

**数据可视化试卷**

姓 名: 董坤

学 号: 201814915

学 院: 软件学院

专 业: 软件工程

山东大学

2019年6月29日

**目 录**

[1. 目的 3](#_Toc12457336)

[2. 系统概述 3](#_Toc12457337)

[2.1 面向用户 3](#_Toc12457338)

[2.2 系统功能 3](#_Toc12457339)

[3. 系统需求分析 4](#_Toc12457340)

[3.1 微博爬虫 4](#_Toc12457341)

[3.1.1 微博热门图片 4](#_Toc12457342)

[3.1.2 生成微博词云 5](#_Toc12457343)

[3.1.3 微博热度分析 6](#_Toc12457344)

[3.1.4 活跃时间分析 6](#_Toc12457345)

[3.2 粉丝爬虫 7](#_Toc12457346)

[3.2.1 粉丝性别分析 7](#_Toc12457347)

[3.2.2 粉丝年龄分析 7](#_Toc12457348)

[3.2.3 粉丝地区分析 8](#_Toc12457349)

[4. 系统设计 8](#_Toc12457350)

[4.1 首页交互设计 8](#_Toc12457351)

[4.2 微博爬虫 9](#_Toc12457352)

[4.2.1 微博热门图片 9](#_Toc12457353)

[4.2.2 生成微博词云 9](#_Toc12457354)

[4.2.3 微博热度分析 10](#_Toc12457355)

[4.2.4 活跃时间分析 10](#_Toc12457356)

[4.3 粉丝爬虫 11](#_Toc12457357)

[4.3.1 粉丝性别分析 11](#_Toc12457358)

[4.3.2 粉丝年龄分析 11](#_Toc12457359)

[4.3.3 粉丝地区分析 12](#_Toc12457360)

[5. 系统实现 12](#_Toc12457361)

[5.1 微博爬虫 12](#_Toc12457362)

[5.1.1 微博热门图片 12](#_Toc12457363)

[5.1.2 生成微博词云 13](#_Toc12457364)

[5.1.3 微博热度分析 13](#_Toc12457365)

[5.1.4 活跃时间分析 13](#_Toc12457366)

[5.2 粉丝爬虫 13](#_Toc12457367)

[5.2.1 粉丝性别分析 14](#_Toc12457368)

[5.2.2 粉丝年龄分析 14](#_Toc12457369)

[5.2.3 粉丝地区分析 15](#_Toc12457370)

[6. 数据分析 15](#_Toc12457371)

[参考文献 17](#_Toc12457372)

# 1. 目的

随着网络的快速发展和普及，微博作为一个信息分享、获取以及传播的平台，以其快速传播、实时搜索等特点，已成为当代人们抒发情感，分享信息和获取资源必不可少的工具[1]。对于每个用户来说，微博中记录了他们的日常生活等各种信息，同时通过微博的关注和被关注，人们的关系可以通过微博联系起来[2]。

系统的设计目的是通过微博爬虫根据微博用户ID获取微博信息，获取微博中热度最高的图片，从而分析发表什么样的图片风格更受粉丝喜欢；获取微博词云，从而分析博中最容易出现的词语和人名，分析到与该用户关系密切的人；获取每条微博的点赞、评论和转发数目作为微博互动数目，分析哪一段时间或发表什么内容微博粉丝更加容易进行互动；获取每条微博的发表时间，从而分析得到该用户经常的微博活跃时间。

通过微博爬虫获取粉丝信息，获取粉丝的性别信息，从而分析粉丝的男女比例；获取粉丝的年龄信息，从而分析粉丝的年龄分布；获取粉丝的地区信息，从而分析粉丝的地区分布及分布数目[3]。

分析后的数据有很大的作用[4]，首先，根据粉丝的分布情况可以选择开演唱会最合适的地区，然后根据粉丝年龄段分析来适当调整门票价格。其次，根据微博的互动数目和热度可以了解粉丝的偏好风格，可以多发这种风格的微博来达到涨粉的目的。最后，粉丝可以根据关注明星的微博数据分析对其有更深入的了解。

# 2. 系统概述

## 2.1 面向用户

系统的设计面向任何想要进行微博爬虫及数据分析的用户，根据某个微博用户的ID便可进行爬虫，然后对该用户的微博和粉丝进行爬取并进行数据分析[5]。不仅可以用来对自己的微博数据进行分析从而观察微博的数据和粉丝的属性，还可以分析偶像的微博数据来增加更深的了解[6]。系统展示使用html进行交互，操作简单方便，显示清晰简洁，对于不是计算机专业的同学也很容易操作[7]。

