豆瓣高分纪录片分析报告

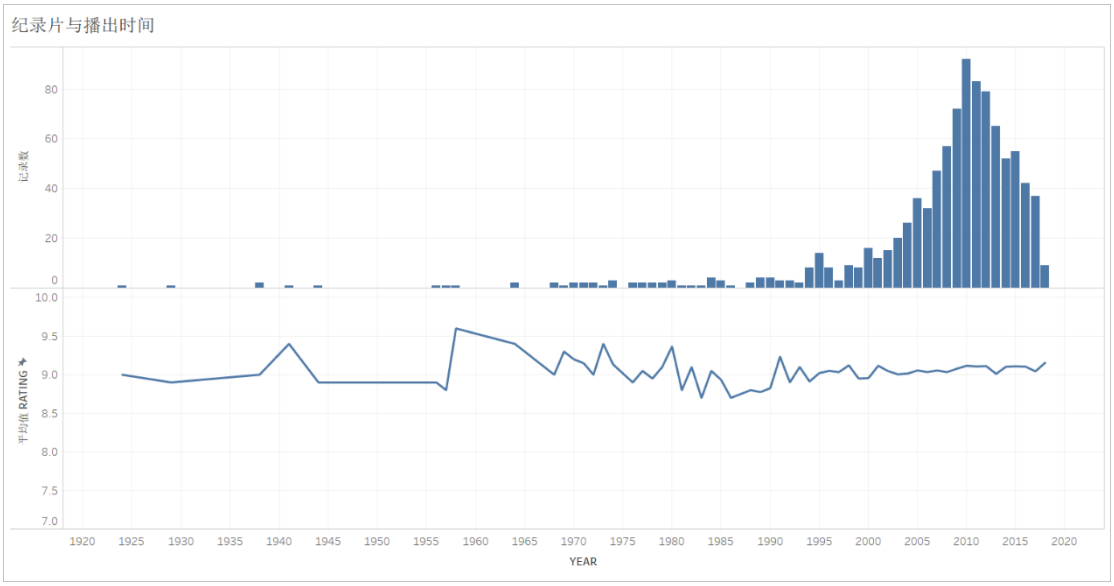
作者：王廷风
邮箱：wangtingfeng@bupt.edu.cn
源代码：github.com/dafengzai

本人在网上看到过许多的豆瓣高分电影数据的分析，便萌生了也做一份豆瓣数据的分析报告的想法，由于本人平日也喜爱看记录片，故在此做出一份豆瓣高分记录片的分析报告。需要注意的是，本次所用数据仅代表豆瓣用户心目中的高分纪录片的情况，并不完全代表这些纪录片的优劣。

