
北京航空航天大學
BEIHANG UNIVERSITY

# 基于 PageRank 的中国机场和航空公司评价

高星宇　姜亦申　卢统凯

李振东　吴金波　许昊

# 摘要

本次实验中，我们通过爬虫采集了携程网 5 月 30 日的全国 204 各机场的航班数据并采用 TextRank[1] 算法得出全国 204 个城市机场基于城市和乘客信息的排名。此外我们还综合考虑了全国航空公司的准点率，票价等信息，对全国的航空公司进行了评价。

**关键词：** Pagerank　机场评价　航空公司评价

# 目录

# 1 简介

## 1.1 研究背景

当前针对机场排名有多种评价方式，也建立了较为完善的评价体系，不同的评价方式的关注点也迥然不同。在本次实验中，我们将机场以及航班线路看成网络结构从而引入 PageRank 算法进行机场排名。PageRank 算法 [2] 是 Google CEO 拉里佩奇提出的一种基于图形的排序算法，用来对通过关键词搜索到的网站进行排名。PageRank 算法是一种通过考虑从整个图表中递归计算的全局信息来决定图形中顶点的重要性的方法，而不是仅仅依赖于本地顶点特定的信息，这样可以综合性的考虑所有对象的特点并获得较好的排序，在引文分析，社交网络和万维网链接结构的分析等领域取得了极好的成果。然而在互联网中，一个页面包含到另一个页面的多个或部分链接是不常见的，因此基于图形的排名的原始 PageRank 定义假设未加权的图形，而这正是我们在进行机场排名中所遇到的问题，因此我们采用 PageRank 的改进算法 TextRank[2]。TextRank 算法同样是一种用于文本的基于图的排序算法, 其基本思想来源于谷歌的 PageRank 算法, 通过把整体分割成若干组成单元并建立图模型, 利用投票机制对个点进行重要性排序。

。

# 2 实验过程

## 2.1 相关配置及软件版本

数据库：Mysql 8.0.16
爬虫：Python 3.5 ; 核心库：pymysql,request, re, json, numpy, pandas etc
可视化核心库：pyecharts

## 2.2 获取数据

在本次实验中我们利用爬虫程序来获取携程网的相关数据，并将获取的数据用 Mysql 数据库进行保存。

### 2.2.1 爬虫介绍

网络爬虫指按照一定的规则（模拟人工登录网页的方式），自动抓取网络上的数据，把站点返回的 HTML 代码，JSON 数据或二进制数据（图片、视频）爬到本地，进而提取用户需要的数据，存放起来使用的程序。平时在我们查询资料获取数据时常常采用人工方式手动获取数据；但在遇到需要大批量从互联网中获取数据的情况时，我们不能人工的去收集数据，这样会很浪费时间与金钱。而爬虫可以批量、自动化的获取和处理数据，这就是本实验采用爬虫获取数据的原因。

### 2.2.2 爬虫过程

爬虫过程大致可以分为发起请求，获取响应内容，解析数据，保存数据。发送请求是指使用 http 库向目标站点发起请求，即发送一个 Request（Request：用户将自己的信息通过浏览器（socket client）发送给服务器（socket server））；如果服务器正常响应，会返回 Response，里面包含的就是该页面的内容，这就是获取相应内容过程，当然在此过程中由于某些原因并不会获得服务器的响应，此时成为异常情况，不同的异常情况会有相应的处理方法；服务器的返回信息需要进行解析处理以获得可使用的数据，解析数据的方法需要根据返回的类型而定，例如如果返回内容是 HTML，则可以用正则表达式、网页解析库进行解析，如果是 Json 类型，可以直接转换为 Json 对象解析；最后则是将解析后的信息进行保存，可以存储为文本，也可以保存至数据库，或其他特定类型文件，但由于数据量的庞大我们通常采用数据库保存数据，在本次实验中我们采用的是 Mysql 数据库进行数据保存。

### 2.2.3 具体实施

在本次实验中我们采用自己编写的 python 爬虫程序采集携程网站中每个机场的航班信息，为反制携程官网的“反爬”措施，我们采用了代理 ip 池的办法，每 8 分钟自动更换代理 ip; 我们使用 python 中的 json 库函数进行文件的解析，并通过相关数据的键值来索引所需数据；获得数据后我们使用 pymysql 库将获取的数据导入数据库。在数据的选择中，我们经过综合考量，最终选择了以下数据进行抓取：航空公司名称、航班号、起飞时间、降落时间、准点率、最低票价、飞机型号、飞机大小。