## 2.2 系统功能

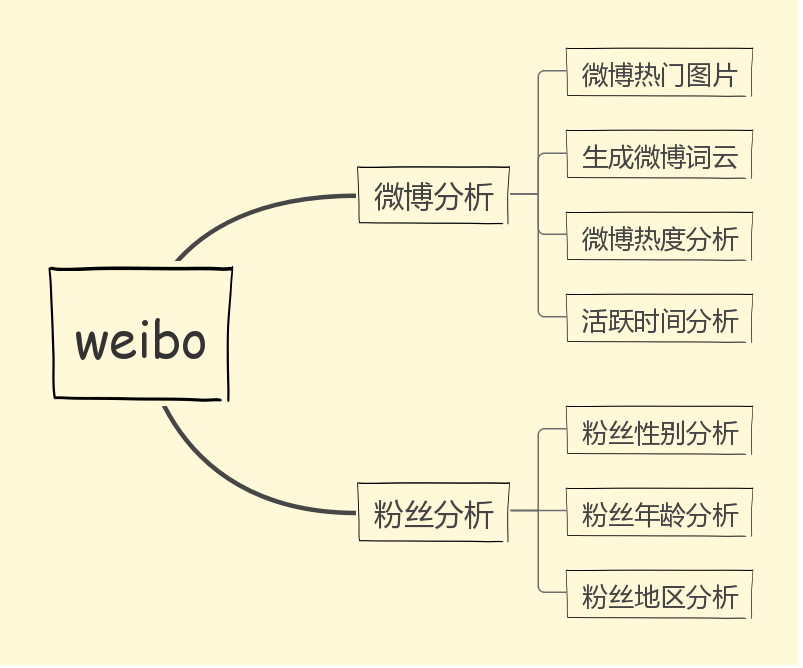


图2-1 系统功能模块

系统主要有七个功能[8]：

1. 微博热门图片：根据微博ID爬取当前用户最热门的微博图片并进行展示。

2. 生成微博词云：根据微博ID爬取该用户的所有微博内容来生成词云进行展示。

3. 微博热度分析：根据微博ID爬取该用户的所有微博的点赞、评论和转发数目作为微博的热度并使用折线图的形式进行展示。

4. 活跃时间分析：根据微博ID爬取该用户发布微博的时间，进行统计分析进行展示。

5. 粉丝性别分析：爬取该用户的粉丝信息，统计性别信息进行分析和展示。

6. 粉丝年龄分析：爬取该用户的粉丝信息，统计年龄信息进行分析和展示。

7. 粉丝地区分析：爬取该用户的粉丝信息，统计地区信息进行分析和展示。[9]

# 3. 系统需求分析

## 3.1 微博爬虫

根据给定的微博用户ID，使用爬虫获取热门微博的微博图片，以及所有微博的信息，包括微博ID、微博正文、发布位置、发布时间、发布工具、点赞数、转发数、评论数。但由于发布位置与发布工具信息较少，所以不具有统计性[10]。

### 3.1.1 微博热门图片

根据获得的热门微博的微博图片，一个热门微博可能有多张图片，在每个微博中选取一张图片，一共获取18个微博图片，并进行展示，从而分析该ID发布的最受粉丝欢迎的图片风格。当进入该功能模块时，可以查看所有微博热门图片，点击图片可以查看大图。

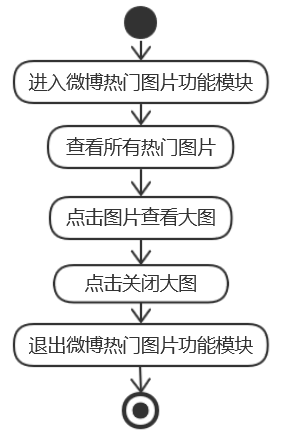


图3-1 微博热门图片功能模块

### 3.1.2 生成微博词云

根据获取的所有微博的微博正文，生成微博词云，从而分析该微博ID发布最多的词语和人名，推测出与该微博ID关系较为密切的人以及该微博ID的语言习惯。当进入该功能模块时，可以查看在该微博ID中出现最多的词语，词语的频次越多，字体越大，鼠标划过时可以查看该词语的出现次数，点击按钮可以重新生成微博词云。

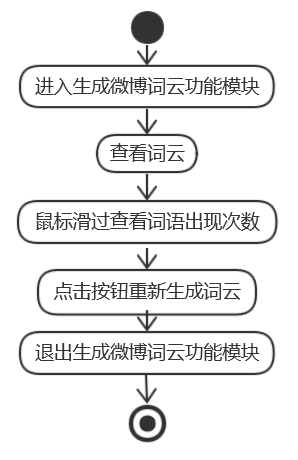


图3-2 生成微博词云功能模块

### 3.1.3 微博热度分析

根据获取的所有微博的点赞数、评论数、转发数作为微博的热度，使用折线图来表示每条微博的三个数据，从而分析粉丝互动数目比较多的微博类型。当进入该功能模块时，可以查看由所有微博的点赞数、评论数、转发数组成的折线图，鼠标滑过时可以查看某个微博的三个详细数据。

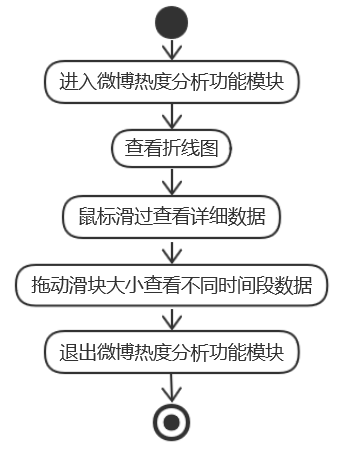


图3-3 微博热度分析功能模块

### 3.1.4 活跃时间分析

根据获取的所有微博的发布时间，以一小时为一个时间段，计算出每个时间段发布微博的数量，形成雷达图，每个时间段的数值表示该时间段内发布微博的数量，从而分析该微博ID最经常的微博活跃时间，大致推测其作息时间。当进入该功能模块时，可以查看每个时间段微博发布的数目组成的雷达图，鼠标滑过时可以查看每个时间段微博发布的数目。