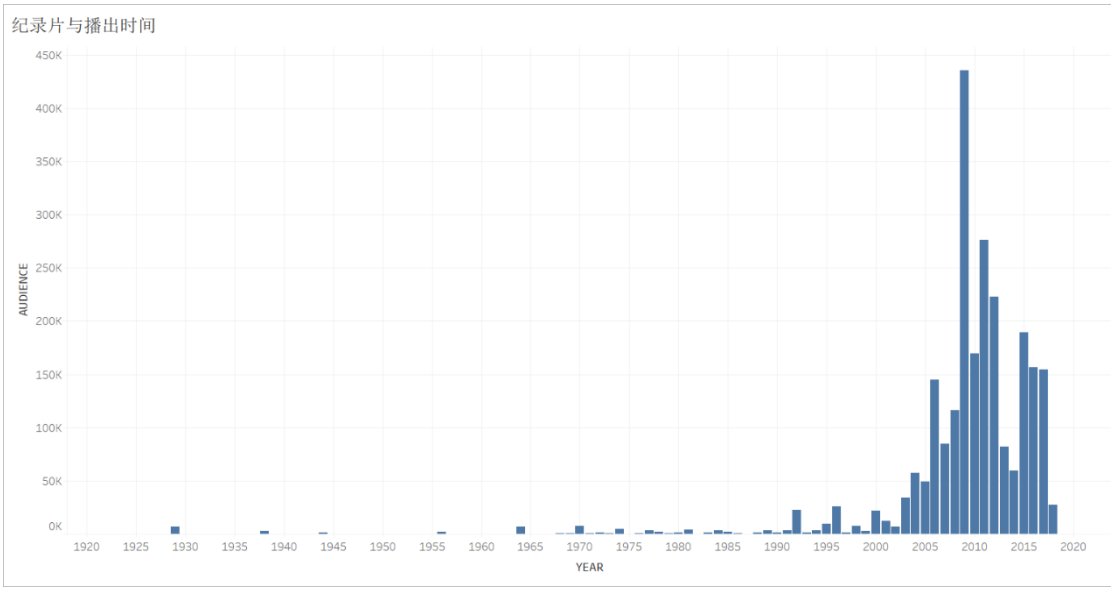
本次数据来源与豆瓣网页的抓取，选取了记录片的电影标签后按照评分排序，由于豆瓣页面显示的限制，共计抓取了按评分排序后的所有 50 页共计 968 部高分记录片的数据，评分由 9.9 分至 8.6 分，需要注意的是依照评分排序时豆瓣已经帮助我们过滤掉了无效的数据（评分人数过少），因此抓取的这些高分记录片都具有一定的代表性。抓取的纪录片数据包括：电影的名称、播出时间、豆瓣评分、评分人数、制片国家/地区、还有被豆瓣用户标记的标签。利用以上的几个维度，下面做一些分析。

（一） 高分纪录片的播出时间

得到数据后我第一时间就想知道高分纪录片与时间的关系，看看豆瓣用户喜欢的高分纪录片都是在哪些时候播出的，便做出下图：



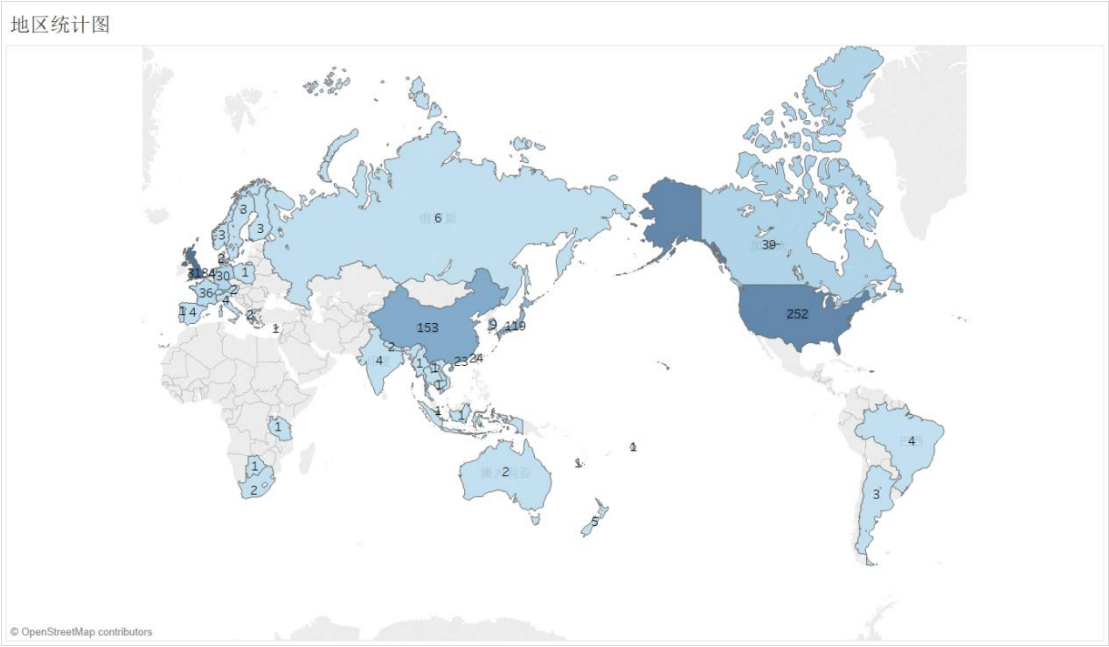
上图表示高分纪录片与播出时间之间的关系。共有 2 个数据，上面是播出时间与高分纪录片数目的关系，下图是播出时间与平均评分的关系。首先我们能注意到大多数的高分纪录片都是在 2000 年后播出的，而在 1990 年以前的数据更是寥寥无几，由于靠前的年份数据过少，平均评分的波动很大，而在 2000 年后的高分记录片里评分便稳定在 9.0 以上。从上图我们可以看出高分记录片最早可以追溯到 1925 年，记录片也是一个十分悠久的类型，那么究竟是什么原因造成 1990 年前的高分记录片数据这么少呢？下面进一步分析。