在数据获取的过程中，遇到了以下错误与异常：

1. 网络连接异常与代理 ip 异常。

2.json 文件解析异常。

3. 缺少数据异常。(当两个飞机场之间没有连接是会报错)

经过不断的修改与优化，我们采用了 sleep 的方式，跳过当前异常数据的处理，并采取的多次爬取的方式补全因为网络原因而未能获得的数据。与此同时，标记已经收集过的页面，从而大大优化了爬虫的获取数据速度。最终版本的爬虫可以实现全天 24 小时、不间断采集数据。

## 2.3 排名

PageRank 算法是由 Larry Page 等人于 1999 年提出的一种网页排序算法。该算法同时考虑了网页的流行性和权威性。也即，如果一个页面 P 被更多的页面引用，如 C1,C2,C3…，同时，当这些页面 C1,C2, C3…也都是被很多其他页面引用的优质网页的时候，那么网页 P 则是一个优质的网页。由于飞机航线网络和网页引用网络的结构是非常相似的，而 PageRank 算法作为一种链接关系排序算法，所以，PageRank 被移植到航线网络中用于机场和城市的排序。一般认为，飞机从一个城市到另一个城市，往往象征着社会资源的流动（例如劳动力）。如果一个城市 M 能接收到来自全国各地的航班，且如果这些抵达 M 的飞机也来源于经济程度很高的城市，那么显然 M 应该也是一个非常繁华的城市。

此外由于城市之间的航班关系相比网页之间的引用关系而言，更加现实且需要顾及经济利益，不会像网页引用那样随意，所以将 PageRank 应用于城市的经济程度排序是比较可行的。TextRank 算法是一种用于文本的基于图的排序算法。其基本思想来源于谷歌的 PageRank 算法, 通过把文本分割成若干组成单元 (单词、句子) 并建立图模型, 利用投票机制对文本中的重要成分进行排序, 仅利用单篇文档本身的信息即可实现关键词提取、文摘。和 LDA、HMM 等模型不同, TextRank 不需要事先对多篇文档进行学习训练, 因其简洁有效而得到广泛应用。在网页的上下文中，页面包含到另一个页面的多个链接是不常见的，因此基于图形的排名的原始 PageRank 定义的是未加权的图形。但是，在我们的模型中，图形是由机场航线构建的，所以可能包括机场之间的多个航班。因此，在模型中确定两个顶点之间连接的“强度”可能是有用的。因此，我们引入了基于航线“强度”排名的新公式，该公式在计算图中顶点的分数时考虑航线权重。图 1 绘制了相同样本图的

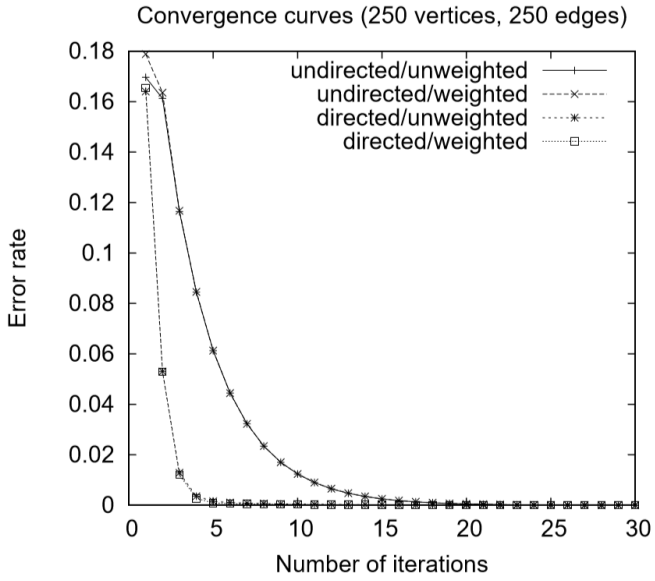


图 1: textrank

收敛曲线。虽然与未加权的替代方案相比，最终顶点得分（以及因此排名）显著不同，但对于加权和未加权图形，收敛的迭代次数和收敛曲线的形状几乎相同，这表明我们对于 TextRank 的采用是合理的。

