****



**本 科 毕 业 设 计**

院 系 计算机科学与技术

专 业 计算机科学与技术

题 目 人机结合的领域实体信息提取和补全系统

年 级 2011级 学 号 111220051

学生姓名 孔令禹

指导老师 柏文阳 职 称 副教授

论文提交日期 2015.06

**南京大学本科生毕业论文（设计、作品）中文摘要**

题目： 人机结合的领域实体信息提取和补全系统

计算机科学与技术 院系 计算机科学与技术 专业 2011级

本科生姓名： 孔令禹 指导教师（姓名、职称）： 柏文阳副教授

摘要：

众包作为一种新的商业模式，缘起于互联网的迅速发展。这一概念虽然提出至今不过10年，但在许多领域中它已经越来越普及，成为了一种行之有效的科学方法。众包作为一种解决问题的模式，基于互联网的高效性，扩大了创新的边界，充分利用了外界的资源。众包的智慧就是普通大众的群体智慧。

在使用众包模式解决问题时，虚假和不准确的信息在众包过程中难以避免，但它们又往往会引起严重的后果，所以众包结果的质量控制已经成为了一个非常关键和亟待解决的问题。

本次毕业设计主要有两个目标：一是要完成一个引入了众包模式的高考招生系统，本系老师可以在一个良好的用户界面上查阅需要的相关资料，有助于他们的招生工作。另一个目标是设计一种全新的质量控制模型，它将在一定程度上有效控制低质量结果的产生。

关键词：众包；高考招生；质量控制

**南京大学本科生毕业论文（设计、作品）英文摘要**

THESIS: An extraction and completion system for domain entity information in a human-computer cooperation way

DEPARTMENT: Computer Science & Technology

SPECIALIZATION: Computer Science & Technology

UNDERGRADUATE: LingYu Kong

MENTOR: WenYang Bai

ABSTRACT:

Crowdsourcing, a new business model, is originated in the rapid development of Internet. Though there is only about 10 years since its appearance, it has become increasingly popular in many areas as an effective scientific method. As a problem-solving mode, crowdsourcing is based on the efficiency of Internet. It expands the boundaries of innovation and makes full use of the external resources. The wisdom of crowdsourcing is collective intelligence of the general public.

When using the crowdsourcing model to solving problems,the appearance of false and inaccurate information is inevitable during the process, which can often lead to serious consequences. In this case,quality control of crowdsourcing has become a very critical and urgent issue.

This graduation project is aimed at two main objectives. One is to built a system of college enrollment with the help of crowdsourcing model. Teachers can review relevant information within an user-friendly interface to help complete the enrollment. The other one is to design a brand new model of quality control. To some extent, it can effectively control the generation of low-quality results.

KEY WORDS: Crowdsourcing; College enrollment; Quality control

**目录**

第一章 引言

1.1 众包 1

1.1.1 众包的常见模式 1

1.1.2 大众接受众包的原因 1

1.1.3 众包兴起的原因 2

1.1.4 大众参与众包的动机 2

1.2 众包平台 3

1.2.1 常见的众包平台 3

1.2.2 PyBossa众包平台 3

1.3 高考招生系统 4

1.3.1 高考招生的重要性 4

1.3.2 高考招生的必要性 6

1.4 论文的组织结构 7

第二章 PyBossa平台

2.1 选择PyBossa平台的原因 8

2.2 PyBossa平台的安装 8

第三章 高考招生系统

3.1 高考招生系统的功能与结构 12

3.2 信息提取与补全的结构和整理 14

3.2.1 信息提取与补全的结构 14

3.2.2 信息提取与补全的整理 16

3.2.2.1 底层信息的拓展 16

3.2.2.2 多级众包的管理 17

第四章 众包结果的质量控制

4.1 质量控制的必要性 20

4.2 影响众包质量的因素 20

4.3 众包质量控制的方法 22

4.3.1 常见的众包质量控制方法 22

4.3.2 众包质量控制的难点 23

4.3.3高考招生系统的众包质量控制方法 23

4.3.3.1 与专业知识相关的算法 25

4.3.3.2 与声誉相关的算法 27

第五章 总结与展望

5.1 总结 28

5.2 展望 28

文献引用 Ⅰ

致谢 Ⅱ

**第 1 章 引言**

* 1. **众包：**

“众包”（Crowd-sourcing）的定义是一个公司或机构把过去由员工执行的工作任务，以自由自愿的形式外包给非特定的（而且通常是大型的）大众网络的做法。“众包”（crowdsourcing）这一概念是由美国《连线》杂志的记者杰夫·豪在2006年6月提出的。[1]但其所描述的商业实践，却早已存在。

1. **众包的常见模式[2]**

金钱报酬解决问题。这是最初也是最原始的众包模式，用户需要为其解决问题的人提供一定的报酬。基于这种模式的众包平台是现在最为常见的了，例如亚马逊的Mechanical Turk”和国内的威客网。

在大部分情况下，众包出去的是任务的设计和研发阶段，而对于最终的众包结果则是由发包方自己进行判断。www.Threadless.com这个由杰克兄弟创办的设计T恤的网站打破了这个常规，对于所有T恤设计结果的评判将会再度利用众包的思想，由用户选出他们喜爱的。相较于其他众包模式，这一模式更为全面和完善。

UGC是User Generated Content的缩写，意即用户生成内容，它将创造制作的工具交给用户，每一个用户都可以生成自己的内容，而网站只需要对其进行一定的规范和管理即可，它的主要收入则是通过投放广告来获得。这种模式是现在大家最为熟悉的众包模式了，像是国外的Facebook，YouTube，又或是国内的优酷网，人人网都是非常典型的例子，甚至百度贴吧，天涯论坛，新浪微博这些都可以算在这一类别中。  
 纯粹的志愿者项目。维基百科毫无疑问就属于这一众包模式。其实将维基百科归于上一类UGC模式也并无不可，那么为什么本文要单独再列出这一类别呢？一是因为维基百科的内容是完全免费提供的，任何一个志愿者，只要他有这方面的兴趣爱好，就可以参与到其中来，几乎不存在什么门槛，而且也完全不会涉及到金钱;二则是因为本次毕业设计所选用的众包模式基本上是沿用了维基百科的思想。

1. **大众接受众包的原因[3]**

众包的三个要素使它能够迅速被大众接受：一是用户的问题可以在较短的时间内得到解决，花费也在用户的接受范围内。二是当用户得到满意的答案时才需要付出报酬，有时甚至会是无偿的。三是众包是面向全世界的，用户可以接触到更多的资源。

1. **众包兴起的原因[4]**

众包的蓬勃发展主要有三方面的因素。

伴随着互联网的普及，无论是从时间还是从空间上来说，人们之间的距离都被极大的拉近了，大众参与到众包中来也成为了一件非常方便的事情。同时，互联网逐渐成为了大众进行信息交换最主要的平台，这些都是众包不断发展的重要推动力。