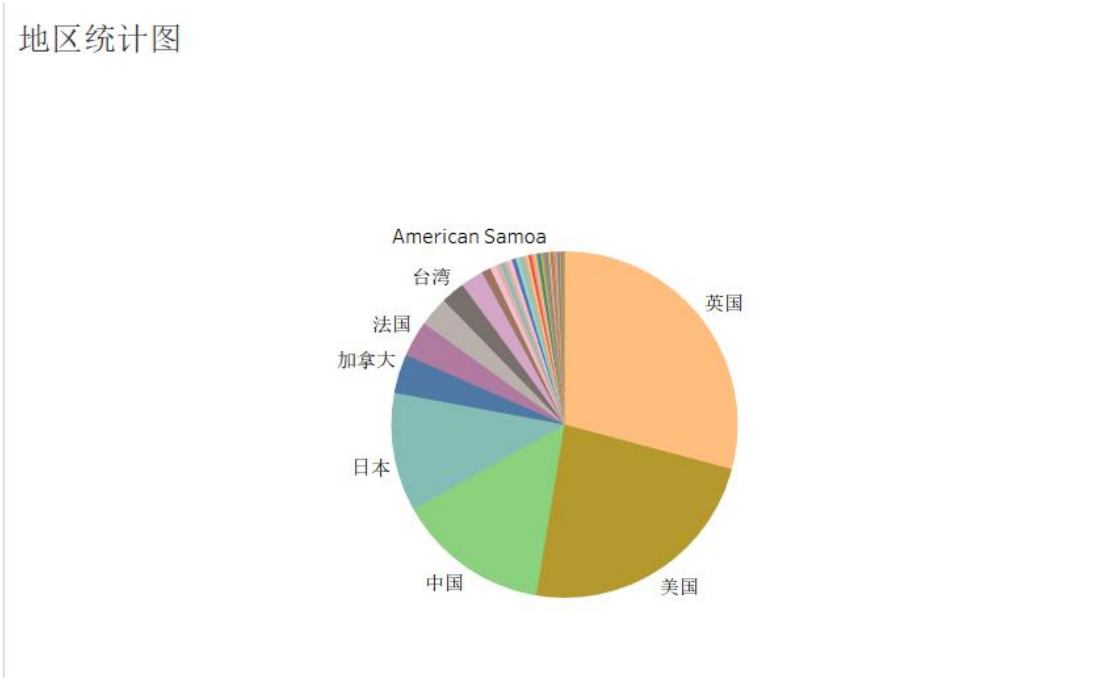
上图表示豆瓣用户评分人数与年份间的关系。从中可以看出，豆瓣用户都十分喜爱 2000 年后播出的纪录片，评分的用户数与 2000 年之前的纪录片完全不在一个数量级。豆瓣用户十分喜爱近年来播出的纪录片，而对于年代久远的则不太感冒。这也能解释为什么上榜的高分记录片中 1990 年以前的这么少，因为豆瓣用户不太热衷于老旧的记录片，许多优秀但比较老的记录片因此没能进入豆瓣用户的视野，自然上榜的与近年的相比便十分的少。