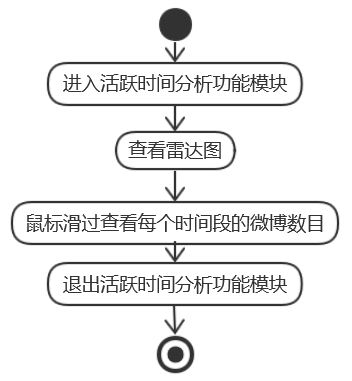


图3-4 活跃时间分析功能模块

## 3.2 粉丝爬虫

根据给定的微博用户ID，使用爬虫获取该微博ID粉丝的部分信息，包括昵称，性别，年龄，地区，简介，关注方式等信息。为获取高质量的统计数据，我们只使用性别、年龄、地区信息进行分析。

### 3.2.1 粉丝性别分析

根据获取的所有粉丝的性别信息，计算得粉丝中男生和女生数目和比例，形成饼状图，每个部分大小表示该性别所占比例，从而分析出该微博ID粉丝的性别比例。当进入该模块时，可以查看男女粉丝所占比例，鼠标滑过时可以查看每个部分的人数及比例。

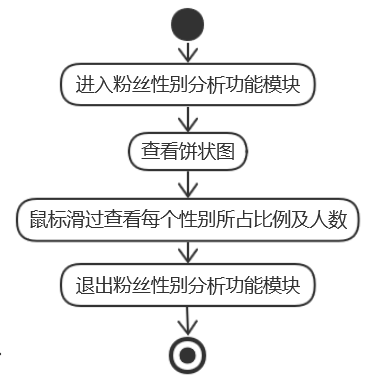


图3-5 粉丝性别分析功能模块

### 3.2.2 粉丝年龄分析

根据获取的所有粉丝的年龄信息，计算得粉丝中各年龄段数目，形成条形图，每个部分大小表示该年龄段粉丝人数多少，从而分析出该微博ID粉丝的主要年龄区间。当进入该模块时，可以查看不同年龄段粉丝人数，鼠标滑过时可以查看每个部分的人数。

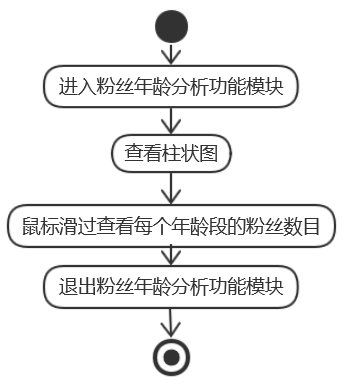


图3-6 粉丝年龄分析功能模块

### 3.2.3 粉丝地区分析

根据获取的所有粉丝的地区信息，计算得各个省份粉丝数目，在地图上每个省份标注每个省份人数比例，以圆形大小表示人数多少，其中top5进行特殊标记，并计算每个省份粉丝的平均年龄，从而可以分析得到该微博ID粉丝的地区分布，进而选取合适的演唱会地点及门票价格。当进入该模块时，可以查看不同地区粉丝人数，鼠标滑过时可以查看每个省份的粉丝数目和粉丝平均年龄。

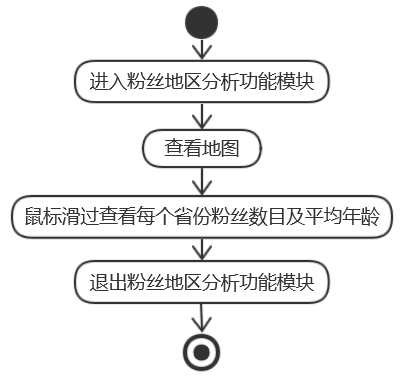


图3-7 粉丝地区分析功能模块

# 4. 系统设计

## 4.1 首页交互设计

根据七个主要功能模块，设计七行文字分别链接到每一部分的功能模块。左下角加入系统名称及数据来源。



图4-1 首页设计

## 4.2 微博爬虫

### 4.2.1 微博热门图片

该页面显示获得的热门微博所包含的图片，因图片大小原因显示缩略图，点击每个图片可以查看大图。



图4-2 微博热门图片功能页面设计

### 4.2.2 生成微博词云

该页面显示根据获取的微博正文信息生成的词云，可以查看在该微博ID中出现最多的词语，词语的频次越多，字体越大，鼠标划过时可以查看该词语的出现次数，点击按钮可以重新生成微博词云。