互联网改变了世界，人们的生活已经离不开网络，庞大的网民基数是众包出现和发展的必要条件。

马斯洛理论的最高层次就是自我实现，而参与众包的过程就是自我实现的过程。大众渴望能够展现自己的才华与能力，众包则为他们提供了这样一个平台。

1. **大众参与众包的动机**

大众参与众包的原因是多样化的，这其中有三种最为普遍：满足心理需求，金钱，学习新知识和技能。

大众参与到众包中的原因常常是出于自己的兴趣爱好，而并非是出于获取报酬的目的。但无论是有偿的还是无偿的，他们肯定都希望自己的想法和思路能够被别人接受，能够被别人认同。一个人如果设计研发出了一款产品，他绝不会满足于产品给自己带来的方便，他更希望自己的产品被大众使用后而获得赞扬，这样他就实现了自我的价值。

当然不是每个人都愿意无偿提供自己的好点子，也不是每个众包任务都能够攫取大家的目光，这种时候金钱无疑是吸引大众参与众包最为简单高效的因素了。

不单单是众包的发包方在这一过程中获得了利益，众包的接包方也可以说是受益匪浅。当身份背景大相径庭的用户共同参与时，众包网络中也就囊括了最为庞大的信息与知识，群体的智慧胜过群体中的任意一个人，用户在众包过程中将能够学习到许多新鲜的知识和技能，大众的智慧得到了普遍的提高。

* 1. **众包平台**

1. **常见的众包平台[5]**

在众包模式愈发火热的今天，越来越多的众包平台出现在人们的日常生活中，下面就来介绍一些知名的众包平台。

InnoCentive,创新中心网站创立于2001年。宝洁，杜邦，波音等世界500强公司都在这个网站中扮演了“寻求者”的身份，他们将各自的难题交给这个网站，等待用户们前来解决，最终被采纳的用户将获得一笔不菲的报酬。

与创新中心极为类似的是亚马逊的Mechanical Turk，他们都是现在最为成熟的商业众包模式。与其不同的是，在创新中心上寻求帮助的都是世界知名公司，而在Mechanical Turk更为常见的则是小型商户和个人用户。

Wikipedia，维基百科与上面两个平台不同，它是非商业化众包的一面旗帜。作为世界上最大的百科全书，它囊括了数百种语言，包含了数百万篇文章。作为一家非盈利性的网站，它的页面浏览量远超其他商业网站，而维护费用又相当之低。

Twitter Translation Center，推特翻译中心，与之相似的还有Facebook和Google的翻译社区，他们都提供了一个语言服务的平台供各自用户进行翻译和本土化的工作。

国内也有许多成功的众包案例，例如：知乎问答，人人猎头，微差事等等，但与国外还存在一定差距，有相当大的进步空间。

1. **PyBossa众包平台**

PyBossa是一个免费的，开源的众包微任务平台.它允许人们创建和运行在线协助执行需要人类认知，知识或智力的任务，例如图像分类，转录，地理编码等。PyBossa帮助调查者，民间黑客和开发者创建项目，世界上的任何一个有时间，感兴趣并且能够联网的人都可以贡献。

PyBossa的灵感来源于Bossa众包引擎，同时又是使用Python语言编写的，因此得名PyBossa。它可以用于任意的分布式项目任务，但最初它是用来帮助科学家和其他研究者采用众包技巧来解决问题的方式。

已经有大量的此类项目运用了PyBossa，包括像FreeBMD这样记录英国的民事出生，婚姻和死亡，并且长期运行的项目，PyBossa对它有极大的帮助;再比方说一个更近的项目，像是GalaxyZoo，它是极其成功的，在这个项目中，志愿者会努力的对由哈勃望远镜所拍摄下的照片进行星系的分类。[6]

通过PyBossa，人们想要尽可能容易的创建出潜在的，具有变革性的项目，因此PyBossa和现有的努力是不尽相同的。

PyBossa是100%开源的。

PyBossa和Mechanical Turk这种风格的项目不同，它不是依靠报酬或是金钱机制进行运作的，而是一个纯粹的志愿者项目。

PyBossa本身并不是一个众包应用程序，它是用来开发部署众包和小任务应用程序的平台和框架。个人众包应用程序被编写为简单的Javascript和HTML片段,然后部署在PyBossa实例上。通过这种方式，一个人可以很容易地使用PyBossa平台去存储数据，管理用户，处理工作流，以此来帮助开发自定义应用程序。