（二） 高分纪录片与制片国家/地区

与电影类似，优秀的纪录片的拍摄也是需要极大的资源的，比如知名的 BBC 的纪录片《蓝色星球》，便动用了当时最先进的摄影机与水下设备，为观众带来了无比震撼的水下美景。因此我认为高分记录片也应该集中于几个有着优秀摄影制作团队的国家/地区内。下面便来验证下我的想法。



上图表示国家/地区制作发行的高分记录片的数目，蓝色由浅到深对应数量的少到多。从中可以确认我的判断，地图中除了少数几个深蓝以外，大多数国家/地区制作的优秀纪录片数目都与这些“纪录片大国”不在一个数量集，下图是详细数据。



可以看到头部国家/地区与其余的相差巨大，豆瓣的高分纪录片近 80%都出自于几个国家/地区。从先前的分析可以看出：要想挤入豆瓣记录片高分榜，作品质量与热度缺一不可。头部制片国家/地区拥有强大的综合国力与优秀的制片团队。特别是第一的英国，更是具有 BBC 电视台这样的全球顶级团队，产出的作品质量与热度兼备，自然能够让豆瓣用户所喜爱。

（三） 高分纪录片的标签频率统计

在每部影片的页面内，都有一个“豆瓣成员常用的标签”栏，其内包含了许多豆瓣用户标注的该部影片的标签。我认为这些标签能后很好地反应出该影片的特征。因此下面我将对这些高分记录片的标签进行探究。

上图为高分纪录片的所有高频标签的统计图。每个标签出现次数的计算方式为：

SUM(被标注影片*该影片评分用户数)。从上图中我可以再次确认之前的结论：如大多数纪录片播出年份集中在 2000 年后、制片国家/地区主要集中在几个国家内（上图巨大的中国、美国、英国的标签）。

下面再来查看标签的总体情况：

count	1862.000000
mean	6098.974758
std	26468.961240
min	124.000000
max	552797.000000

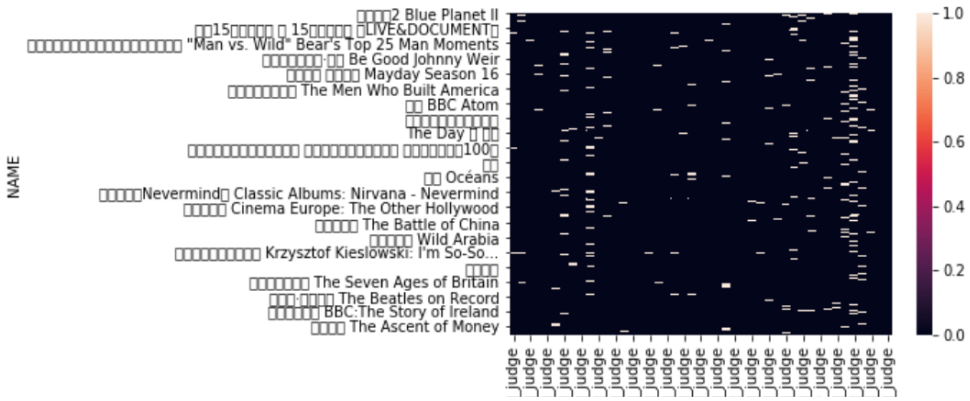
可以看出标签分布不均匀的情况实在是太严重了，平均数也才为 6099，与最大值整整差了 2 个数量集。绝大多数标签出现的次数很少，与 1800 多的标签种类数相比，只有几十个标签出现次数与最大值 552,797 处在同一数量集内，我称这些标签为“热门标签”。它们代表着那些受豆瓣用户喜欢的纪录片类型。

