推荐系统研究与认识

摘要

互联网技术的迅猛发展把我们带进了信息爆炸的时代. 海量信息的同时呈现, 一方面使用 户很难从中发现自己感兴趣的部分, 另一方面也使得大量少人问津的信息成为网络中的 “暗信息”无法被一般用户获取. 个性化推荐系统通过建立用户与信息产品之间的二元关系, 利用已有的选择过程或相似性关系挖掘每个用户潜在感兴趣的对象, 进而进行个性化推荐, 其本质就是信息过滤. 个性化推荐系统不仅在社会经济中具有重要的应用价值, 而且也是一个非常值得研究的科学问题. 事实上它是目前解决信息过载问题最有效的工具, 现有主流的推荐系统分为协同过滤系统、基于内容的推荐系统、混合推荐系统等,本文主要介绍协同过滤算法的研究和认识.

**关键词：推荐系统; 系统过滤; 基于内容的推荐; 基于网络的推荐**

# 1. 引言

伴随着互联网的快速发展, 目前我们正处于一个信息爆炸的时代. 我们在享受着互联网带来获取知识、信息的便利性的时候, 同样苦恼于如何从浩瀚的知识海洋中快速有效地汲取到我们所需要的信息. 近年来, 推荐系统作为一种有效地解决信息泛滥的方法受到了大量国内、国外研究者的广泛研究. 其核心目标是通过分析用户行为、兴趣、需求等信息[1] , 在海量的数据中挖掘用户感兴趣的信息.

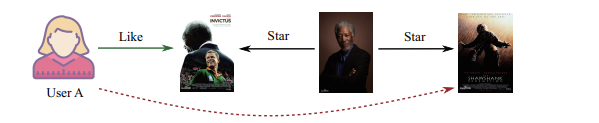


图1 一个基于可解释性图谱的推荐示例

目前, 推荐系统在各种各样的信息系统中都扮演着至关重要的角色, 用来促进商务业务发展以及辅助决策过程, 并且广泛地应用于各大电子商务 (淘宝、Amazon、eBay 等) 和多媒体 (MovieLens、豆瓣) 等网站中[2] .

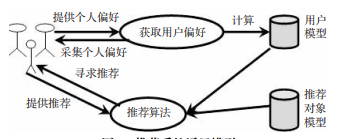
个性化推荐算法是推荐系统的核心, 其主要可以被分为 3 类, 即基于内容的推荐算法、基于协同过滤的推荐算法和混合推荐算法. 其中, 协同过滤方法通过利用用户历史的行为偏好数据构建模型, 取得了巨大的成功. 这种方法的优势在于不需要类似基于内容的推荐算法那样对物品进行复杂的特征提取与建模. 尽管这类协同过滤方法通常是有效且普适的, 但其依旧存在着多种问题. 主要包括用户和物品之间的行为关系数据的稀疏问题和对新用户或者新物品进行推荐时存在的冷启动问题[3] . 为此, 研究者尝试将协同过滤推荐算法和其他辅助信息相结合 (例如, 用户与物品的属性特征、用户社 交网络信息等) 搭建混合推荐系统来解决以上问题, 从而提升推荐效果[4] .

图2 推荐系统通用模型

一个完整的推荐系统由3个部分组成：收集用户信息的行为记录模块、分析用户喜好的模型分析模块和推荐算法模块. 行为记录模块负责记录用户的喜好行为, 例如问答、评分、购买、下载、浏览等. 问答和打分的信息相对好收集, 然而有的用户不愿意向系统提供这些信, 那么就需要通过其他方式对用户的行为进行分析. 例如购买、下载、浏 览等行为, 通过这些用户的行为记录分析用户的潜在喜好产品和喜欢程度. 这就是模型分析模块要完 成的工作, 模型分析模块的功能能够对用户的行为 记录进行分析建立合适的模型来描述用户的喜好 信息. 最后是推荐算法模块利用后台的推荐算法, 实时地从产品集合中筛选出用户感兴趣的产品 进行推荐, 其中推荐算法模块是推荐系统中最为核心的部分.

# 2. 协同过滤推荐系统

## 2.1 协同过滤

协同过滤推荐（collaborative filtering recommendation）是推荐策略中最成功的策略, 它于20世纪90年代开始研究并促进了整个推荐系统研究的繁荣. 大量论文和研究都属于该类别. 比如Grundy书籍推荐系统、Tapestry邮件处理系统, Grouplens、Ringo等推荐系统都属于该类推荐[5] .