在下文中，我们首先将分析并推导所使用的算法：以城市为投票标准的和以乘客为投票标准的两种 TextRank。接下来我们将会阐述我们的程序的架构：包括爬虫架构、算法架构、绘图架构以及整体架构。最后我们会通过图片和数据库展示我们的结果以及我们组内的分工。

# 3 算法分析和推导

在 PageRank 的原始推导中，作者采用的公式 [1] 为：

$$S(V_i) = \frac{1-d}{N} + d * \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j) \tag{1}$$

其中 Out(Vj) 为 j 城市的出度。该公式的实质是对每个联系城市分配影响力，而且分配给所有联系城市的比例是相同的。

我们的第一个算法改进了这一算法，即将每一个城市的影响力通过航班数带权重的分配给抵达城市。公式改进为：

$$WS(V_i) = \frac{1-d}{N} + d * \sum_{j \in In(V_i)} \frac{w_{ji}}{\sum\limits_{V_k \in Out(V_j)} w_{jk}} WS(V_j) \tag{2}$$

其中 wji 表示 j 城市抵达 i 城市的航班数目。更进一步地，我们在第二个算法中继续细化了权重，即将乘客数目作为决定权重的标准，此时公式中的 wji 表示 j 城市抵达 i 城市的乘客数目。我们假定每次航班乘客数目与本次航班的最大载客量的比例几乎为常数，所以在这一过程中，我们通过飞机型号决定的载客量来代替乘客数目。

但是囿于飞机型号的数据缺失，我们定义了一套补全数据的规则。即：首先，如果当天的有别的航班数据，我们优先补充时间最相近的型号。否则，我们定义第一次中的航线分数为：

$$score = WS(V_i) * \frac{n}{N(V_i)} \tag{3}$$

其中 WS(Vi) 是城市 Vi 的分数，n 表示该航线中的航班数，N(Vi) 表示以城市 Vi 为出发点的航班总数。接下来寻找第一次 PageRank 中分数最相近的航线并依照时间相近优先的原则进行数据补充即可。

