# “知乎数据爬虫”软件系统的需求构思及描述

## 背景介绍

随着互联网的普及和发展，现在网民数量已经达到了9.89亿人，其中相当一部分人对知识是十分渴求的，所以知乎这类问答平台变得火爆起来了。知乎作为一个问答软件，许多人将其当做阅读、学习、和业内人士沟通的工具。知乎的月活跃用户高达9430万。虽然知乎的用途广、用户多，但是缺少用户分区排行、用户检索系统、绘制统计图。

## 欲解决问题

知乎的缺点如下：缺少垃圾回答屏蔽系统；答主排序、检索系统；信息搜集、可视化系统。

在知乎查资料的时候经常会遇到垃圾回答，比如广告、引战、辱骂、搞笑段子。这些问题可以通过数据清洗来去除，因为垃圾信息往往同质化严重，当重复度高于某一数值时，就可以列为垃圾回答。

答主排序检索系统是可以将所有答主的人气信息搜集起来的系统，人气包括：赞数、评论数、粉丝数、发表数、收藏数、转发数、引用数。通过这些数据综合排序就可以或者相对高质量的作者。

现在知乎想要知道某板块的热度或者某些热词的出现频率，只能通过购买数据搜集网站的服务来统计热度。而我们可以解决它，用项目来给用户带来便利。

## 软件创意

我们的软件创意点在可视化数据，通过云图、动态柱状图等多种形式的可视化，增加了该软件的趣味性和直观性。趣味性指的是将枯燥的数字和表格变为更加生动的柱状图、饼状图、折线图，甚至可以生成云图，将高频词放在图的中央，越高频字体越大，而相较于出现频率更低的词则会排在高频词的周围或缝隙中，字体也会小得多。

## 系统的组成和部署

介绍系统由哪些要素（包括计算机软件、其他设备、服务和系统等）组成，这些要素间的相互关系，如何部署这些系统要素，它们需要依赖哪些软硬件系统或者计算基础设施。

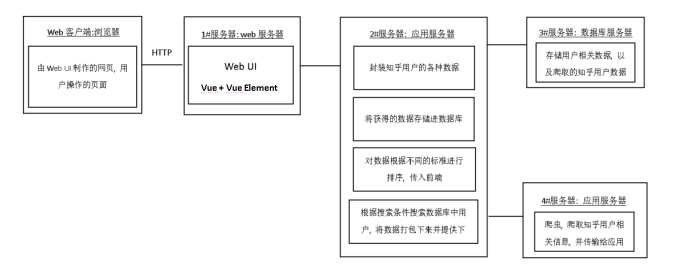


图 1 软件部署图

本软件客户端主要以web客户端为主, 用户可以利用浏览器登录网页并通过相应的操作来执行不同的功能, Web客户端的前端以Vue + Vue Element架构来编写, 制作相应的界面, 安卓也可以通过浏览器进入。

后端则以Spring Boot + Maven + My Batis架构编写, 作为整个程序的中介。 通过jbdc与数据库进行连接, 从爬虫程序爬取的数据进行整理存储进入数据库, 在展示时, 通过利用不同的排序方式, 将数据展示到网页上。

爬虫则主要以python为主, 从知乎上爬取用户数据以支持本程序运行。

## 软件系统的功能描述

### 5.1 软件功能分析

1. 对百万级数据进行爬取、清理、分析、统计及可视化，做到短周期内同步更新，确保数据的真实性以及有效性。
2. 对分析数据进行图标可视化展示
3. 查看相关数据排名
4. 对知乎中用户的详细数据（应包含其回答，视频，专栏等一切可见数据）提供下载功能，供使用者进行下载（数据会以文件包的形式提供给用户）
5. 简单知识图谱构建