图4-3 生成微博词云功能页面设计

### 4.2.3 微博热度分析

该页面显示可以查看根据获取的所有微博的点赞数、评论数、转发数组成的折线图，鼠标滑过时可以查看某个微博的三个详细数据,拖动下方滑块可以查看不同时间段的详细信息。

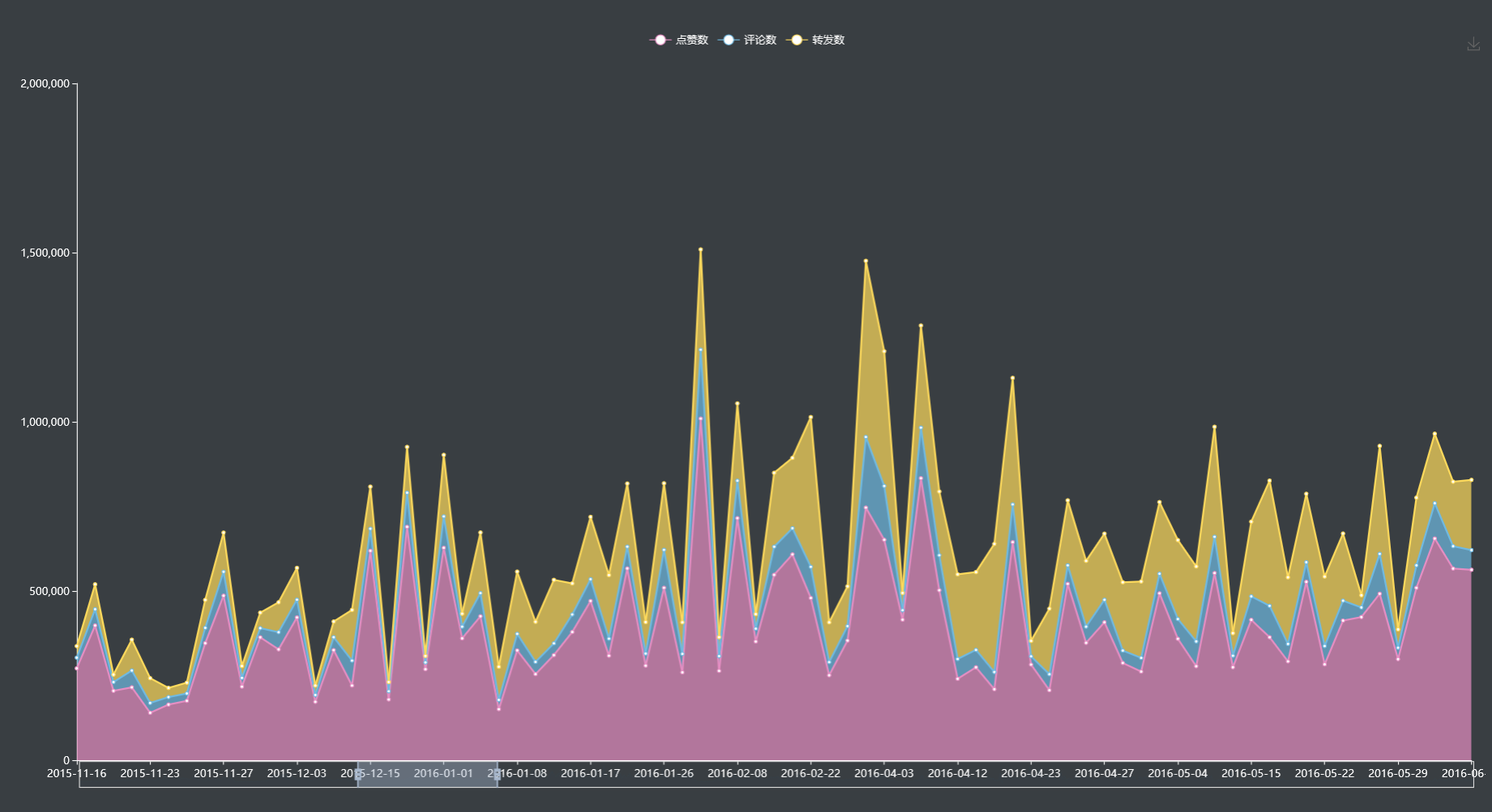


图4-4 微博热度分析功能页面设计

### 4.2.4 活跃时间分析

该页面显示根据获取的所有微博的发布时间，以一小时为一个时间段，计算出每个时间段发布微博的数量，形成的雷达图。可以查看每个时间段微博发布的数目组成的雷达图，鼠标滑过时可以查看每个时间段微博发布的数目。