* 1. **高考招生系统：**

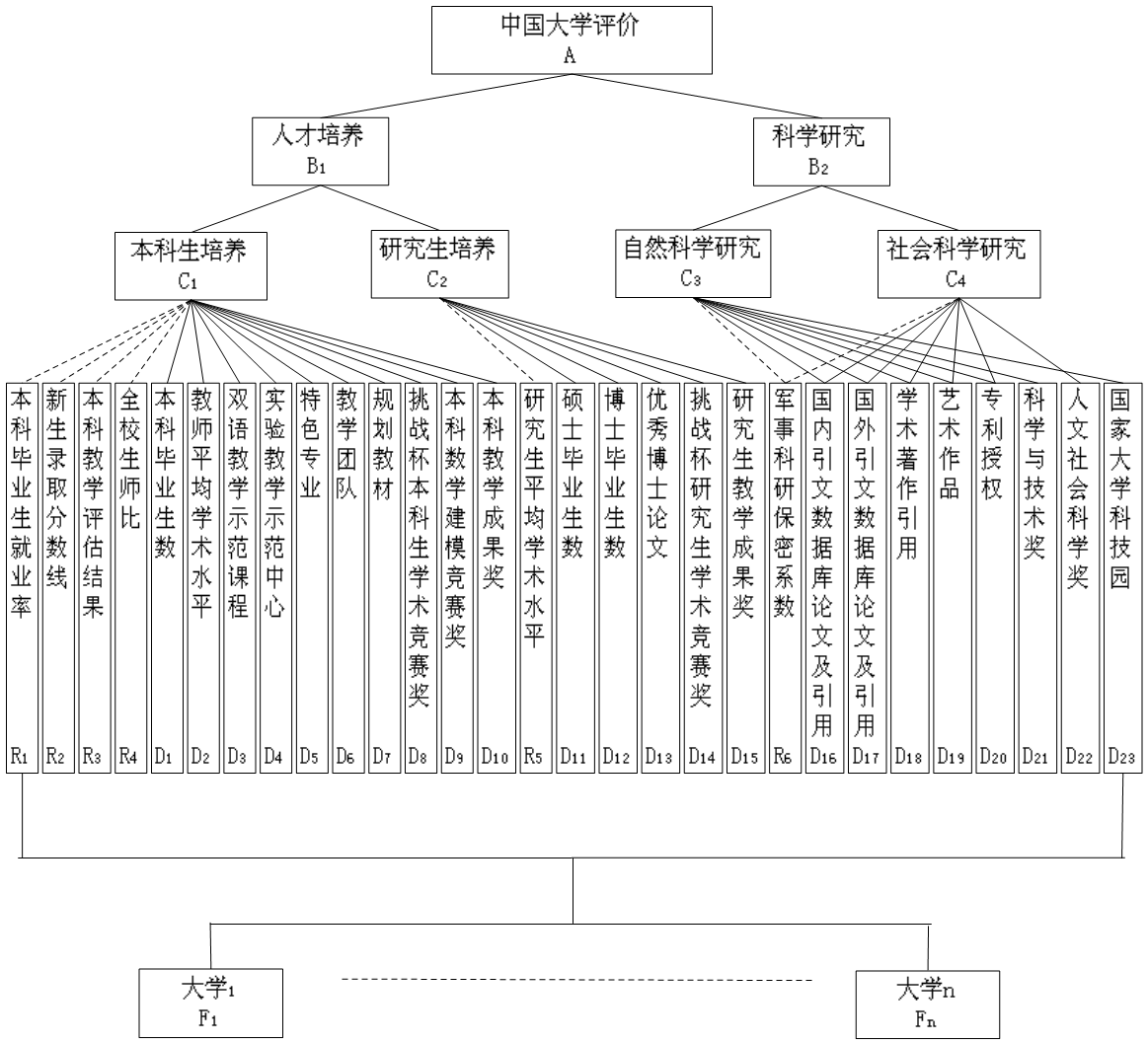
1. **高考招生的重要性**

高考是“普通高等学校招生全国统一考试”的简称，这场考试是学生步入大学的主要途径，它在一位学子人生中的重要性是毋庸置疑的，这里本文就不再过多的叙述了。那么学生在面对高考时，是否只要奋发学习，努力考出一个理想的成绩就足够了呢？当然不是，一个优秀的成绩当然是必要的，但其他方面也不容忽视，这其中非常关键的一点就是志愿填报了，有过切身体会的学生们自不必多说，便是早已过了高考年龄的人也时常会看到这样的新闻，或是高考成绩颇为理想，但填报志愿时过于谨慎而错过了更佳的学校，亦或是陷入自己的高考成绩未能达到填报学校最低分数线的尴尬处境。高考不但存在学生选择学校的难题，同时对于学校来说也可能是一场极为严峻的考验。

为什么要说高考的招生工作对于学校来说形式也极为严峻呢？大家都知道每年都会有中国大学排名这样的榜单出炉，虽然榜单的排名并不是完全准确的，但毫无疑问，这些榜单在高考生对学校进行选择时必然会产生极大的影响，那么每一张榜单必然有它自身的评价标准，每年的排名榜可以说是层出不穷，它们的评价标准自然也不尽相同，这里本文选取了中国校友会网和武书连这两个较为权威榜单的评价标准。

表1-1 2015中国大学排行榜评价指标及权重表[7]

|  |  |  |  |
| --- | --- | --- | --- |
| 一级指标 | 二级指标 | 三级指标 | 指标权重 |
| 人才培养 | 教学质量 | 教学水平 | 10.40% |
| 杰出校友 | 16.00% |
| 师资队伍 | 杰出师资 | 12.00% |
| 培养基地 | 学科建设 | 9.60% |
| 科学研究 | 科研成果 | 高端科研成果 | 16.00% |
| 科研基地 | 创新基地 | 9.60% |
| 科研项目 | 基础科研项目 | 9.60% |
| 社会影响 | 办学定位 | 办学层次 | 2.00% |
| 社会声誉 | 校友捐赠 | 5.00% |
| 生源竞争力 | 2.20% |
| 媒体影响力 | 3.20% |
| 国际影响 | 国际影响力 | 4.40% |



**图1-1 2010中国大学评价结构图[8]**

表1-1是由艾瑞深中国校友会网大学研究团队于2015年提出的中国大学排行榜评价指标和权重，图1-1则是由武书连等研究者在2010年提出的中国大学评价结构图。从这两幅图表中可以看出对大学进行排名时涉及到的因素是非常多的，可想而知学生在对学校进行选择时也会考虑到多方面的因素，这时招生工作的重要性就显露无疑了，其实招生工作就是宣传工作，招生工作者需要做的就是向学生宣传自己学校的信息，让学生能够更加了解学校，对学校的印象更加深刻，吸引质量更高的学生投报自己的学校。在这两幅图表中分别存在生源竞争力和新生录取分数线这样两个指标，这两个指标将会直接受益于招生工作的成功与否，那么高考的招生工作对于一所大学的意义也就不言而喻了。

1. **高考招生的必要性**  
    上文已经说明了高考招生工作的重要性，那么对于竞争力比较相近的两所学校来说，招生工作的好坏往往可能会导致两所学校的新生生源质量出现较大的差距，所以用战争来形容每年的招生工作一点也不为过。那么学校要打好这一场战争，策略是至关重要的。现在高考的招生是按照省份来划分的，每个省份的高考政策不尽相同，而同一个省份每隔几年高考政策又往往会发生改变，所以高考招生着实是一项复杂的工作。计算机科学与技术系每年都会负责南京大学在河南省的招生工作，所以毕业设计的系统暂时将只会面向南京大学在河南省的招生工作。那么哪些信息是招生工作者关心的，哪些信息又会对招生工作起到一定的帮助呢，本文认为一所高校主要会关心高中和其余高校的信息。

高中的信息。高中是大学进行招生工作时所面向的对象，但是单就河南省而言，高中的数量就有数百所之多，无论是从人力上，时间上或是效率上来说招生工作者都不可能亲赴每一所学校进行招生工作，那么这个时候就需要招生工作者从大量的高中中甄别出一小部分能够获得大量质量较高学生的，例如郑州市第一中学，开封高级中学等，并到这些高中去进行全面和具体的宣传。

大学的信息。其余的大学在进行招生工作时都是本校的竞争对手，当然招生工作者也并不会关注其余所有的大学，他们需要关注的主要是那些与本校实力相近的高校，因为这些高校会和本校争夺同一质量层次的生源，对于南京大学来说，需要关注的就是浙江大学，中国人民大学等等。

本文设计的系统中，本身将会存储有大量从网络中搜索出来的高中和大学信息，然后采用众包的方式，对本系统已经存在的信息进行修改和补全，最后招生工作者将可以按照他们的需要从本系统中得到需求的信息，以此来制定更为健全的招生宣传策略，使学校得到更高质量的生源。