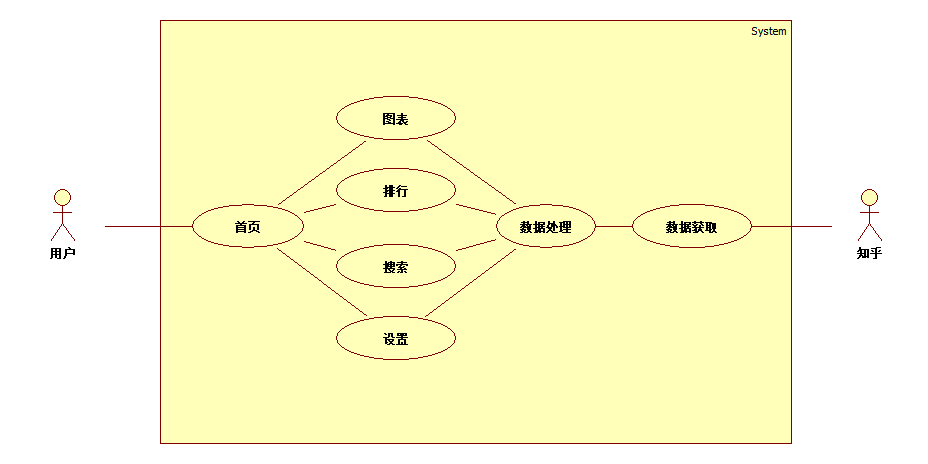


图1.1

***1、用户进入首页***

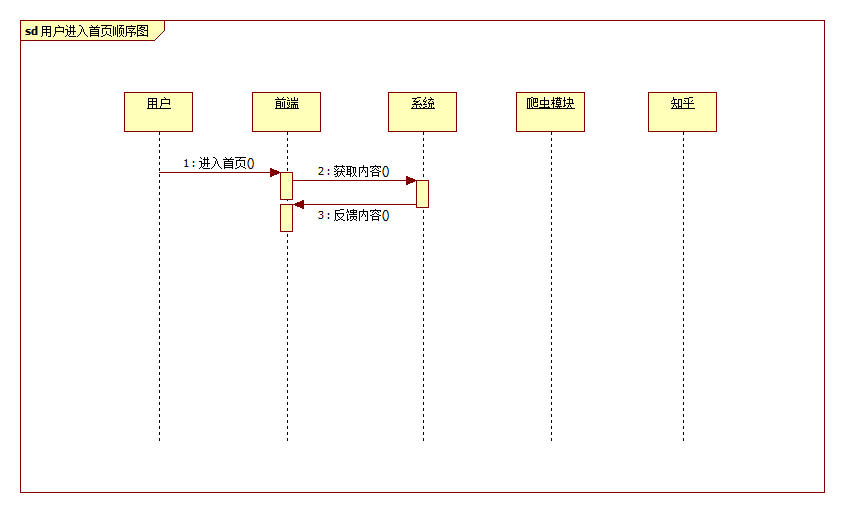


图2.1

|  |  |
| --- | --- |
| 用例名 | 用户进入首页 |
| 用例描述 | 用户进入首页 |
| 参与者 | 用户 知乎 |
| 过程 | 1.用户进入首页  2.系统获取信息  3.系统反馈信息 |

### 5.2 软件功能列表

根据系统的use case图，识别和描述软件系统的各项功能，说明其工程特征

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 编号 | 功能名 | 功能描述 | 优先级 | 重要性 | 工作量 |
| 1 | 搜索查看指定用户 | 搜索查看指定用户信息 | 第二次迭代前完成 | 高 |  |
| 2 | 下载相关内容 | 对指定用户回答等其他内容进行下载 | 第二次迭代前完成 | 中 |  |
| 3 | 查看数据统计 | 查看可视化数据统计 | 第一次迭代前完成 | 高 |  |
| 4 | 查看知乎用户关系 | 对存在关系的知乎用户进行关系整理，存为图知识库 | 第三次迭代前完成 | 中 |  |
| 5 | 推荐商品整理 | 对知乎用户推荐的商品进行整合对比 | 第三次迭代前完成 | 中 |  |

表2：软件系统的各项功能及描述

## 可行性及潜在风险

该项目可行性是非常高的，第一，有前人完成的例子可以进行学习，各网站也有足够多的相关教程。上述框架已经有相当大量的应用于此方面的先例，在兼容性方面是不用担心的。组员分工明确，各个组员工作量适中，不用担心猝死，而且可以高效率工作。

潜在风险：1.隐私泄露。我们的项目会将数以百万计的答主的数据收集来，数据量会达到TB级别，这些数据一旦泄露，我组按照法律会被罚款。信息是有归属权的，我们大量爬取信息的行为有可能面临知乎公司的起诉。2.性能骚扰。过于大量的信息涌入校园网服务器会导致网络拥挤、瘫痪。造成其他同学的使用不便。3.封号风险。知乎网站会在用户登录后对cookie的调用进行检测，如果调用次数过于频繁，账号有可能会被封禁。

针对隐私泄露的风险我们计划使用加密数据库对数据进行保存，装有数据的电脑只有经过项目负责人允许才能进行使用。第二项可以使用robots协议来限制我们爬虫的大小，防止校园网的路由器过载。第三项我们会尽量限制对cookies的调用次数，或者使用某种软件来躲避网站对cookies的监测。