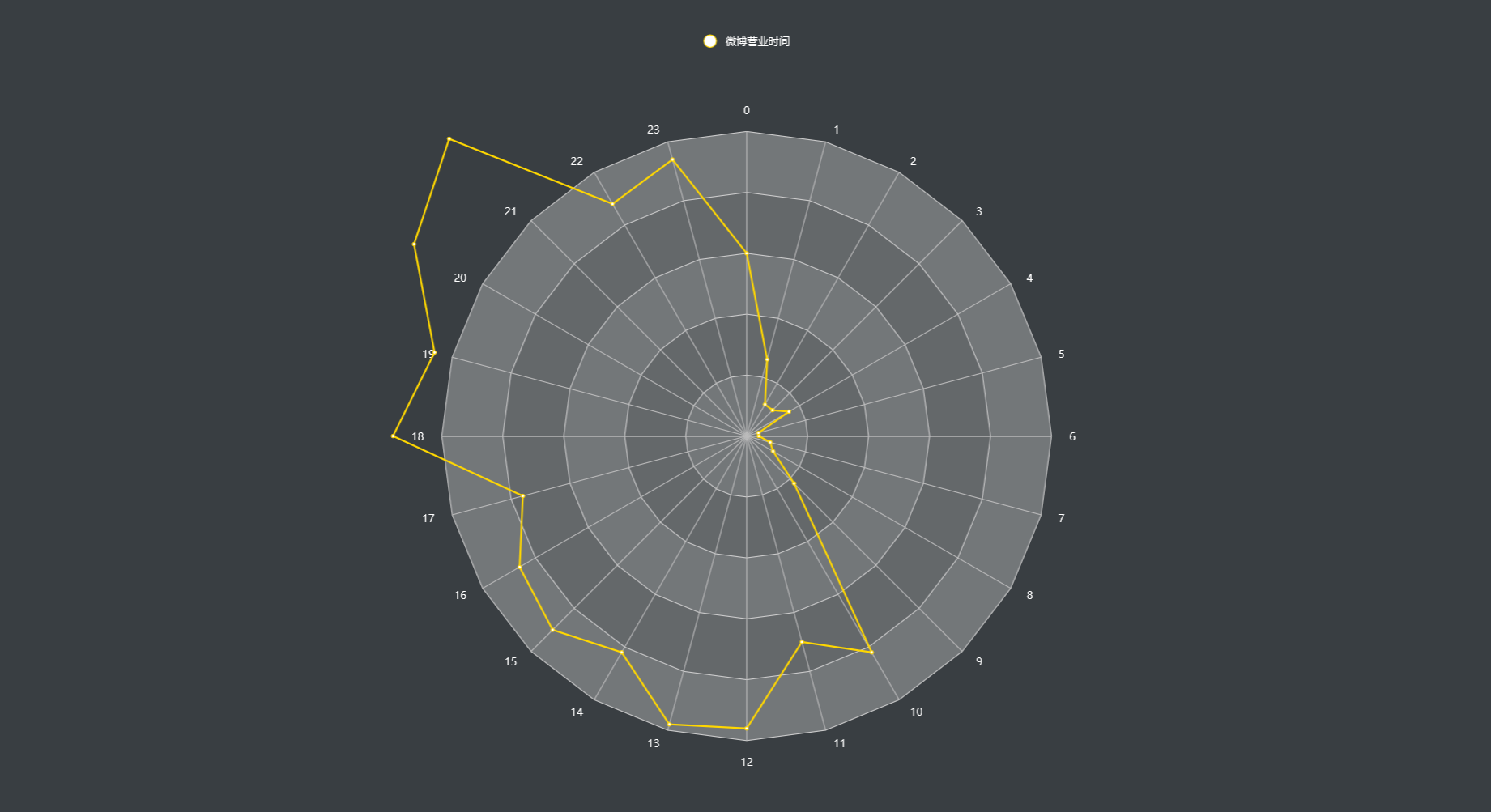


图4-5活跃时间分析功能页面设计

## 4.3 粉丝爬虫

### 4.3.1 粉丝性别分析

该页面显示根据获取的所有粉丝的性别信息，计算得粉丝中男生和女生数目和比例，形成的饼状图，其中每个部分大小表示该性别所占比例。可以查看男女粉丝所占比例，鼠标滑过时可以查看每个部分的人数及比例。