* 1. **论文的组织结构：**

本篇论文主要分为四个部分，

第一部分为绪论，介绍本篇论文涉及到的背景知识和相关领域内容。

第二部分为PyBossa平台，介绍本文选择它的原因和它的安装流程。

第三部分为高考招生系统，介绍基于PyBossa平台设计的系统将会具备的功能和系统进行信息提取和补全的结构。

第四部分为众包结果的质量控制，介绍为什么要对众包结果的质量进行控制和如何对众包结果的质量进行控制。

第五部分是对本文的总结和不足之处的改进思路。

**第 2 章 PyBossa平台**

* 1. **选择PyBossa平台的原因：**

通过最初的引言相信大家已经对众包平台有了一个较为清晰的认识，本次毕业设计的目标之一就是设计一个基于众包平台的高考招生系统，所以众包平台的选择是非常重要的。在主流众包平台的基础上都可以比较轻松的设计出一个高考招生系统，但不可避免的会对系统有一定的要求和限制。最终本文选择了PyBossa平台作为毕业设计系统的基本框架，主要原因有以下三个：

上文中曾经提到过高考招生系统的设计思想与维基百科的模式是极为类似的，而PyBossa平台作为一个免费的，开源的众包平台，它的门槛是非常低的，只要用户有时间，有兴趣，并且能够连入网络，就可以自由的参与进来，这完全符合高考招生系统的本质要求。

第二个原因则是PyBossa的技术是较为成熟的，并且有许多项目已经使用了PyBossa平台，并取得了可喜的成果。它帮助大英博物馆和伦敦大学学院整理青铜时代的藏品，巴西空间研究所用它来评估热带雨林的砍伐程度，它也被用于伦敦纳米技术中心对纳米分子的分类工作。[9]

最为重要的一点在于PyBossa本身并不是一个众包应用程序，而是一个用来部署和开发众包应用程序的平台，这样高考招生系统就不会是一个一成不变的东西，它就没有了局限，具有了可扩展性，可以按照需求和喜好自定义的来开发一个众包应用程序。

* 1. **PyBossa平台的安装**

要知道PyBossa并不是一个应用程序，而是一个服务平台，所以相应的安装也就不会那么简易，PyBossa的安装是非常复杂的过程，在安装过程中，还会涉及到许多运用在不同领域的软件，在实际的安装过程中会遇到不少麻烦和困惑，最终安装成功后也将会学到了很多过去在课程中没有接触到的事物。

在登录PyBossa的官网之后，立即就能看到一个只需要5步的快速安装教程，它说在电脑上安装PyBossa最简便的方法就是使用虚拟机。使用这种方法，所有库的安装和配置过程都将会在虚拟机中完成，完全不会影响到日常使用的系统，当用户不再需要它们的时候，只需要删除虚拟机即可。[10]在最初使用这种方法进行安装的过程中出现了许多波折，虽然克服了诸多的难题，但始终无法成功的配置PyBossa。PyBossa平台是开源的，在github上可以寻找到它的源代码，于是开始阅读PyBossa的源代码。在结合源代码中的部分配置文件信息和安装过程中得到的错误反馈信息后，发现PyBossa的快速安装教程实在是过于简短了，这其中完全没有包括一些需要协同使用的软件信息。通过一段时间的查找之后，终于在PyBossa的官方文档中发现了一个相对健全的安装过程。下面就来介绍一下在安装PyBossa过程中会涉及到的一些资源：[11]

Ubuntu是一款基于Linux的免费开源桌面PC操作系统。PyBossa平台正是运行在Ubuntu系统之上的，现在PyBossa支持的系统版本为Ubuntu 14.04 LTS。

Git是一款免费、开源的分布式版本控制系统，用于敏捷高效地处理任何或小或大的项目。PyBossa使用git分布式版本控制系统来管理PyBossa服务器的源代码以及模板项目。

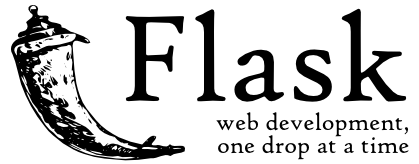
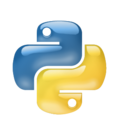
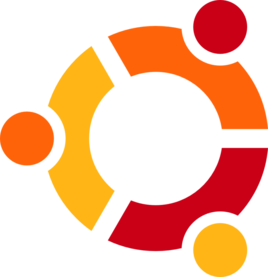
PostgreSQL是一个强大的、开放源代码的对象关系型数据库系统。它有着超过15年的活跃发展和经过验证的体系结构,它的可靠性、数据完整性和正确性也为其赢得了良好的声誉。PyBossa使用PostgreSQL作为存储所有数据的主要数据库。PyBossa支持PostgreSQL9.3以上的版本。

virtualenv是一个Python环境管理工具，它可以帮你创建一个独立的Python环境,你也可以使用它来管理不同的依赖关系和版本问题而不必在你的服务器及其中处理根权限的问题。virtualenv创建了一个拥有自己的安装目录,而不必与其他virtualenv共享库的环境，你甚至可以选择不访问全球安装库。如果你想要在系统级别安装软件，你就必须拥有根权限，但这常常又会导致你操作系统中的Python包出现破碎的依赖关系。如果你想避免这种糟糕的情况，virtualenv不失为一种明智的选择。当然，这不是你必需的步骤，不过为了你的电脑着想，最好学会使用它。

Python是一种面向对象、解释型计算机程序设计语言，语法简洁而清晰，具有丰富和强大的类库，这些优点让它成为了一门易读、易维护，并且被大量用户所欢迎的、用途广泛的语言。PyBossa的命名正是源于它是使用Python语言编写的。PyBossa支持Python2.7.6之后的所有Python2.x的版本。这里需要特别注意一个安装和管理Python包的工具pip，PyBossa要求它的版本须在6.1以上。

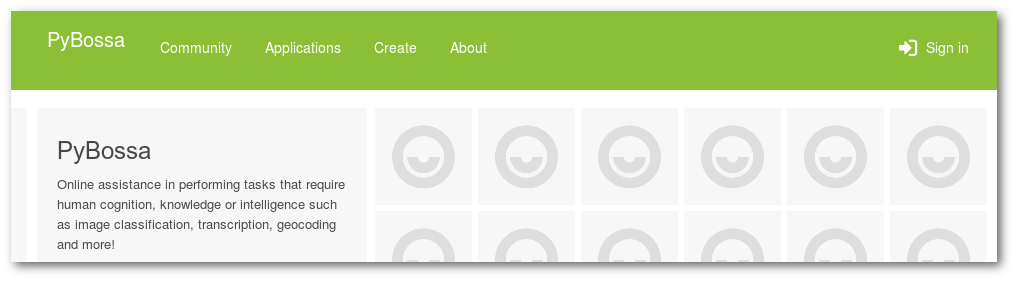
Redis 是一个高性能的key-value型数据库。Redis被PyBossa用来缓存对象，加速网站和限制API请求的使用。Redis-server和Redis-sentinel是必须要运行的两个实例，前者作为一个主节点用来接收读和写的操作，而后者则作为一个标记节点被PyBossa用来使得查询负载均衡和自动配置主从节点。Redis的安装相对来说比较复杂，要注意将关键文件拷贝到目标目录，在启动服务之后需要进行验证。[12]PyBossa支持Redis2.6以上的版本。