## 3.1 算法流程展示

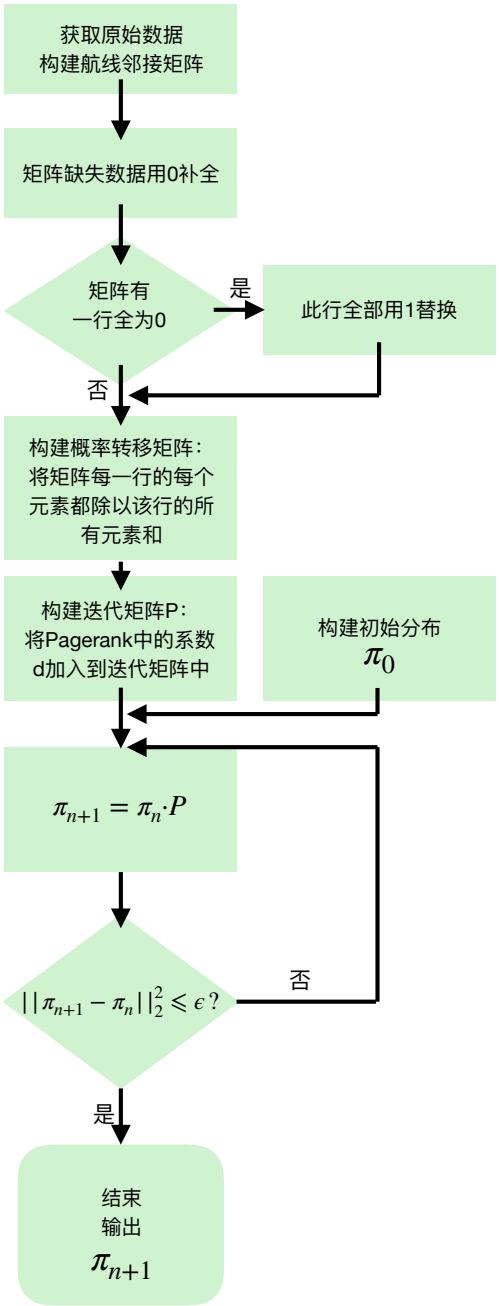


图 2: FlightRank 算法展示

# 4　结果展示

## 4.1　城市机场排名


图 3: 全国各大机场评分

上图为全国所有机场评分，从青色到红色表示分数递增。上图为全国较重要机场评

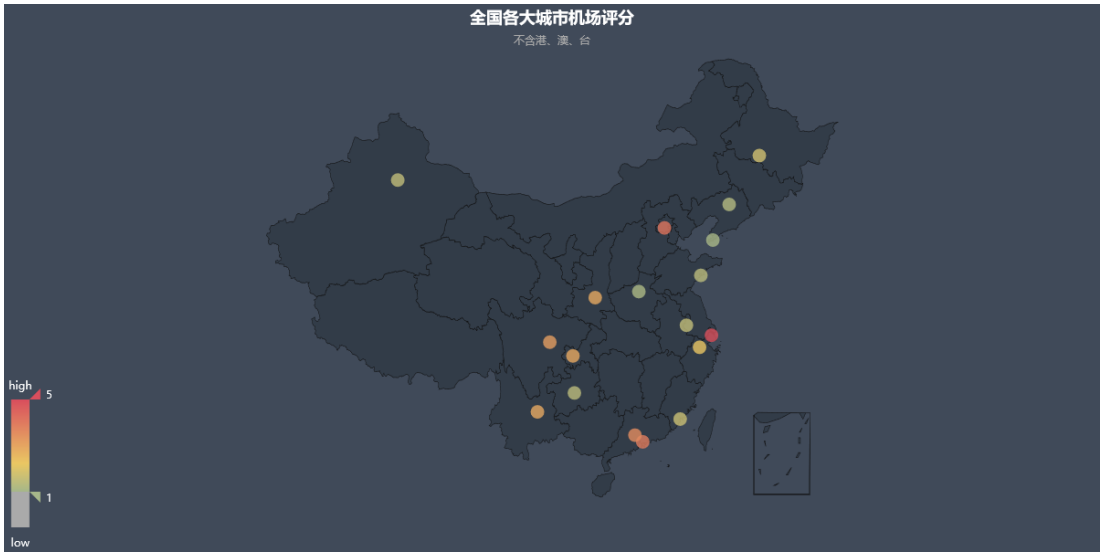


图 4: 全国重点机场展示

分，为在 1 分以上的所有城市。

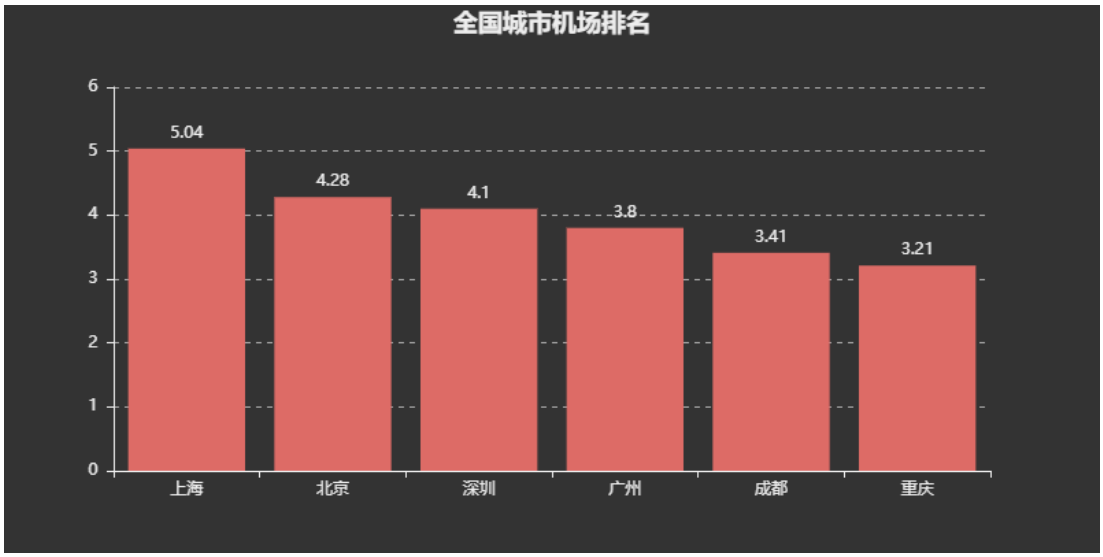


图 5: 全国排名前五机场 (按航线)