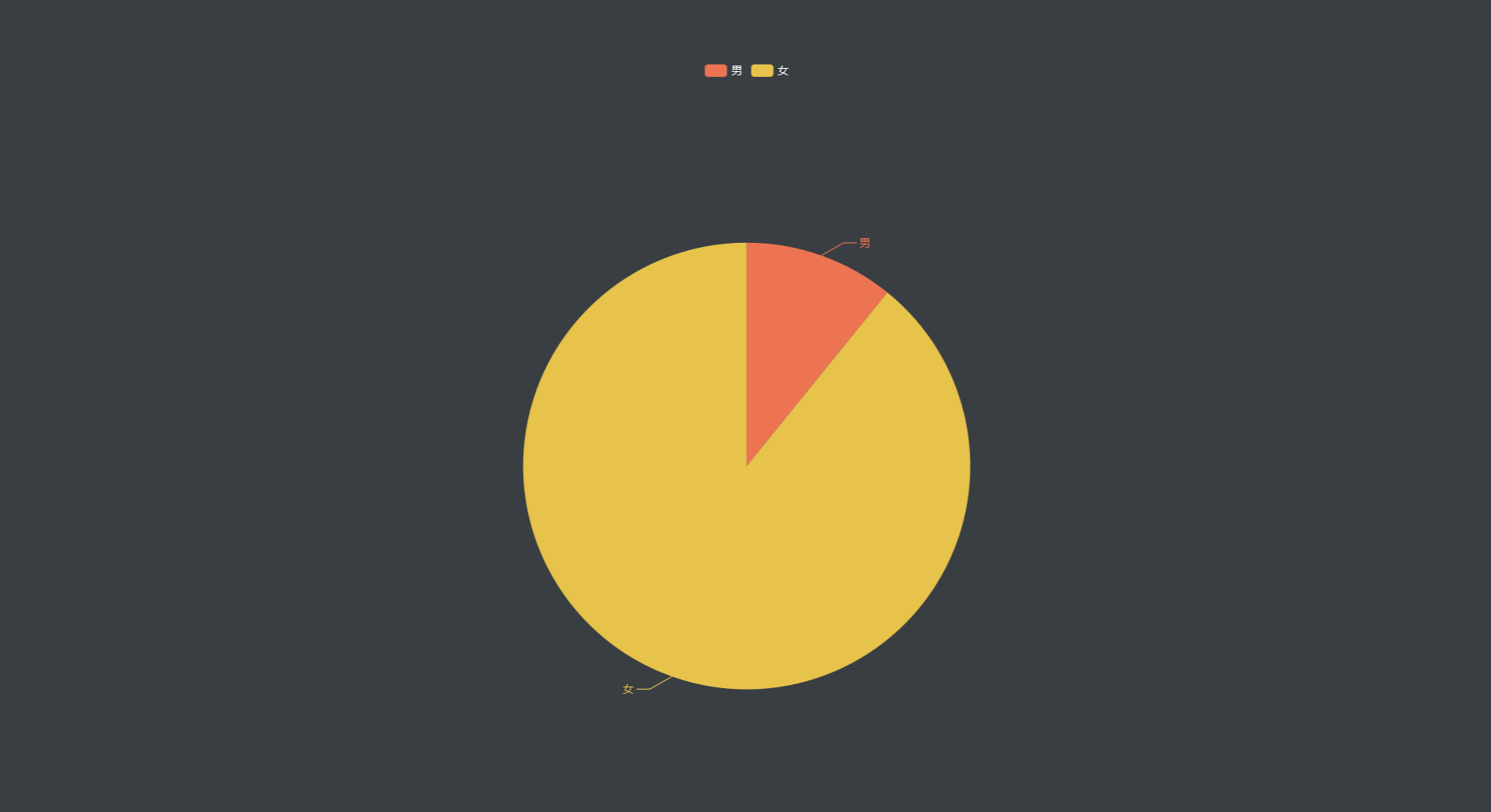


图4-6 粉丝性别分析功能页面设计

### 4.3.2 粉丝年龄分析

该页面显示根据获取的所有粉丝的年龄信息，计算得粉丝中各年龄段数目形成的条形图，其中每个部分大小表示该年龄段粉丝人数多少。可以查看不同年龄段粉丝人数，鼠标滑过时可以查看每个部分的人数。

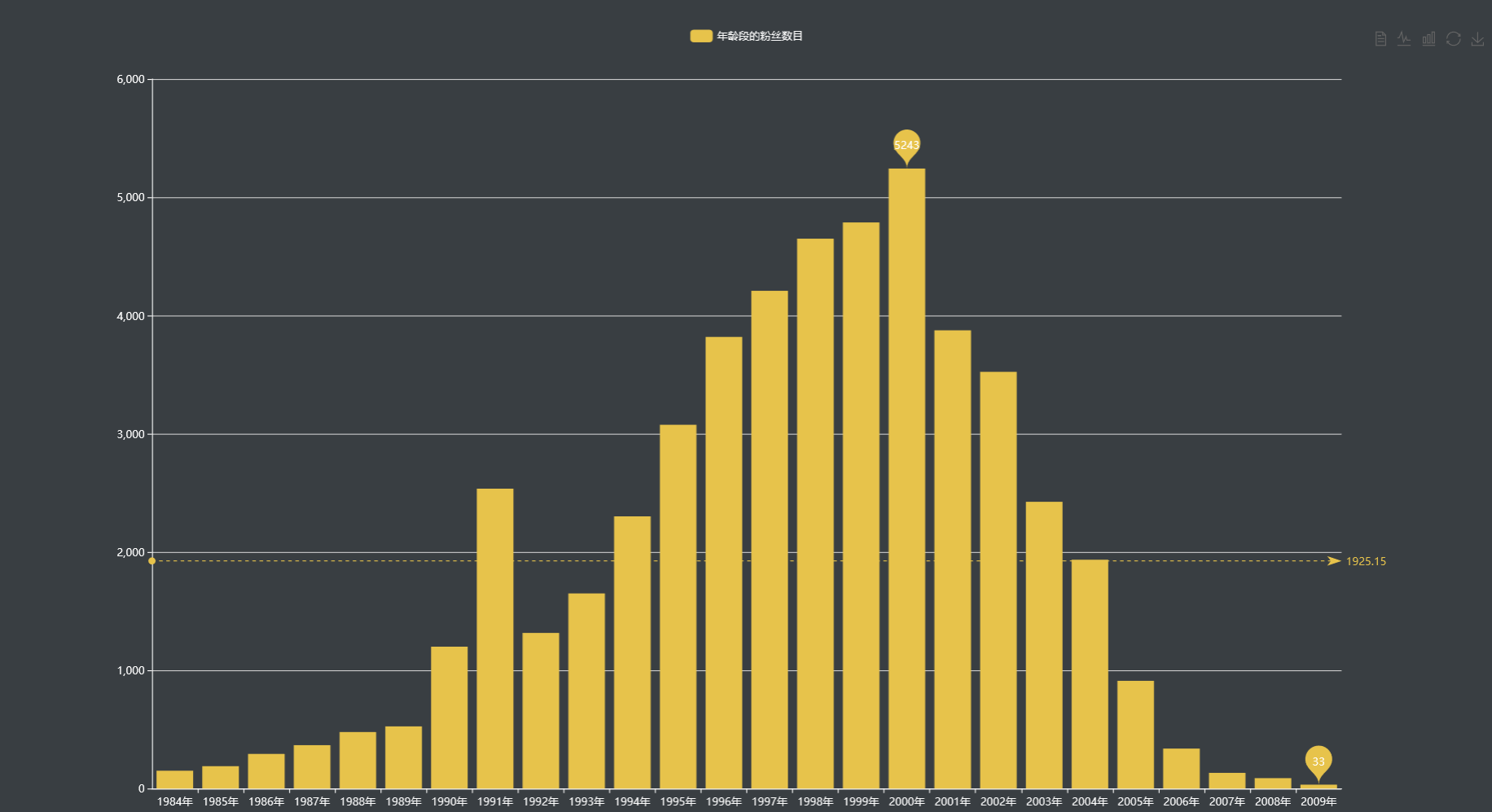


图4-7 粉丝年龄分析功能页面设计

### 4.3.3 粉丝地区分析

该页面显示根据获取的所有粉丝的地区信息，计算的各个省份粉丝数目，其中在地图上每个省份标注每个省份人数比例，以圆形大小表示人数多少，top5进行特殊标记，并显示每个省份粉丝的平均年龄。可以查看不同地区粉丝人数，鼠标滑过时可以查看每个省份的粉丝数目和粉丝平均年龄。



