# “百万级知乎用户数据分析与搜索系统”

# 软件系统的需求构思及描述

## 背景介绍

知乎是中文互联网高质量的问答社区和创作者聚集的原创内容平台。知乎以问答业务为基础，经过近十年的发展，已经承载为综合性内容平台。

知乎也面临着一些问题，一旦有人从中灌水，就会引发一连串的连锁反应，这就是[破窗效应](https://baike.baidu.com/item/%E7%A0%B4%E7%AA%97%E6%95%88%E5%BA%94" \t "https://baike.baidu.com/item/%E7%9F%A5%E4%B9%8E/_blank)。知乎也意识到这种威胁，一方面继续采用严格的邀请制度，另一方面采取了许多措施来抑制破窗效应。比如对提问类型进行了限制，允许用户举报重复低质量的问题和隐藏“没有帮助”的答案等。但是，破窗效应的威胁依然存在，随着不断开放和用户数增长，如何保持回答的质量和社区的氛围依旧是不小的威胁。

## 欲解决问题

解决数据质量驳杂冗余的问题

解决有效数据的获取困难问题

## 软件创意

按人气对答主进行排序

下载答主回答及其他内容

知乎用户数据分析可视化

## 系统的组成和部署

爬虫python

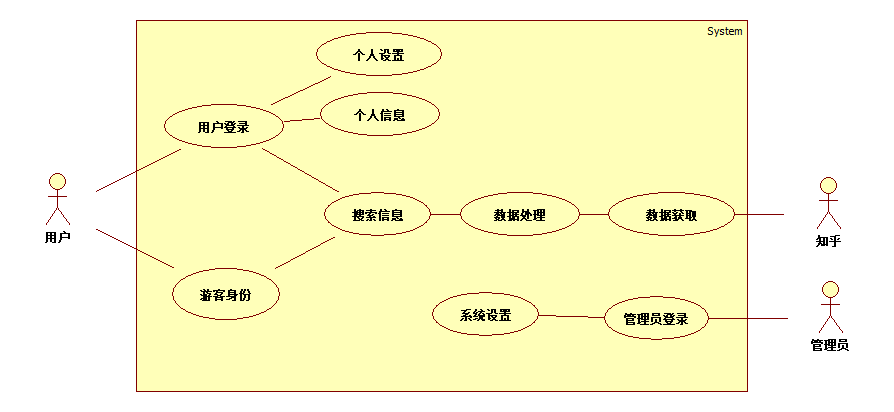
后端 spingboot+mybits+maven

前端 vue+element+echarts

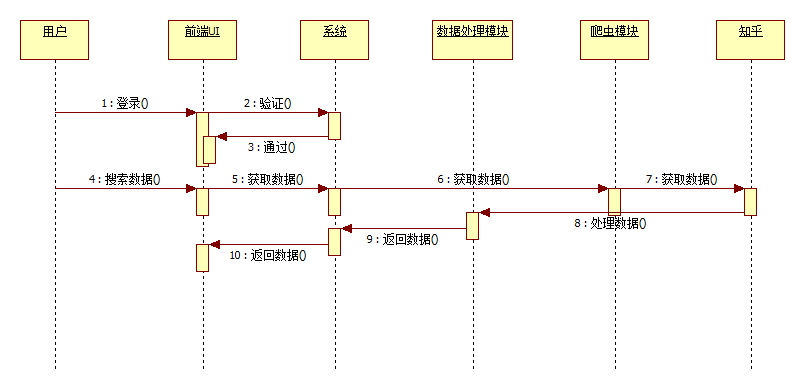
数据库 lucence/mysql

## 软件系统的功能描述

1. 对百万级数据进行爬取，同时对获取到的数据进行清理，分析，统计以及可视化，同时进行可能做到短周期内同步更新，确保数据的真实性以及有效性。
2. 对知乎上的大v基于粉丝数量进行排序（由于知乎本身已经有了热门话题的排序功能，所以这里暂且不做）
3. 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）



***1、用户登录搜索***



|  |  |
| --- | --- |
| 用例名 | 用户登录搜索 |
| 用例描述 | 用户登录后进行搜索信息 |
| 参与者 | 用户 知乎 |
| 过程 | 1. 用户登录 2. 系统验证登录 3. 用户搜索信息 4. 系统请求信息 5. 系统获取信息 6. 系统处理信息 7. 系统反馈信息 |

## 可行性及潜在风险

技术上运用的都是已知且较为成熟的技术，需要处理和分析的数据量很大，规模大，需要的时间更多对算力和算法的要求更高。

可能存在的风险：搜集的信息中可能会包含隐私信息或者版权问题或者敏感信息，要有严格的审查机制。知乎对于信息的管控可能会导致业务难以进行。