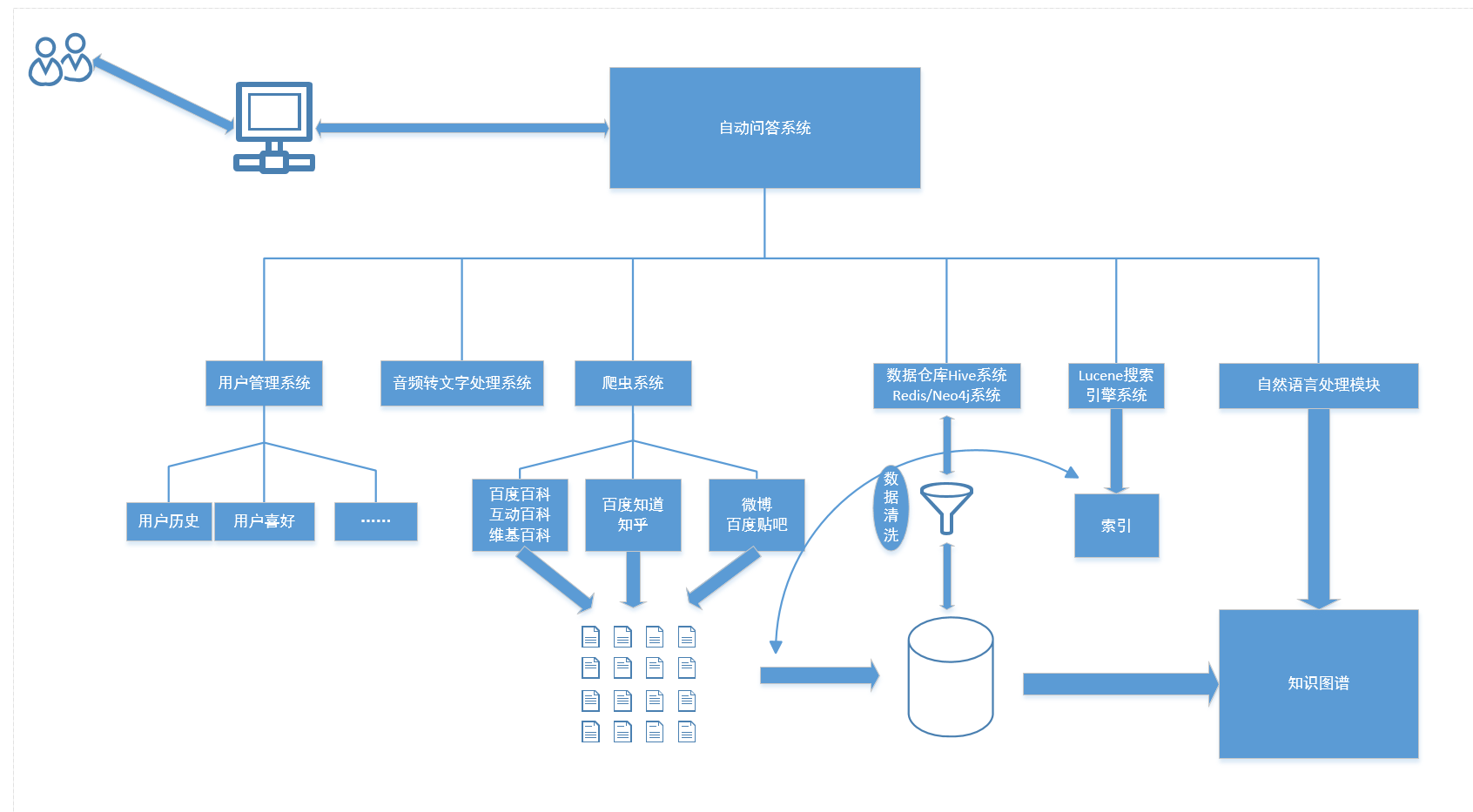
需方：用户；

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。

项目思路如图，原图见开发文档文件夹系统开发设计.vsdx



## 1.3术语和缩略词

### 1.3.1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

SQL：一种用于访问查询数据库的语言

SQL：Structured Query Language(结构化查询语言)