图4-7 粉丝地区分析功能页面设计

# 5. 系统实现

在该系统的实现过程中，以张艺兴微博（@努力努力再努力x）为例进行微博和粉丝爬虫，并进行数据分析。

## 5.1 微博爬虫

根据给定的微博ID（2706896955），使用爬虫获取前18个热门微博的图片，以及所有微博的信息，总共762条（截止到2019年6月4日），其中每条微博信息包括微博ID、微博正文、发布位置、发布时间、发布工具、点赞数、转发数、评论数。

### 5.1.1 微博热门图片

1.根据获取的前18个热门微博的图片，一条微博有多个图片时选取其中一张图片，从而获取热门微博的18张图片。

2.数据处理：对图片进行裁剪批量处理，得到18张500\*500的标准大小图片。

3.绘制图表：编写html代码，显示该18张图片，因图片数目较多，分两部分进行展示，首先展示前9张图片，然后点击load加载剩余9张图片。

### 5.1.2 生成微博词云

1.根据获取的微博信息，将微博正文单独保存在txt文件中。

2.数据处理：对文本进行分词处理，并计算每个词语出现的频次。

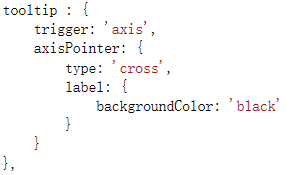
3.设置限定条件，去除系统自带词语，比如“微博”、“转发”、“原文”等词语，设置用户词典，比如“极限挑战”、“青春制作人”等关键词，防止在分词过程中进行拆分。

4.绘制图表：根据由大到小的词语出现频次，选取前300个词语生成词云并进行显示。

### 5.1.3 微博热度分析

1.绘制图表：以获取的微博的发布时间作为横坐标，x轴上分别表示762个微博发布时间，以该微博的点赞数、转发数、评论数作为纵坐标绘制折线图。

2.鼠标滑过时可以显示该条微博的发布时间，以及该微博的点赞数、转发数、评论数。



3.点击上方图标，可以选择是否进行显示，从而可以单独查看某一项的变化趋势。

4.拖动下方滑块,可以查看不同时间段的详细信息。

### 5.1.4 活跃时间分析

1.数据处理：根据获取的微博的发布时间，以每一个小时作为一个时间段，统计每个时间段内微博发布的数目。

2.绘制图表：根据24个时间段以及每个时间段的微博数目，绘制雷达图。

3.鼠标滑过时可以显示每个时间段的微博数目。

## 5.2 粉丝爬虫

由于微博的反爬虫机制，程序不能够直接对粉丝页面进行爬取，最多只能爬五十个左右的粉丝，这对我们的操作十分不利。于是选择爬取每条微博中进行评论的粉丝ID，这样有许多好处，相比于点赞和转发，认真进行评论的粉丝更能够作为粉丝代表，同时，也去除了粉丝列表中的众多僵尸粉。但是这样也存在某些问题，比如大量数据重复，需要后期进行去除重复粉丝ID的操作。

对于每条微博，我们爬取热门评论的前100页也就是前1000个评论的粉丝ID，这样不仅能够减少爬虫的数据量，也能够筛选大部分粉丝，从而获取最能够代表粉丝群体的一部分。由于微博反爬虫机制，一旦爬取信息速度过快，会被微博视为机器操作，所以在爬虫过程中，需要设置暂停时间，每爬取一段信息就暂停几秒钟的时间，才能够顺利爬取。同时，这也使得爬虫速度十分慢，加上数据量大，总共需要花费几天左右的时间才将所有微博评论中的粉丝ID获取下来。

1.根据每条微博的ID获取微博评论下的粉丝ID。

2.对所获得的所有ID去除其中的重复值。

3.根据粉丝ID爬取粉丝信息，包括粉丝昵称、粉丝性别、粉丝年龄、粉丝地区等信息。

4.根据这些信息进行数据分析。

理论上讲，在762条微博中爬取粉丝ID，平均每条微博可以爬取1000个左右的粉丝ID，但由于粉丝ID大量重复性，加上有些各种原因导致的无法爬取的粉丝，所以实际上通过粉丝爬虫过程，共爬取609196个粉丝ID，其中重复的粉丝ID有475593个，删除重复值后剩余133603个粉丝ID，然后根据133603个粉丝ID爬取其微博信息。

排除各种原因（微博限制，爬虫速度）导致无法获取的粉丝微博信息，实际上共爬取粉丝信息90924条，接下来对其进行数据分析。

### 5.2.1 粉丝性别分析

1.数据处理：根据获取的90924个粉丝的性别信息，分别计算得男生人数为9848和女生人数为81076，然后分别计算两部分所占的比例。

2.绘制图表：根据计算出的男生和女生人数以及所占比例，绘制饼状图。

3.鼠标滑过时可以显示每个性别的人数及比例。

### 5.2.2 粉丝年龄分析

1.数据处理：根据获取的90924个粉丝的年龄信息，将每一年作为一个年龄段，分别计算每个年龄段的人数。

首先排除虚假年龄信息，一个是2010年以后的用户年龄，一个是20世纪初期的年龄，显然这些用户不可能使用微博。然后从1970年开始计算，直到2009年，计算每一年的粉丝数量。最后进行展示，由于1984年之前的人数较少所以不做展示，从1984年开始进行展示。