Flask是一个使用Python编写，基于Jinja2模板引擎，采用Werkzeug WSGI工具箱的轻量级Web应用框架。Flask于2010年年中开始被大众普遍使用，它主要面向一些需求简单，规模较小的微型项目。它会提供统一的小型工具，你并不需要进行过多的配置就可以快速的制作出一个简单的Web接口或是在一个已经存在的API上创建微型的网络接口。[13]PyBossa是基于Python-flask微型框架编写的。



**图2-1 PyBossa平台安装过程中涉及到的资源**

当对上述涉及到的知识有了深刻理解之后，再结合安装教程，就可以成功的安装PyBossa平台了。

**图2-2 安装成功后的截图**

**第 3 章 高考招生系统**

* 1. **高考招生系统的功能与结构**

 在成功安装PyBossa后，将会在PyBossa平台的基础上进行高考招生系统的设计，因为PyBossa平台本身只支持英语和西班牙语两种语言，所以汉化工作将是不可或缺的一步。PyBossa平台允许你注册为实名用户，你也可以不注册而只是一位匿名用户。当然实名用户比匿名用户的权限是要高的。为了能够提高众包结果的质量，本文设计的高考招生系统必须要进行注册，你只需要填写很简单的信息就可以成为注册用户了。系统为用户主要提供了两种不同的功能。

**图3-1 系统注册用户图 图3-2 系统创建项目图**



**图3-3 系统贡献任务图**

创建项目，此时你的身份是众包中的发包方。当遇到一些难以解决的问题时，你只需要创建一个项目，在项目中详细说明问题，然后你就可以等待其他对这个问题感兴趣的用户来帮助你了。

贡献任务，此这时你的身份是众包中的接包方。在你浏览高考招生系统时，你可以看到用户们提出的许多问题，在这些问题中或许会有某个是你感兴趣的，试着去解决它，然后把你的想法和做法提交上来，看看会不会被问题的提出者采纳吧。

在说明用户的两类主要功能时，文中使用了项目和任务两个不同的词汇，这两个词的意思有些相近，但在高考招生系统中确实是不同层次的两个概念，下面就来阐述一下系统的结构。

项目

任务

用户

创建

贡献

包含

类型

**图3-4 高考招生系统结构图**

根据图3-4可知高考招生系统主要分为项目，任务和用户这三个层次。用户在系统中提出了一个问题，这个问题就是一个项目，网站中肯定有大量的项目存在，为了便于网站的管理，系统会对这些项目进行分类，那些具有共性的项目将会被划分在同一类型当中，这同时也将有助于用户们寻找自己感兴趣的项目，但类型并不足以单独算作一个层次，而应该看作项目的衍生物。任务其实就是项目的一部分，可以将它看作是一个小项目，因为用户发布出去的一个项目可能会包含非常多的内容，其他试图来解决这个项目的用户也许只能回答其中的部分内容，那么此时他们就可以针对性的找到这一部分任务并加以解决。当然有些项目并没有太多内容，所以也存在项目和任务是一对一的情形。最后一个层次是用户，在创建项目时，用户面向的对象是项目；在贡献任务时，用户面向的对象则成了任务。任务和用户之间是多对多的关系，也就是说，一个任务可以由不同的用户来共同完成，一位用户也可以试图去完成许多的任务，无论这些任务是否属于同一个项目。

* 1. **信息提取与补全的结构和整理：**

1. **信息提取与补全的结构**

2.1节中曾经提到过，本文选择PyBossa平台最为重要的因素就是PyBossa只是一个框架，可以根据自己的需求来自定义的设计最终的应用程序。毫无疑问，原始的PyBossa平台不可能完全符合高考招生系统的设计需求。这个荐系统是为高考招生工作服务的，那么就来看看招生工作者需要哪些信息以及这些信息与PyBossa平台结合后的组织结构。

高中的信息。对于南京大学的招生工作者来说，他们可能比较关心一所高中的一本率，在省市排名靠前的学生有多少，今年高考工作负责人的联系方式等等。

大学的信息。在本次毕业设计中，主要关注浙江大学，中国人民大学等15所高校的信息。大学信息主要可以分为两个部分，一是招生计划，二是高考分数，而这两个部分又都包含理科与文科两类。其中理科的招生计划可以分为非国防生，国防生，贫困计划等等，更低一层又会包括专业、学制、计划招生人数、收费标准、专业方向等信息;理科的分数线也会具体到每个院系专业的最高分，最低分，平均分等数据。因为文科与理科在更细层次的划分方式基本一致，所以这里就不再重复了。

当然，无论是高中还是大学，高考招生系统都会存储一份基本信息。基本信息的内容主要为学校的人文历史，地理环境，教学水平，师资力量，办学层次，社会影响，杰出校友等等。

将这些信息与PyBossa平台本身的运作结构相结合之后，最终形成了如下图所示的信息提取和补全系统的结构。

**图3-5 信息提取和补全系统结构缩略图**

对图3-4和图3-5进行比照，其中图3-4的类型层次对应于图3-5第一层次的中学与大学，图3-4的每一个项目对应于图3-5第二层次中列出的每一所学校，图3-5第三层次根节点到叶节点整体对应于图3-4的任务层次，这里说明一下，由于大学招生信息的分类过于繁杂，图中将更低层次的细节省略了。

1. **信息提取与补全的整理**
2. **底层信息的拓展**

图3-5是一张缩略图，是因为它省略了大量的底层信息，这主要是出于两个方面的考量：一是在3.2.1节中本文已经用文字比较详尽的列出了底层信息，可以看出底层信息的种类是极为庞大和复杂的，将所有底层信息都在图中展示是一种不明智的行为；二是本文虽然已经列出了大量的底层信息，但这仍然不可能包括了全部的信息，甚至只是底层信息的一小部分。针对这种情况，本文在图3-5的基础上对底层结构进行了拓展与细化。

**图3-6 高中底层基本信息拓展图**

图3-6对应的是一下两种情形：第一种情况是一位用户作为接包方承接了人文环境这项任务，但这位用户在完成这项任务时出现了困难，此时他就可以将人文环境这项任务作为一个新的众包项目在高考招生系统中发布出去，由其他用户完成，最后他将从新的众包项目中总结和整理出的信息作为自己的众包任务提交即可；第二种情况是一位用户作为接包方承接了师资力量这项任务后，他认为师资力量这一项任务还可以进行更细的划分，可以分为正高级教师，特级教师，学科带头人等等。

