世纪佳缘用户画像

All by LucasX（微信&知乎）

# 前言

为了了解我国18~28岁年轻人群体在婚恋方面的行为偏好，同时获取海量图片用于深度学习神经网络的训练。利用网络爬虫技术自动地从我国最大的婚恋交友网站——世纪佳缘采集用户已发布的择偶信息，并存入MySQL数据库以便做进一步的分析。声明：本项目所有数据均来自“世纪佳缘”网站用户已发布的择偶数据，仅仅作为学习交流使用，本项目所有文档、源代码、数据原则上对所有用户开源，任何人不得利用其进行违反法律或道德之用途！如若有侵犯您的权益，请及时告知本人予以删除，谢谢合作！

# 技术架构

## 数据采集

Python在网络爬虫领域具有非常广泛的应用，且代码清晰易读，第三方支持模块众多。因此本项目数据采集部分使用Python3.5.2作为开发平台，网络请求使用requests库，HTML页面使用BeautifulSoup4进行解析，并实时将爬取的每一页JSON数据先行存储到本地磁盘。

## 数据存储

MySQL是一种免费、开源的关系型数据库，在综合考虑数据的数据量大小、可移植性、存储成本后，本项目使用MySQL来存储从网站爬取的数据。

## 数据清洗

Java在Web开发中有着最为广泛的应用，MyBatis是一种非常成熟的数据持久层框架，通过该框架，我们可以直接通过配置XML来操作数据库，而不需要自己编写代码进行驱动的加载、数据库连接池的打开与关闭等操作。因此本项目的数据清洗入库使用Java语言。

## 数据分析

在经由数据清洗过程之后，此时得到的已经是较为干净的结构化数据了。由于R语言在统计分析的广泛应用。因此本项目数据统计分析采用R语言来作为支撑。

## 数据可视化

在综合考虑平台一致性、数据可视化领域的应用案例，本项目采用R语言ggplot2包实现数据可视化。

## 系统架构图

**互联网**

**JSON文件**

**JSON文件**

**JSON文件**

**数据清洗**

**MySQL**

**R统计分析**

**ggplot2可视化**

# 实验过程

# 结果分析

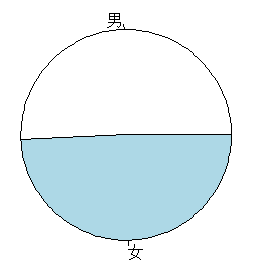
在对世纪佳缘网站718395名用户信息进行抓取后，利用R语言进行分析，得到如下的用户画像，其中数据可视化部分作者会介绍R语言中的ggplot2包。

> install.packages("ggplot2")

> library("ggplot2")

## 男女用户人数比例（R语言实现）

> pie(c(male\_number, female\_number), labels = c("男", "女"))



可以看到，世纪佳缘男女用户比例基本接近1：1（363839：354556 | 50.65%：49.35%）。

## 主要指标体系分析（SQL实现）

> write.csv(data\_female[, c(1, 3, 4, 6, 7, 8, 9, 13, 14)], 'e:/data\_female.csv')

### 地区省份分析

SELECT COUNT(uid) AS '数量', province AS '省份' FROM candidates GROUP BY province ORDER BY COUNT(uid) DESC

### 婚姻状态分析

SELECT COUNT(uid) AS '数量', marry AS '婚姻状态' FROM candidates GROUP BY marry

### 地区平均年龄分析

SELECT province AS '省份', AVG(age) AS '平均年龄' FROM candidates WHERE sex = '男' GROUP BY province ORDER BY AVG(age) DESC

### 教育程度分析

SELECT COUNT(uid) AS '数量', education as '文化程度' FROM candidates GROUP BY education ORDER BY COUNT(uid) DESC

### 地区身高分析

SELECT province AS '省份', AVG(height) as '身高' FROM candidates WHERE sex = '男' GROUP BY province ORDER BY AVG(height) DESC

SELECT province AS '省份', AVG(height) as '身高' FROM candidates WHERE sex = '女' GROUP BY province ORDER BY AVG(height) DESC

## 分词

### 择偶标准

> write.csv(data\_female\_hb[, 10], 'd:/matchcondition-hb.txt', quote = F, row.names = F)