2.绘制图表：根据每个年龄段的人数，绘制饼状图。

3.鼠标滑过时可以显示每个年龄段的人数。

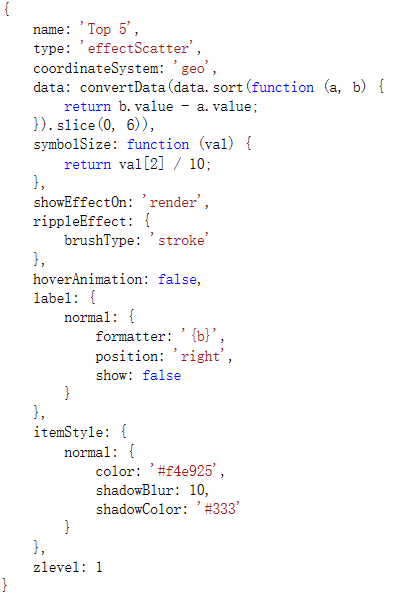
### 5.2.3 粉丝地区分析

1.数据处理：根据获取的90924个粉丝的地区信息，按每个省份进行分类，分别计算每个省份的人数，并计算出该省份的平均年龄值。

同样对每个省份的粉丝年龄做如上的筛选，除去明显不符合实际的信息，计算从1970年至2009年的粉丝年龄平均值进行展示。

2.绘制图表：根据每个省份的人数，绘制饼状图。

3.用不同大小的圆形表示不同数量粉丝的省份，其中top进行特殊标记，鼠标滑过时可以显示每个省份的人数及该省份粉丝的平均年龄。



# 6. 数据分析

1.根据微博热门图片展示结果，可以分析出发布活动出席或拍摄的高清图片、分享的日常生活图片，以及演唱会图片，更容易作为热门微博，而发布的广告信息或品牌宣传则热度不高。

2.根据生成的微博词云展示结果，分析出张艺兴微博中出现最多的词语有努力、大家、谢谢、音乐、希望、加油，可以了解到张艺兴是一个非常努力、热爱音乐并且充满正能量的人，另外出现比较多的明星名字还有孙红雷、黄渤、罗志祥、黄磊，可以了解到张艺兴与这几位明星关系更加密切，微博互动也相对较多。

3.根据微博热度分析展示结果，可以看到自从2013年开通微博以来，到2019年，张艺兴的微博粉丝互动是一直呈上升趋势，尤其是2018年以后，基本上大概率达到峰值。

对于点赞数，达到峰值的微博基本上是生日微博、登上春晚微博、以及发布坚定立场言论的类似微博，另外新剧播出，新专辑发布，发自拍的类似微博也能收获很多的赞。

对于评论数，达到峰值的微博基本上是生病期间、生日微博，另外新音乐发布、综艺节目结束、发自拍的类似微博也有很多评论。

4.根据活跃时间分析展示结果，可以看到微博的发布时间一般在10点到凌晨1点之间，分析出张艺兴经常晚睡晚起，也有少量微博是发布在凌晨，说明在准备音乐阶段经常通宵工作，是个很认真努力工作的人。

5.根据粉丝性别分析展示结果，可以看到张艺兴的粉丝中女生大约占89%，男生大约占11%。

6.根据粉丝年龄分析展示结果，可以看到张艺兴的粉丝中大部分属于95后，出生于2000年左右，这个年龄段的粉丝基本上都在上大学，所以比较有时间追星。另外，可以明显看到1991年的粉丝数量也是一个小高峰，那是因为张艺兴出生于1991年，所以很多粉丝都将出生日期改为了1991年10月7日所导致的。

7.根据粉丝地区分析展示结果，首先每个地区的粉丝年龄差距不是很大，基本上都在22岁左右，所以不具有很大的参考价值。然后，观察地区分布，粉丝分布的前五名分别是广东省，海外，北京，湖南，山东，分别占粉丝总数的8%，6%，5.3%，5.3%，4.7%。其中湖南省粉丝较多是因为张艺兴家乡是湖南，所以很多粉丝将地区改为了湖南长沙，但是也不能排除真正湖南的粉丝很多的情况。另外，江苏，四川，河南，浙江等地粉丝分布也比较多，如果开演唱会可以考虑在以上地区选址。

# 参考文献

[1] 罗一纾. 微博爬虫的相关技术研究[D]. 哈尔滨工业大学.

[2] 吴剑兰. 基于Python的新浪微博爬虫研究[J]. 无线互联科技, 2015(6):93-94.

[3] 卢杨, 李华康, 孙国梓. 一种基于P2P技术的分布式微博爬虫系统[J]. 江苏大学学报：自然科学版, 2016(3):296-301.

[4] 王艳阁. 面向微博爬虫系统的分析[J]. 河南科技, 2013(2X):16-17.

[5] 陈政伊, 袁云静, 贺月锦, et al. 基于Python的微博爬虫系统研究[J]. 大众科技, 2017(8).

[6] 叶婷. 基于关键词的微博爬虫系统的设计与实现[D]. 2016.

[7] 于营 . 面向微博的网络爬虫数据采集[J]. 信息系统工程, 2017.

[8] 刘晶晶. 面向微博的网络爬虫研究与实现[D]. 复旦大学, 2012.

[9] 范巍. 微博爬虫研究与实现[J]. 硅谷, 2013(22):54-54.

[10] 陈琳, 任芳. 基于Python的新浪微博数据爬虫程序设计[J]. 信息系统工程, 2016(9):97-99.