（四） 通过标签探究高分纪录片的类别

下面对这些标签数据进行进一步的分析，我认为标签能够反应出一部电影的特征，那么通过这些标签我就能研究下豆瓣高分纪录片的特征，以便进一步分析豆瓣用户们都喜欢看哪些类型的纪录片。

首先我将查看除去时间与国家/地区内容后标签的分布情况，相同的国家与摄影团队制作的纪录片可以包罗万象，比如纪录片高分榜的常客 BBC 电视台，就制作过自然类：《人类星球》（9.8 分）、科普类：《神秘的混沌理论》（9.1 分）以及历史类：《二战全史》（9.4 分）。因此我认为如果将上述的这几部影片都归结到“BBC”类型的记录是有失偏颇的。下面查看出去时间与国家/地区标签后的词频图：

首先可以看到“自然”、“人性”、“历史”这几个标签占据着很大的比重，那么我们是否就可以断定高分纪录片主要有这几类组成的呢？还需进一步的分析，下面先查看标签热力图：



每一行表示纪录片的名字，而每一列代表每一个标签，白色（1.0）表示出现该标签，黑色（0.0）则表示没有。通过上图可以看到，即便我选取的已经是“热门标签”了，但上图仍十分的“稀疏”（绝大多数标签没在影片中被标注），除了一些极其热门的标签（某一列许多白色，代表着影片被标注有该标签）。下面再进行进一步的建模操作，对标签数据进行聚类，探究豆瓣用户喜欢的高分纪录片究竟可以被分为哪些类别的。

首先我意识到我的样本共有 968 个而提取出来的标签则有 45 个，且与样本组成的特征空间十分的稀疏，（热力图中一行可以看作样本在特征空间中的一个“向量”）因此对特征进行降维十分有必要，使用 PCA 主成分分析法将特征减少到 26 个，这些主成分的累计贡献率约为 0.95，可以用作下一步的聚类操作。

前面提到每一个样本都可以看作是特征空间里的一个向量，每一个维度都是一个标签，我认为这些距离越相近的那些“向量”，其影片的特点也越相近。本次我将采用 K 均值聚类的方法来探究数据的分类情况。

进过本人的多次实验，最终决定将数据分为 6 大类别，模型的总的轮廓系数达到了 0.4。（总的轮廓系数取值为[-1, 1]，越接近 1 说明各个类别中的样本越近似）各个类别所占的比重为：

The rate of centers0 is 0.562436:
The rate of centers1 is 0.123839:
The rate of centers2 is 0.092879:
The rate of centers3 is 0.098039:
The rate of centers4 is 0.043344:
The rate of centers5 is 0.079463:

可以看出 centers0 类的纪录片占据了半壁江山，而其他类别的纪录片分布较为均匀，除 centers4 外都在 10%左右。下面就要来探究各个类别究竟代表着怎样的纪录片类型。

