Netflix 的推荐系统

施朱鸣

2022 年 7 月

Netflix 对推荐系统的需求

- 用户可以顺序浏览或者根据关键词搜索电影，比如电影类型、导演和演员
- 用户可以给看过的电影打 1-5 分
- 即使不打分，Netflix 仍然知道你订阅过哪些电影
- 相比于 amazon 等的推荐系统，Netflix 用户更有积极性打分
- 在用户给一定数量电影打分后，Netflix 开始推荐电影

Netflix 的推荐系统比赛

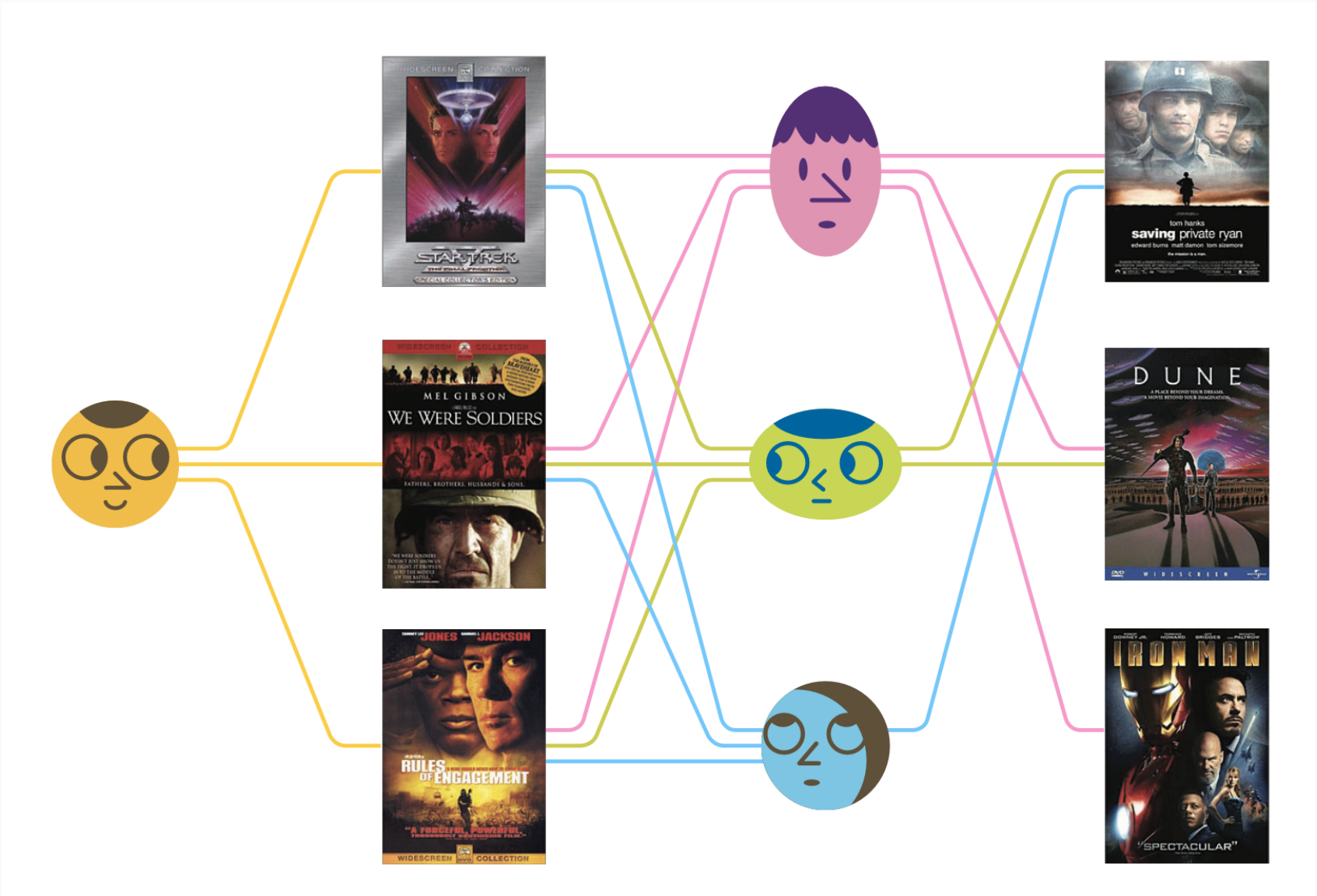
- 提供 480000 个匿名用户对 17000 部电影的 1 亿条打分
- Netflix 用 3 百万条数据，让参赛者预测打分
- Netflix 比较参赛者打分和真实打分的均方根误差

协同过滤 collaborative filtering

一种策略：最近相邻策略 nearest-neighbor approach

- 什么是 neighbor？一部电影的 neighbor 是，由同一个观众打分时，得分最相似的一类电影，比如拯救大兵瑞恩的 neighbor 是其他一些战争片、斯皮尔伯格导演的电影和阿汤哥演的电影
- 如何根据一部电影的 neighbor 预测该观众对该电影的评分？对该电影的 50 个以内的 neighbor 的打分加权平均