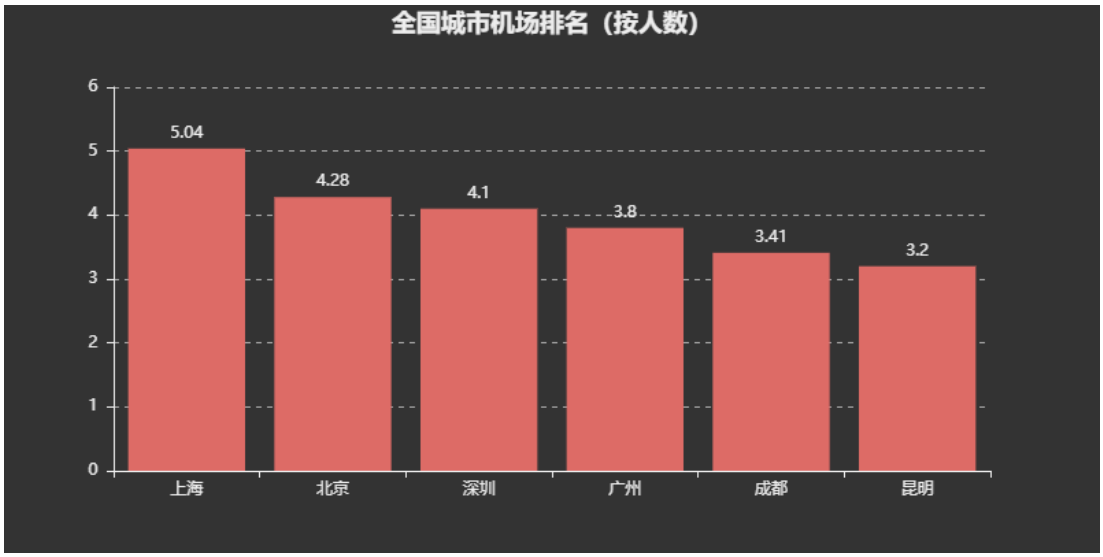


图 6: 全国排名前五机场 (按客流量)

上图为全国分数（按乘客数）排名前 6 的城市，分别为：上海——5.04，北京——4.28，深圳——4.1，广州——3.8，成都——3.41，昆明——3.2。上图为全国分数（按航班数）排名 1 到 10 名，11 到 20 名，21 到 30 名和 31 到 40 名的城市分数占比。上图为全国分数（按乘客数）排名 1 到 10 名，11 到 20 名，21 到 30 名和 31 到 40 名的城市分数占比。

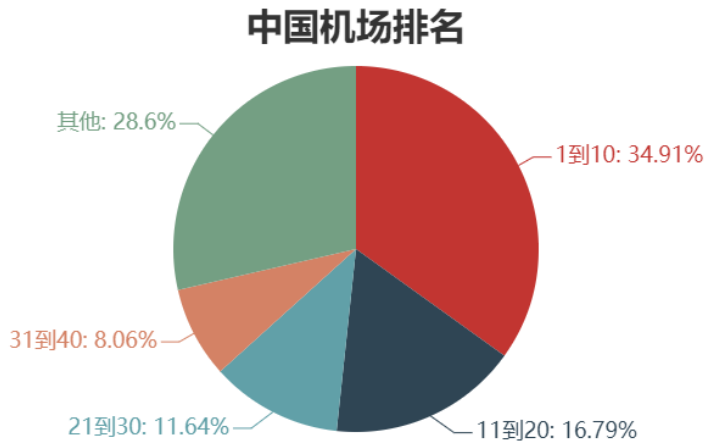


图 7: 中国机场排名 (按航班数)

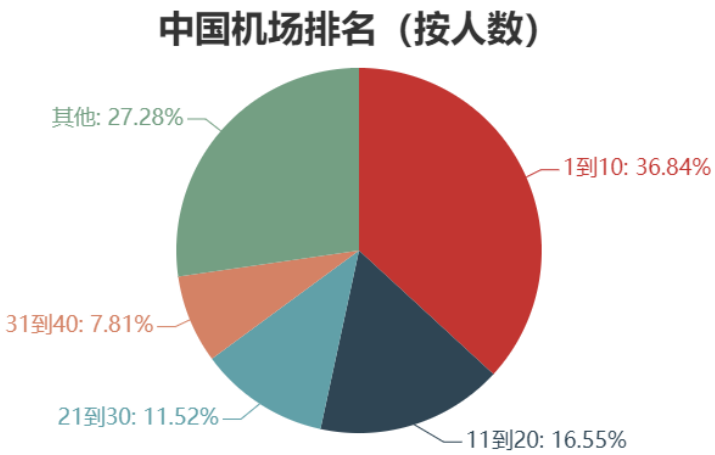


图 8: 中国机场排名 (按客流量)

