图分析大作业文档

清华大学软件学院 李肇阳 2014013432

2015 年 12 月 26 日

目 录

§1 数据说明

§1.1 数据采集

自行采集数据。

用JavaScript和Python编写了爬虫，抓取豆瓣电影[1]的Top250榜单、该榜单中每一部电影下的长影评。

于是，每个电影都对应于一个“观众集合”：给该电影写过长影评的用户集合。

§1.2 模型建立

用以上数据构建**正权无向完全图**。

结点表示电影，边权表示两部电影的关系。为尽量使分析结果有实际意义，在不同分析项目中，边权的取值不同。

- 在作“最短路径”、“最小生成树”分析时，边权表示两部电影的**相异**程度，为其观众集合的**杰卡德距离**。
- 在作“连通分量”、“社群结构”分析时，边权表示两部电影的**相似**程度，为其观众集合的**杰卡德相似系数**。

该图含有246个节点，28971条边。

[1] http://movie.douban.com/

§2　分析项目

§2.1　最短路径

求给定两个节点之间的最短路径。

采用正权图上的单源最短路径Dijkstra算法。以C++实现。

封装在类中，对外提供API：接受邻接矩阵、起终点编号，返回最短路径的长度、途径节点的列表。

实现了查询最短路径的交互式图形界面，可显示最短路径长度、依次列出其上所有节点。

§2.2　最小生成树

求图中的一颗最小生成树。

采用求最小生成树的Prim算法。以C++实现。

封装在类中，对外提供API：接受邻接矩阵，返回包含于最小生成树中的边的集合。

以sigmajs[2]库实现了最小生成树的可视化，采用力导向算法布局。

§2.3　连通分量

以结点所有关联边的权之和进行过滤。相当平凡，在此略去。

§2.4　社群结构

希望通过找到图中的社群结构，来实现不依赖任何已有知识地、自动化地对电影分类。

用Wolfram Language先后尝试了模块度聚类、中心度聚类、小团体渗透、层次聚类、谱聚类等五种聚类算法。

§3　分析结果

§3.1　交互式可视化

多项分析结果以HTML、CSS、JavaScript之Web前端三剑客进行了可视化。请使用现代浏览器（如版本号足够高的Google Chrome）打开front/index.html查看。

用户可以进行视图的放大缩小平移，可以点击节点查看详细信息，可以拖动节点改变布局，可以链接跳转到豆瓣电影相应页面上。

用户可以通过动画查看最小生成树，可以制定任意两个节点查看其最短路径，可以查看路径上任意节点的详细信息。

§3.2　电影自动分类

寻找社群结构的分析结果，并未找到合适的方法进行可视化。以文字形式展现如下：
豆瓣电影TOP250被分成了11类：

- A：无耻混蛋, 禁闭岛, 盗梦空间, 机器人总动员, 贫民窟的百万富翁, 记忆碎片, 让子弹飞, 本杰明·巴顿奇事, 朗读者, 致命魔术, 阿凡达, 穆赫兰道, 蝙蝠侠：黑暗骑士, 飞屋环游记, 当幸福来敲门, 撞车, 入殓师, 三傻大闹宝莱坞, 天堂电影院, 搏击俱乐部, 黑天鹅, 窃听风暴, 国王的演讲, 月球, 蝴蝶效应, 鬼子来了, 大鱼, 玛丽和马克思, V字仇杀队, 源代码, 岁月神偷, 一一, 放牛班的春天, 暖暖内含光, 这个男人来自地球, 恐怖游轮, 海上钢琴师, 告白, 香水, 真爱至上, 大话西游之大圣娶亲, 死亡诗社, 浪潮, 断背山, 人工智能, 楚门的世界, 荒野生存, 恋恋笔记本, 忠犬八公的故事, 傲慢与偏见, 曾经, 蝙蝠侠：黑暗骑士崛起, 初恋这件小事, 廊桥遗梦, 爱在暹罗
- B：教父, 猜火车, 低俗小说, 发条橙, 重庆森林, 阿飞正传, 七宗罪, 辛德勒的名单, 燃情岁月, 西西里的美丽传说, 英国病人, 美国丽人, 阿甘正传, 勇敢的心, 爱在日落黄昏时, 美丽心灵, 剪刀手爱德华, 这个杀手不太冷, 阳光灿烂的日子, 心灵捕手, 东邪西毒, 爱在黎明破晓前, 闻香识女人, 美丽人生, 肖申克的救赎, 天使爱美丽, 飞越疯人院, 霸王别姬, 泰坦尼克号, 甜蜜蜜, 活着, 碧海蓝天, 情书, 蓝色大门