1. **多级众包的管理**

3.2.2.1节介绍的方法实质是众包的拓展与延伸，即一个众包任务的子任务仍然可以使用众包模式去解决它，这种多级众包的难点在于理清一个任务的子任务有哪些，它又是否是从其他任务中划分出来的子任务。结合图3-5与图3-6，可以看出这两幅图都是一颗多叉树。所以多级众包的管理也就变成了树的编码问题，多级众包的难点就是通过编码判断一个节点是根节点，中间节点还是叶子节点以及节点之间的关系。下面具体介绍本文使用的多叉树编码规则：

1. 一棵树的每一级节点占两级编码，从编码00至编码98，编码99具有特殊含义。
2. 父节点的编码作为子节点编码的前缀。
3. 根节点在编码最后加上一位0。
4. 叶子节点在编码最后加上一位1。
5. 当一棵树的中间节点作为新的根节点进行拓展时，新根节点的编码在原编码的最后加上99后，按照前面4项规则继续编码。
6. 当一棵树的叶子节点作为新的根节点进行拓展时，先去掉叶子节点编码的最后一位1，再按照规则5进行编码

此处以图3-7为例，具体说明编码：

**图3-7 多级众包多叉树编码图**

下面说明在知道一项任务的编码之后，如何判定它在整个系统中所处的位置：

1. 若编码中包含99（第一个9在奇数位），则该任务是经过多次众包的，众包的级数为99的数目加一。
2. 若编码为奇数位且最后一位为0，则为根节点；若编码为奇数位且最后一位为1，则为叶子节点；若编码为偶数位，则为中间节点。
3. 在去除根节点和叶子节点的最后一位，以及所有编码99（第一个9在奇数位）后，若节点a的编码是节点b编码的前缀，则节点a是节点b的祖先，即任务b是任务a的子任务。

**第4章 众包结果的质量控制**

* 1. **质量控制的必要性：**

众包这一概念提出至今不过短短10年，但它的发展之迅猛令大家嗔目结舌，在人们的现实生活中，众包这一新兴模式似乎已随处可见。众包能在这极短的时间内被大众所接受并乐于使用，毫无疑问它成功迎合了时代的发展潮流。但是10年的时间还不足以让众包完全成熟，人们在使用众包途径解决问题时，一些麻烦和困难也往往会不期而至，关于众包的争议在社会上仍然是广泛存在的。这其中，最具有挑战性的问题之一就是众包结果的质量控制。一份低质量的众包结果显然不是发包方所期望的，如果长期得到的众包结果的质量都难以令人满意，这无疑将让众包模式陷入一个尴尬的境地。本文设计的高考招生系统采用了众包的思想，所以质量控制也就成为了本次设计必须要克服的一大难题。

* 1. **影响众包质量的因素：**

那么哪些因素会使得众包结果的质量参差不齐呢？下面这幅图给出了本文的想法。

**图4-1 众包结果质量因素分类图**

从上图中可以看出低质量众包结果的产生有两方面的因素，一是参与众包的人，二是众包任务的设计。

参与众包的人可能会在无意中犯下错误。再细心的人也难免会出现差错，一不小心打出了错别字，无意中删改了自己的结果，填写数字时出现了问题，这些情况都是有可能会发生，并且无法避免的。

参与众包的人也可能是受限于自身懂得的知识和掌握的技术水平，无法给出高质量的众包结果。人们参与众包的一个重要原因是兴趣，但是一个人对某个方面感兴趣，却并不意味着他对这个方面的事物就会十分了解和精通，同时参与众包的人群背景也可能会迥然不同，那么他们对于同一个问题有着截然相反的理解与回答也就不难理解了。

参与众包的人还可能受到利益的驱使，而恶意的提供虚假结果。这类人一方面想要获得任务完成后的丰厚报酬，另一方面却又并不愿意在具体任务上倾注心血，投入时间。他们会在很短的时间内对大量的任务做出回答，可想而知，他们提交的结果质量必然是无法令人满意的。

参与众包的人固然会犯下错误，而一个糟糕的任务设计也会导致众包结果质量的低下。

任务的定义其实就是发包方给接报方的关于众包任务的信息。一方面它会告诉你任务的实质和截止时间等等，另一方面它会对接包方有所限制，以此来保障你具备完成这个任务的资质。事实证明一个明确的任务定义将会有助于众包质量的提高。

接包方需要在一个用户界面上完成众包任务。如果用户界面过于简陋，那些致力于寻找系统漏洞，使用欺骗手段来完成任务的人就有了可乘之机。如果用户界面过于复杂，许多真心实意想要来完成任务的人又怕是要望而却步了，而且它会使得你的效率很低。所以，一个用户友好的界面是必不可少的。

任务的粒度。众包在最初被用来解决一些简单的问题，但现在越来越多的复杂问题也开始使用众包的方式寻求答案了。相对于简单的问题来说，解决复杂问题需要花费更多的时间，具备更全面的知识，所以很少会有人对它感兴趣，即使某个用户试图去解决它也很难得到合格的答案。把一个复杂的任务划分成多个简单的子任务，最后再将所有的子任务聚合起来，将会大大提高最终结果的质量。高考招生系统中项目和任务的关系，就是任务和子任务的关系。

人们参与众包的一个重要原因是金钱报酬。如果你的报酬够不上你任务的复杂程度，显然你的任务对人们的吸引力将是极低的。一味地提高报酬虽然会大大激发他人的兴趣，但这并不能保证结果的质量也会相应增加，相反你得到的答案也许会更加糟糕，因为那些恶意提供虚假结果的人最感兴趣的就是这一类报酬极高的任务了。所以报酬应当与你任务的难度相匹配，只有这样才能得到质量较高的答案。

综上所述，众包系统的质量是由接包方的个人素质和发包方的任务设计两方面共同构成的。对于高考招生系统来说，任务的定义完全取决于用户的喜好和水平，很难对其产生影响；任务的用户界面，高考招生系统将会尽最大努力去改善它。任务的粒度在本文第3章中也已经有了充分的考虑；高考招生系统同时沿用的是维基百科的思想，它是一个免费的，开源的平台，这个系统不会与金钱报酬相挂钩，也就不会涉及到任务设计的第四个要素，人们参与到这个平台中的原因将纯粹是因为兴趣。所以，下文中对于质量控制具体方法的研究，基本不会与任务设计方向有关，而主要是从接包方入手，通过提高接包方素质的方式达到提高众包质量的目的。