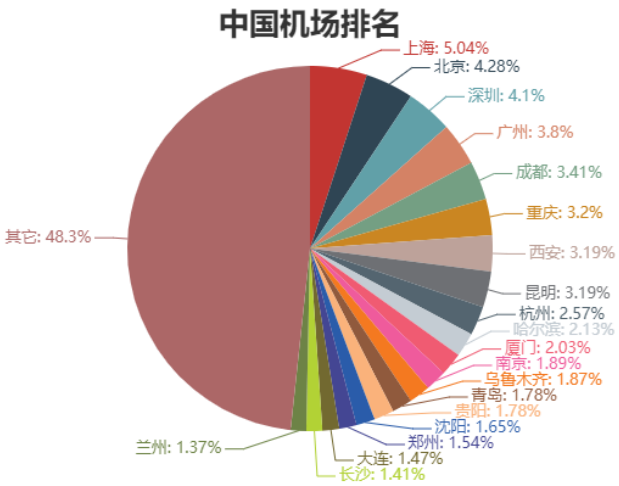


图 9: 中国机场排名 (按航班数)

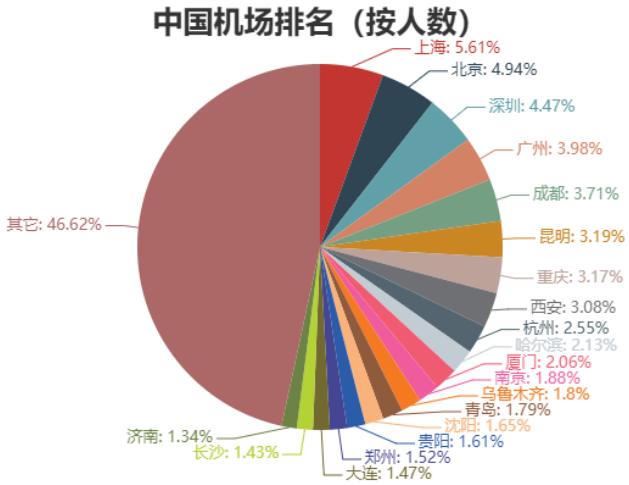


图 10: 中国机场排名 (按客流量)


图 11: 中国前五机场互连情况

从这几幅图可知，我国选用飞机出行的居民集中在较大的城市，而出行方式选用飞机的居民一般而言家庭境况都比较富裕，所以这从一方面说明了我国城市的经济资源在大城市中集中。所以这说明我国城市的经济发展仍然不是很均衡，大小城市之间发展水平差异较大。从而我们推断，我国的城市化还远远没有结束，这也佐证了我国仍然为发展中国家的事实。

而且中国城市的发展程度大约符合二八准则，即大约 20% 的城市拥有大约 80% 的经济资源。在在后一张图中，可以看出，前 20 的城市占用了中国超过一半的经济资源。