UML：统一建模语言、是一套用来设计软件蓝图的标准建模语言，是一种从软件分析、设计到编写程序规范的标准化建模语言

### 1.3.2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

### 1.3.3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

## 1.4引用文档

### 1.4.1书籍

（1）文档格式要求按照我国GB/T8567-1988国家标准和IEEE/ANSI830-1993标准规范要求进行。包括以下文件：

《软件工程国家标准文档》，《软件需求说明书编写规范》

（2）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

### 1.4.2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》

### 1.4.3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

# 2项目概述

（1）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。

（2）项目分解

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. 问答系统问题分析模块；
7. 问答系统信息检索模块；
8. 问答系统答案抽取模块；

（3）项目开发计划

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. 制作和修订项目开发计划；
2. 进行计划跟踪与监控；
3. 配合质量保证工作；
4. 工作产品及时进行受控管理；
5. 按计划提请阶段评审；
6. 提交测试部门评测开发产品；
7. 交付最终工作产品。
8. 项目实施总结
9. 项目验收

## **2.2** 主要参加人员

架构师

数据工程师

前端工程师

测试工程师

|  |  |  |
| --- | --- | --- |
| 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

### 2.3.2 文件

1. 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。
2. 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。
3. 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。
4. 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。
5. 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 | 杨东东 | 李睿霖、方凯 |
| 开发报告 | 杨东东 | 李睿霖、方凯 |
| 需求分析 | 杨东东 | 李睿霖、方凯 |
| 系统分析 | 杨东东 | 李睿霖、方凯 |
| 数据库建立 | 杨东东 | 李睿霖、方凯 |
| 界面设计 | 杨东东 | 李睿霖、方凯 |
| 测试计划 | 杨东东 | 李睿霖、方凯 |
| 测试报告 | 杨东东 | 李睿霖、方凯 |
| 项目开发总结报告 | 杨东东 | 李睿霖、方凯 |
| 用户操作手册 | 杨东东 | 李睿霖、方凯 |
| 后期维护 | 杨东东 | 李睿霖、方凯 |

## **3.2** 接口人员

1. 负责本项目同用户的接口人员；
2. 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；
3. 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 | 系统需求说明书 | 2015.11.30 | 杨东东 |
| 详细设计完成时 | 系统设计说明书 | 2015.12.30 | 杨东东、李睿霖、方凯 |
| 系统编码完成时 | 可运行系统 | 2016.06.01 | 杨东东、李睿霖、方凯 |
| 工作完成 | 项目总结 | 2016.06.30 | 杨东东、李睿霖、方凯 |

## **3.4** 预算

### 3.4.1 劳务费如下表 （人力成本3000/月

|  |  |  |
| --- | --- | --- |
| 参与人员 | 时间（月） | 预算（元） |
| 杨东东 | 10 | 30000 |
| 李睿霖 | 10 | 30000 |
| 方凯 | 10 | 30000 |
| 总共3人 | | 总计：90000 |

### 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

1. 经费和硬件设施有限
2. 用户需求不清，容易产生误解
3. 开发人员没有实际经验
4. 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

1. 硬件

内存：8GB 以上；

硬盘：至少 1TB 以上；

网络适配器：10MB/100MB自适应；

打印机一台

UPS(选配)

工作站：Pentium 4以上微机；

内存：512MB

硬盘：至少80以上；

CD－ROM：32倍速以上；

网络适配器：10MＢ/100MＢ自适应

网络： 至少一台服务器

至少一台工作站

使用TCP/IP协议的局域网

1. 软件

操作系统为 Windows 10/Ubuntu/Mac

集成开发工具 Eclipse,Android Studio

数据库采用 MySQL,Redis,Neo4j

### 4.1.2 运行时需要的支持条件

1. 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。
2. 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 2016-05

## 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。

## 5.3安全保密计划

在从项目开发阶段到最后软件的正式发布期间，做好项目的保密工作，小组成员对所有项目所有相关文档进行加密，做好备份工作。