最近相邻策略 nearest-neighbor approach

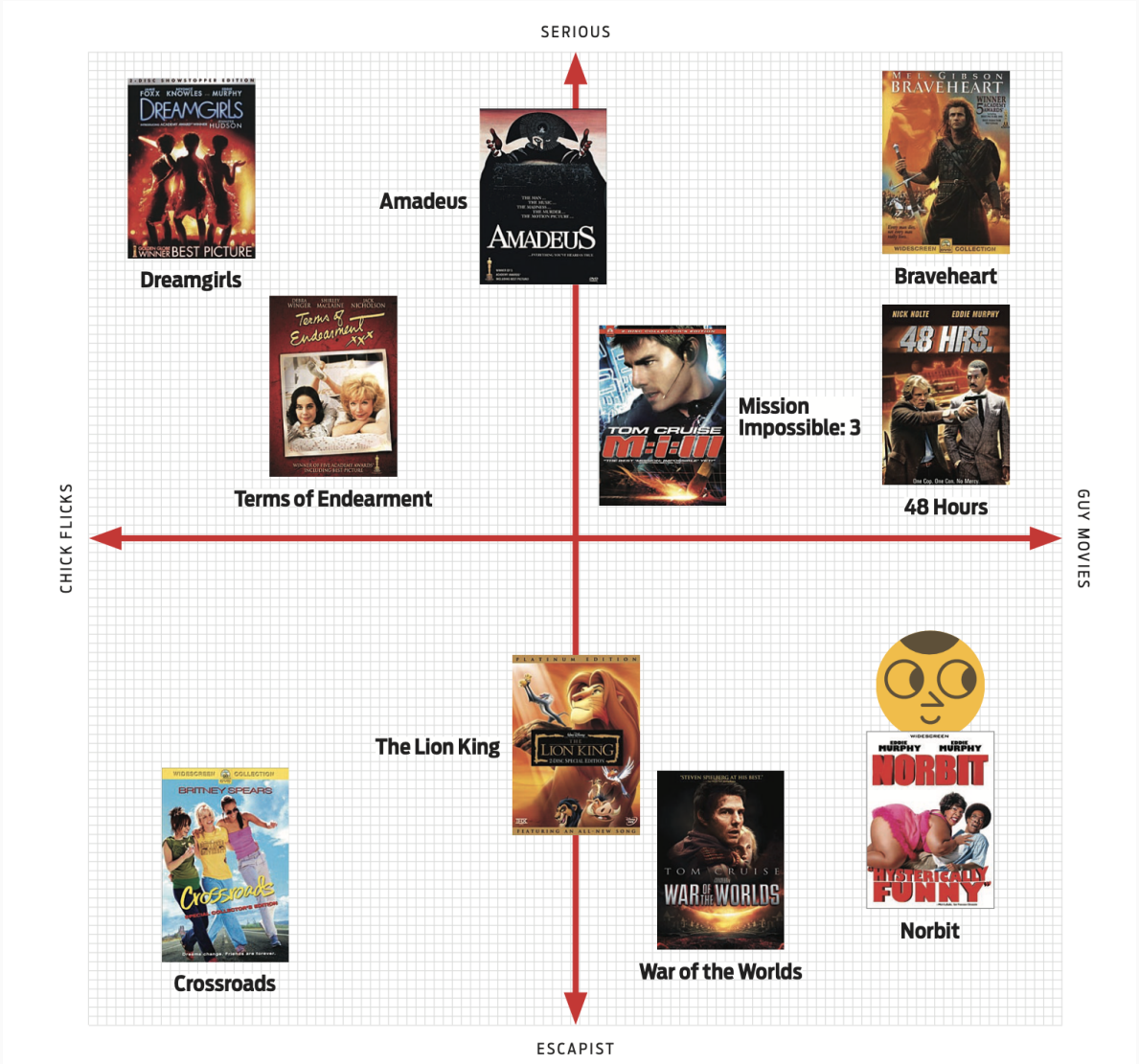


协同过滤 collaborative filtering

另一种策略：隐因子模型 latent-factor models

- 同时对观众和电影建模，用多个维度刻画电影和观众
- 刻画电影的维度，比如喜剧还是情节剧，动作片还是爱情片，面向儿童还是成年人等，这些维度由算法得出，可能难以解释
- 刻画观众的维度，比如观众有多么喜欢喜剧，多么喜欢动作片
- 如何预测评分？使用 20-40 个维度对观众和电影进行定位，根据电影在观众关心的维度上的评分来预测打分

隐因子模型 latent-factor models



算法的问题

- 最近相邻策略只在 50 个以下 neighbor 时表现较好，无法利用其他全部数据
- 隐因子模型没法联系一些关系紧密的电影

作者的最终算法

- 使用整体方法 ensemble approach 来结合这两种方法
- 需要避免过拟合，打分很多的观众的行为模式，不能套到打分很少的观众身上
- 关注观众对哪些电影打过分，即使一部电影在某个维度上不讨喜，观众仍然可能喜欢它（分数由数值变成 0-1 变量）

Netflix 从比赛中得到了什么

- 比赛得到的算法已经整合进 Netflix 的推荐系统
- 给研究人员提供了高质量数据集
- 给 Netflix 带来了很多人才

致谢

谢谢大家聆听！

联系方式，欢迎私聊讨论

邮箱:shizhuming@pku.edu.cn