## 4.2 航空公司

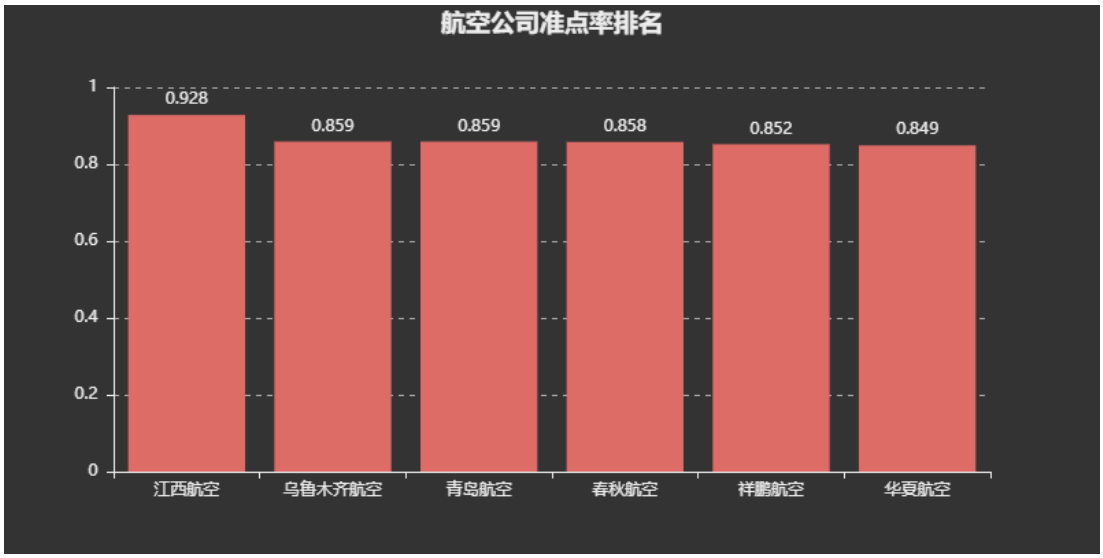


图 12: 航空公司准点率排名

上图为准点率排名前 6 的航空公司排名，分别为：江西航空——92.8%，乌鲁木齐航空——85.9%，青岛航空——85.9%，春秋航空——85.8%，祥鹏航空——85.2%，华夏航空——84.9% 上图为价格由高到低排名前 6 的航线数大于 200 的航空公司的平均票价，分

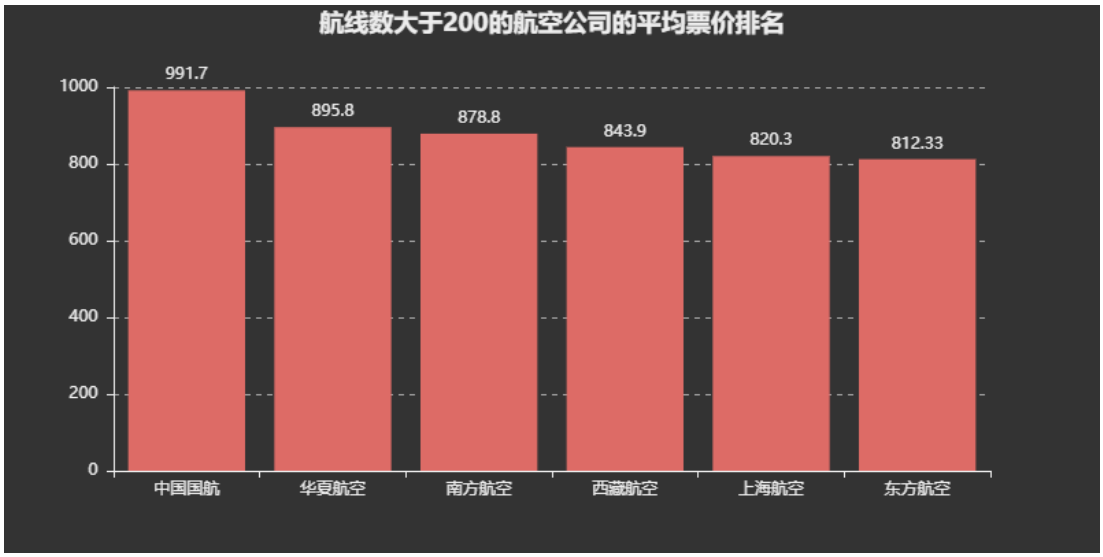


图 13: 航空公司票价排名 (高价票)