[2]http://sigmajs.org/

- C：教父3, 教父2, 指环王1：魔戒再现, 指环王2：双塔奇兵, 虎口脱险, 摩登时代, 哈利·波特与魔法石, 雨中曲, 终结者2, 控方证人, 英雄本色, 音乐之声, 冰川时代, E.T.外星人, 射雕英雄传之东成西就, 变脸, 角斗士, 纵横四海, 哪吒闹海, 上帝也疯狂, 巴黎淘气帮, 未麻的部屋, 伴我同行, 末代皇帝, 寿司之神, 迁徙的鸟, 速度与激情5, 假如爱有天意
- D：加勒比海盗, 惊魂记, 黄金三镖客, 上帝之城, 沉默的羔羊, 罗生门, 七武士, 绿里奇迹, 卡萨布兰卡, 罗马假日, 雨人, 美国往事, 与狼共舞, 乱世佳人, 钢琴家, 魂断蓝桥, 末路狂花, 花样年华, 夜访吸血鬼, 跳出我天地, 东京物语, 帝企鹅日记
- E：怪兽电力公司, 狮子王, 天空之城, 哈尔的移动城堡, 驯龙高手, 秒速5厘米, 风之谷, 幽灵公主, 魔女宅急便, 龙猫, 千与千寻, 玩具总动员3, 神偷奶爸, 萤火虫之墓, 无敌破坏王, 侧耳倾听, 7号房的礼物, 萤火之森, 刺猬的优雅
- F：谍影重重2, 谍影重重3, 偷拐抢骗, 谍影重重, 猫鼠游戏, 拯救大兵瑞恩, 战争之王, 致命ID, 第六感, 卢旺达饭店, 勇闯夺命岛, 血钻, 千钧一发, 恐怖直播, 导盲犬小Q, 唐伯虎点秋香, 黑鹰坠落, 荒岛余生, 喜剧之王
- G：疯狂原始人, 辩护人, 一次别离, 触不可及, 我是山姆, 幸福终点站, 少年派的奇幻漂流, 时空恋旅人, 超脱, 怦然心动, 遗愿清单, 借东西的小人阿莉埃蒂, 叫我第一名, 海洋, 哈利·波特与死亡圣器(下), 地球上的星星, 蝴蝶, 爱·回家, 八月迷情
- H：饮食男女, 十二怒汉, 春光乍泄, 海盗电台, 狩猎, 杀人回忆, 牯岭街少年杀人事件, 燕尾蝶, 穿条纹睡衣的男孩, 穿越时空的少女, 小鞋子, 菊次郎的夏天, 中央车站, 阳光姐妹淘, 海豚湾, 被嫌弃的松子的一生, 素媛, 青蛇
- I：指环王3：王者无敌, 城市之光, 黑客帝国, 黑客帝国3：矩阵革命, 喜宴, 大闹天宫, 无间道, 新龙门客栈, 可可西里, 大话西游之月光宝盒, 麦兜故事, 倩女幽魂, 枪火, 莫扎特传, 两小无猜
- J：两杆大烟枪, 非常嫌疑犯, 追随, 完美的世界, 布达佩斯大饭店, 疯狂约会美丽都, 红辣椒, 再见列宁, 勇士, 我在伊朗长大, 不一样的天空, 忠犬八公物语, 我们俩, 我爱你
- K：洛城机密, 梦之安魂曲, 大卫·戈尔的一生

值得再次强调的是，以上分类没有利用任何先验知识（如地区、导演、类型、风格、关键词等），所根据的仅仅是用户“发表影评”这一行为。

可以看到一些可喜的地方，比如《谍影重重》三部曲被分到了同一类下（F）；有一类下清一色全部是动画片（E）。这表明，根据用户观影偏好（由发表影评的行为反映），对电影做不依赖已有知识的、自动化的分类，这是可行的。

由于精力有限，未对该分类结果的合理性做进一步的评估。

§4 感言

坦率地讲，最短路径、最小生成树、中心度这些分析项目，对于我这个电影作为节点的图来讲，都意义不大，对我也没什么吸引力。

而最迷人、也是令我最得意之处，就在于最后完成了对TOP250电影的自动分类，得到了可喜的结果。通过这一过程，我亲身体验了图论课程所介绍的模型和方法的实用性，看到了它是如何来解决实际问题的。

通过完成此份作业，我对图论课程介绍的若干算法有了更深的理解，实际编程能力也得到了锻炼。收获颇丰。

（完）

参考文献