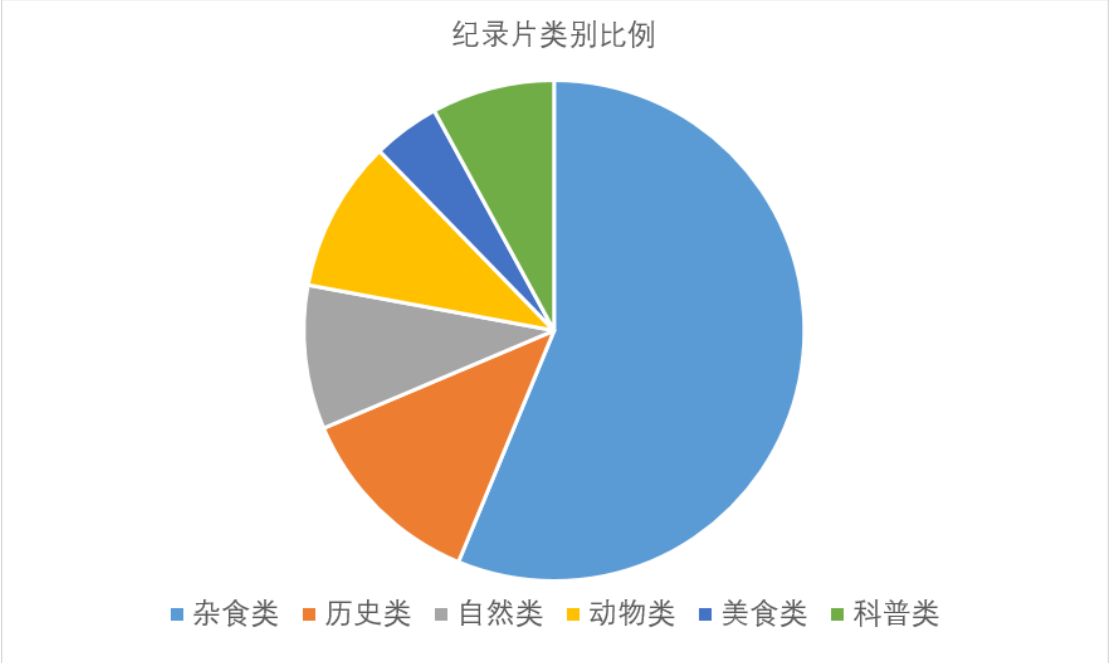
寻找各个类别的数据的中心点，将这些降维后的数据还原，因为我将样本中出现“热门标签”对应 1，没有则对应于 0。所以若一个样本中某个标签对应的数据越接近于 1，则倾向于该样本具有该标签。通过观察这 6 个聚类的中心样本数据，我可以观察到这 6 个聚类分别对应着哪 6 种类型的纪录片。各个中心点数据还原后按标签对应数值从大到小排列后的情况为：

centers 0: 艺术_judge 0.0623713197062 旅行_judge 0.0605519978249	centers 1: 历史_judge 0.999960734488 战争_judge 0.158673196564
centers 2: 自然_judge 1.00019329519 海洋_judge 0.121860041686	centers 3: 动物_judge 0.999949117407 自然_judge 0.652643242407

centers 4: 美食_judge 0.999734073418 吃货_judge 0.0497596757622	centers 5: 科普_judge 0.999907979876 自然_judge 0.129847939441

下面便可以得出我的结论：

1. 第一类（centers 0）为杂食类。在这个类别中标签对应数值的最大值也不到 0.1，远低于 1。这说明该类样本中倾向与不具有任何“热门标签”，即该类纪录片倾向于没有被用户标注任何一个热门的标签，这个类别的纪录片从标签情况上来看不仅与其他类别相去甚远，在其自身内部也不相近，可以认为这个类别里包含了各种各样的纪录片类型，这个类别属于杂食类。且该类别占所有样本的比重超过一半，可以说明豆瓣用户对纪录片的品味还是十分“杂食”的。
2. 第二类（centers1）为历史类。
3. 第三类（centers2）为自然类。
4. 第四类（centers3）为动物类。同时被标注“动物”标签的记录片中，同时出现“自然”标签的情况也比较普遍。
5. 第五类（centers4）为美食类。
6. 第六类（centers5）为科普类。



（五） 总结

豆瓣纪录片高分榜共有 968 个影片，播出时间横跨 20 至 21 世纪，制片国家/地区分布于世界各地，但这不代表豆瓣用户对这些纪录片的关注是分散的。豆瓣用户更加喜欢观看 2000 年后的记录片，喜爱的纪录片大多都出自某几个国家与地区，BBC 电视台更是出产了相当多豆瓣用户喜爱的纪录片。从纪录片的类型来看，豆瓣高分纪录片的类型可以说是包罗万象，超过一半的高分榜纪录片不能归结于某一特定的类型，这也从另一角度说明了豆瓣用户喜好的多样性，而近一半的高分纪录片可以分类为：历史、自然、动物、美食、科普这几大类，可以看出豆瓣用户对这几大类型的纪录片十分的热爱。