别是：中国国航——991.7，华夏航空——895.8，南方航空——878.8，西藏航空——843.9，上海航空——820.3，东方航空——812.33。

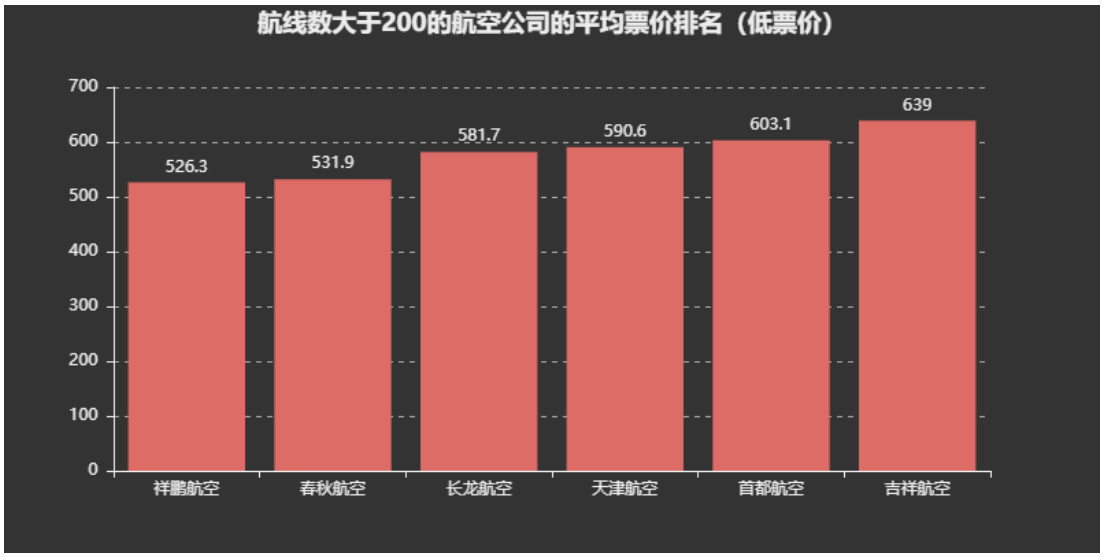


图 14: 航空公司票价排名 (低价票)

上图为价格由低到高排名前 6 的航线数大于 200 的航空公司的平均票价，分别是：祥鹏航空——526.3，春秋航空——531.9，长龙航空——581.7，天津航空——590.6，首都航空——603.1，吉祥航空——639。

### 4.3 分工

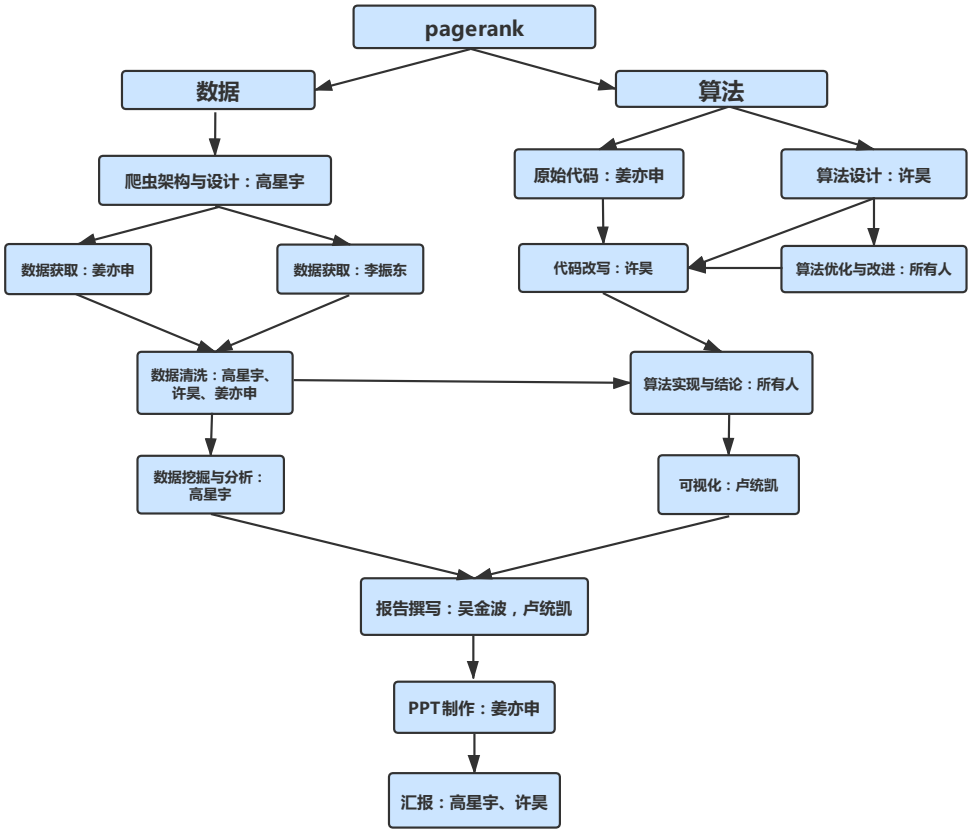


图 15: 分工

## 参考文献

[1] Page L, Brin S, Motwani R, et al. The PageRank citation ranking: Bringing order to the web[R]. Stanford InfoLab, 1999.

[2] Mihalcea R , Tarau P . TextRank: Bringing Order into Texts[C]// Proc Conference on Empirical Methods in Natural Language Processing. 2004.