* 1. **众包质量控制的方法：**
     1. **常见的众包质量控制方法[14]：**

目前，众包质量控制的研究工作已经取得了一定的进展，下面具体介绍一下现在常见的几种众包质量控制算法。

使用黄金标准数据对众包结果的质量进行评判，黄金标准数据是指拥有标准答案的数据。收包方将接包方提交的结果与黄金标准数据进行比较，以此来鉴别众包结果的准确性。

使用最大期望算法（Expectation Maximization Algorithm）可以得到接包方提交结果质量的误差率，在此基础上消除接包方的个人喜好等会影响误差率的外部因素，得到最终的误差大小。

大量众包平台同时允许匿名用户和实名用户的存在。由于匿名用户的不确定性，这里主要考虑实名用户。根据实名用户以往提交的众包结果的质量，系统将会标识出最为可靠的一群实名用户，给予他们管理者的身份与权限，由他们共同承担众包平台的管理开销。发包方将任务交给这些管理者，再由管理者分发给各个接包方，管理者会对接包方进行测试，也会对接包方最终提供的结果进行检验。维基百科的模式与此类似。

* + 1. **众包质量控制的难点：**

尽管4.3.1节中已经介绍了一些常用的众包质量控制算法，但这些算法普遍会受到任务类型，金钱开销，接包方素质，时间等因素的限制。

首先，上述方法并不具备推广到任意类型众包任务中的能力，尤其是一些主观意念占主导的任务，例如对某样事物进行评价或是对一段文字进行翻译。其次，众包平台的监督管理能力不足，在很多情况下众包任务的接包方是匿名用户，系统很难对这一类人进行监管，无法杜绝恶意制造虚假结果现象的产生。此外，金钱报酬一直是一项极为重要的因素, 在4.2节中本文提到过，较高的金钱报酬并不一定由质量较高的众包结果与之对应,相反这一类任务对于恶意工作者的吸引力是极强的。 但是如果金钱报酬较低，可能需要等待相当长的一段时间，才会有接包方选择你的任务。黄金标准数据这一方法是目前使用较多的，然而创造一条黄金标准数据，在时间和金钱上也要付出一定的代价。众包任务的接包方遍布五湖四海，他们的身份背景各不相同，对于众包任务的理解也会产生偏差，最终提交的任务结果当然就大不相同。 现行的一些策略虽然取得了一定的成果，但都还有很大的改进空间。如果任务的参与门槛过高，这会大大降低人们参与的兴趣，如果对接包方的国籍加以限制，那么接包方的人数将会大幅减少，较低的金钱报酬伴随着的将是接包方较低的积极程度，而且这些策略对于恶意工作者的影响是有限的。 就目前而言，还没有一个全面并完善的方案来提高众包结果的质量。

1. **高考招生系统的众包质量控制方法：**

接包方的能力和质量将会影响众包任务结果的质量，那么要如何对接包方的素质进行判断，以此来保证可以寻找到素质较高的接包方呢？接包方的素质主要体现在他的专业知识和声誉两个方面。这两个因素是具有关联性的，不难想象，一个专业技能很强的接包方也将会具有很高的声誉。但本文仍然将这两方面区分开来了，因为从本质上来说，声誉可能比专业知识更能反映接包方能否完成特定的众包任务。专业知识体现在众包结果的质量上，而声誉的判断则要复杂的多，完成工作的时间、对众包质量的评估都会对声誉造成影响。所以声誉是基于整个系统的要素，专业知识则是基于单个任务的要素。举例来说，一个用户他对郑州市第一中学的信息非常了解和熟悉，并且具有很高的声誉，但如果让他去补全开封市高级中学的信息，那结果可能会令人大失所望。

接包方的专业知识告诉了大众他是否能够胜任某个特定任务，而专业知识又由资历和经验两个更细致的指标构成。接包者在众包平台进行注册时会填写许多的信息，这其中包括他获得的各种资格证书，使用的语言，所在的地区等等，这些信息共同构成了接包者的资历。经验则可以通过接包者已完成的众包任务有所了解，在完成众包任务的同时，接包者也会学习到更多新鲜的知识和技能，例如大量的程序员都在使用Stack Overflow这样一个关于IT技术的网站，无论是提出问题还是解答问题，对于他们自身来说都会有很大的进步。所以，完成的众包任务数量越多，质量越高，接包者的经验也就更加丰富。具体到高考招生系统来说，这个系统目前将主要开放给南京大学招收的历届河南考生和其他相关人士正是出于这一方面的考量。该系统是为南京大学在河南省的高考招生工作服务的，河南省的知名高中信息和其他高等院校的信息，历年来学校招生的河南籍学生无疑是最为熟悉的了，因为这其中既有他们的母校，也有他们曾经考虑过的其他知名学府。当然，如果这个高考招生系统显示出了它的作用，那么系统必然会逐步面向整个社会的人群，这个时候就需要使用其他的手段来提高接包方的质量了。

接包方的声誉。发包方对接包方的信任程度反映了发包方对于从接包方获得高质量众包结果的期望。而在众包系统之中，一个用户可能从未和其他用户进行过沟通和交流，此时对他的信任程度只能基于他在系统中的声誉了。假设某个用户已经完成了一定数量的众包任务，其他有权限的用户对这些任务进行评定，该用户的声誉将由综合评定结果决定。有时这种反馈是显示的，因为其余用户会对任务的完成度进行排名或是给出一个明确的分数。在另外一些情况下，反馈则是隐式的，例如当你在维基百科中创建或是修改一个词条时，它并不会立即生效，而是需要等待一位管理者的审核，因为这个审核过程不会公开，所以其他人也就无从得知你的成果是否被采纳了。

针对接包方的专业知识和声誉，本文分别设计了算法在众包的不同阶段保障众包结果的质量。

**图4-2 引入控制算法后的系统流程图**

* + - 1. **与专业知识相关的算法：**

为了尽量减少低质量的众包结果，并不是每一个人都可以参与到高考招生系统中来，对于所有有意对高考招生系统做贡献的人，都会对他们进行一项测试，只有通过测试的人才有权成为系统的注册用户，测试主要考验人们对于相关度的判断和排序能力。

第一轮测试只要对相关度进行判断即可。首先会有三类文章作为判断的对象，他们的标号分别为0,1,2，如果用户判断正确，将会得到较高的分值，如果用户判断错误，将会得到较低的分值。同时已知编号为0的文章最难判断，编号为1的文章其次，编号为2的文章判断难度最低。下面是具体的规则：

标号为0的文章：如果用户标记为0，得15分；如果用户标记为1，得10分；如果用户标记为2，不得分。

标号为1的文章：如果用户标记为0，得5分；如果用户标记为1，得12分；如果用户标记为2，得5分。

标号为2的文章：如果用户标记为0，不得分；如果用户标记为1，得9分；如果用户标记为2，得11分。

用户一共需要对5篇文章进行判断，那么用户得分的总和为：

（4.1）

其中是用户在第i篇文章上得到的分数。

假设用户判断对了全部5篇文章，那么用户得分的总和为：

（4.2）

其中是用户在第i篇文章上得到的分数。

最后定义每一位用户最终归一化的分数为：

（4.3）

本文将归一化分数的阈值定为0.8，只有归一化分数超过0.8的用户才可以进行下一轮测试。

第二轮测试需要在判断的基础上进行排序。假设第一轮中5篇文章正确的标记为0,1,2,0,1，那么它们的正确顺序应该为3,2,1,3,2（越容易判断的文章排名越高，数值越小），假设用户对它们的排序为2,5,3,1,4。下面是一位用户最终归一化分数的计算方法。

（4.4）

其中为每一篇文章的正确标记（即0,1,2,0,1）

（4.5）

其中为用户对文章的排序（即2,5,3,1,4）

（4.6）

其中为文章的正确排序（即3,2,1,3,2）

（4.7）

SNS即为最终归一化的值，本文将SNS的阈值定为0.6，所有分值超过0.6的用户将被允许参与到高考招生系统中来。

* + - 1. **与声誉相关的算法：**

在高考招生系统中，其他用户有权对你完成的任务进行评价。每位用户在系统中最初的声誉值为0.5，对其余用户的评分应为0与1之间的值。在你的任务被其他用户评价之后，你的声誉值将发生变化，计算方式如下：

（4.8）

其中n为评价你这个任务用户的总数，为编号为i的用户的声誉值，为编号为i的用户对这个任务的评分。

（4.9）

其中为更新前的声誉值，为更新后的声誉值。

从公式4.8中可以看出，你的声誉值对你的评价也是有影响的，声誉值较高的人相对于声誉值底下的人来说，他的评价更有可信度，更具有价值。而从公式4.9中可以看出，决定你声誉值变化的不是高考招生系统中用户的平均声誉，而是你在提交这个任务之前的声誉值，因为即使一个人提交的结果仍然水平较低，但只要与之前相比有所进步和成长，就应当认可他的努力，提高他的声誉值。相反，若是一个人以往展现的水准极高，而在某次任务中被认定为中等水平，也应该降低他的声誉值。所以，要获得更高的声誉值，就要尽最大的努力在高考招生系统贡献任务，超越自我。

**第 5 章 总结与展望**

* 1. **总结：**

竞争日益加剧的高考招生工作已经成为了各大高校每年都必须直面的问题。本文的主要目的之一就是为高考招生工作服务，向招生工作者提供他们需求的信息，帮助他们更好的完成招生工作,因此在PyBossa平台的基础之上，搭建了一个基于众包模式的高考招生系统。本文的前三章主要介绍了众包的思想和PyBossa平台的使用方法，高考招生系统的功能与结构以及信息提取与补全系统的结构和整理。众包在网络环境中已经存在了很多年，它的可信度一直是大众非常关注和尚未彻底解决的问题。本文的第四章详细分析了低质量众包结果产生的原因，现行的一些众包质量控制方法以及众包质量控制的难点所在，并且在众包开始之前和众包过程中分别提出了一种提高众包质量的具体方法。可以预见，在开始实行这两种算法之后，众包结果的质量将会得到一定程度的提高，该系统也会在高考招生工作中起到不可忽视的作用。

* 1. **展望：**

一方面，PyBossa平台是使用Python语言中的一个微型框架flask编写的，由于对Python语言的生疏和时间的不足，经过修改后的PyBossa平台在界面的友好程度和功能的多样化方面，与理想中的高考招生系统仍然存在一定的差距，在之后需要对系统进行进一步的修改。

另一方面，高考招生系统暂时只进行了α测试，还没有能够正式上线提供给广大用户使用，因此虽然文章中提出了两种控制众包结果质量的算法，但只是停留在了理论阶段，等到系统正式上线运行之后，将会使用得到的具体数据进行实际检验，以此来对算法的可行性等方面进行分析和修正。

**文献引用**

[1]百度百科，众包，http://baike.baidu.com/view/729695.htm，2015-03-03.

[2]人人小站，众包综述，http://zhan.renren.com/h5/entry/3602888498035058715, 2012-12-25.

[3]WIKIPEDIA A.Crowdsourcing[EB/OL].

http://en.wikipedia.org/wiki/Crowdsourcing, 2014-11-17.

[4] 张利斌, 钟复平, 涂慧. 众包问题研究综述[J]. 科技进步与对策, 2012, 29(6): 154-160.

[5]沈备军，众包—群体工作的未来，http://blog.sina.com.cn/s/blog\_4cc6846d0102v23z.html, 2014-08-24.

[6]Open KnowLedge Blog,Introducing PyBossa–the open-source micro-tasking platform, http://blog.okfn.org/2012/06/08/introducing-pybossa-the-open-source-micro-tasking-platform/, 2012-06-08.

[7]新浪教育，2015中国大学排行榜评价指标体系概述，

http://edu.sina.com.cn/gaokao/2015-02-04/1120456986\_3.shtml, 2015-02-04.

[8] 武书连, 吕嘉, 郭石林. 2010 中国大学评价[J]. 科学学与科学技术管理, 2010, 31(4): 5-13.

[9]PyBossa, Powered by PyBossa, http://pybossa.com/, 2015-01-25.

[10]PyBossa, Getting Started, http://pybossa.com/getting-started/, 2015-01-30.

[11]PyBossa, Installing PyBossa—PyBossa v0.2.2 documentation，http://docs.pybossa.com/en/latest/install.html, 2015-02-04.

[12]Linux公社，Ubuntu12.10下安装Redis（图文详解）+Jedis连接Redis，http://www.linuxidc.com/Linux/2013-06/85816.htm, 2013-06-11.

[13] 开源中国社区，选择一个 Python Web 框架：Django vs Flask vs Pyramid，

http://www.oschina.net/translate/django-flask-pyramid, 2014-11-06.

[14] 张志强, 逄居升, 谢晓芹, 周永. 众包质量控制策略及评估算法研究[J]. 计算机学报, 2013, 36(8): 1636-1649.

**致谢**

首先我要感谢我的指导老师柏文阳副教授。在完成实验，编写论文的过程中，我遭遇了许多困难和挫折。柏老师一直不厌其烦的为我指点迷津，他严谨的科学态度，扎实的知识储备，开阔的视野思路令我受益匪浅。在此对柏老师表示我最衷心的感谢。

其次我要感谢杨勇学长和徐诚同学。从PyBossa平台的搭建到控制众包质量算法的构思，这些都离不开学长的无私指导与帮助，徐诚同学则对我论文的写作起到了巨大的作用。还有我的好朋友，黄茂森、黄羿、蒋天博等人，在我的生活和学习中给予了我许多关怀与帮助，在此对大家表示感谢。

最后我要感谢我的父母，是他们对我的理解和支持，鼓励与期许让我能够顺利的完成毕业论文。