协同过滤推荐的基本思想借鉴了日常在选购商品、选择 用餐饭店、选择看哪部电影等等的方法. 如果自己身边的很 多朋友都选购某种商品, 那么自己就会很大概率的选择该商 品. 或者用户喜欢某类商品, 当看到和这类商品相似商品并 且其他用户对此类商品评价很高时, 则购买的概率就会很大. 协同推荐的用户模型为用户-项目评价矩阵, 如表1中表 示Ri, j第i个用户对第j个项目的评分.

|  |  |
| --- | --- |
|  |  |
| 图3 基于用户的协同过滤推荐过程 | 图4 用户评分矩阵 |

基于用户的协同推荐主要工作有两个, 首先是查询最近 邻居, 然后就是产生推荐. 其核心就是最近邻居查询. 要找 到最近邻居就要度量用户之间的相似性, 相似度越高, 用户就 越相近. 把用户i和用户j之间的相似性记为sim（i, j）. 每个用户对推荐对象的评分可看作是一个m维向量, 度量用户间 的相似度就可以用不同的m维向量间的相似度进行度量, 如 图4所示, 度量用户间相似性的方法主要有如下三种方法

1. 余弦相似性（Cosine）：设用户i和用户j在m维对象空 间上的评分表示为向量 i, j , 则sim（i, j）的相似性计算方法如式（1）：



1. 相关相似性（Correlation）：设用户i和用户j共同评分 的对象集合用 Iij 表示, 则用户 i 和用户 j 之间的相似性通过 Pearson 相关系数度量, 如式（2）所示：



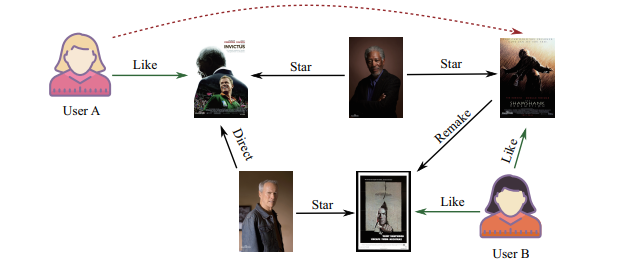
1. 修正余弦函数弦相似性（Adjusted Cosine）：在余弦相 似度量方法中没有考虑不同用户的评分尺度问题, 修正余弦 相似度量方法通过减去用户对推荐对象的平均评分改善上述 缺陷. 设用户i和用户j共同评分过的推荐对象集合用  表 示,  和 分别表示用户i和用户j评分过的推荐对象集合, 则 用户i和用户j之间的相似性  如式(3)所示



图5 基于用户的协同过滤

一些研究者通过原有的用户 – 物品的交互数据构成一个数据, 利用挖掘图上用户到物品之间不同的路径, 提升推荐效果[6] . 图 5 展示了这样一个异质信息网络图应用到推荐场景中的 示例.

## 2.2 相关算法

许多改进算法已经被广泛研究并且应用到标准的 相关性计算和夹角余弦公式中. 例如缺席投票(default voting)事例引申(case amplification)和加权优势预测[7] 等.

其中缺席投票是基于记忆方法的一种 扩展. 如果用户明确评分的产品数目很少上面提到 的算法得到的用户相似度都不准确. 原因在于这种相 似性的计算是基于用户 x 和 y 共同评过分的产品集合. 实证数据表明如果给一些没有打分的产品赋予 一些缺省的打分值那么预测分数的准确性将大幅度 提高.

应用相关性和夹角余弦方法计算产品之间的相似性. 这个思想被推广到基于产品相似性的 top-N 推荐算法 中即在进行推荐的时候只考虑相似度最高的 N 个产品并非所有的产品. 实验证明这种方法不仅比传统的基于用户邻居的推荐算法快1—2个数量级, 而且具有更好的推荐准确性.

利用用户产品列表中的先后次序计算用户之间的相似性排名靠前的产品在计算用户相似性的时候具有较高的权重. 利用用户的行为信息异构用户的兴趣点, 利用兴趣点计算用户之间的相似性. 实验证明, 这种方法比经典的协同过滤算法的推荐结果要好. 结果表明基于产品相似性的算法能够比基于用户相似性的算法得到更好的计算结果.

其他基于模型的协同过滤推荐系统有概率相关模型, 极大熵模型、线性回归、基于聚类 的 Gibbs 抽样算法、Bayes 模型等等.

前沿研究还有基于 Markov 决策过程的序列决策过程, 利用已有信息预测用户以后的喜好产品的概率, 在找到相应的产品进行推荐. 其他的概率模型技术包括概率潜层语义分析语义生成模型（aspect model）, 另外用一个简单的概率模型说明每个用户相对小的数据在协同过滤中是非常重要的. 从输入数据处理的角度提出了改进协同过滤的其他方法.

包括除噪音技术、选择用户打分集技术、余度分析和打分数据的稀疏性处理等. 数值结果显示这些方法可以提高基于模型的协同过滤算法的准确性和效率.

## 2.3 技术优缺点

### 2.3.1 协同过滤的优点

1. 复杂的非结构化的对象可以应用协同过滤, 比如电 影、音乐、图像等推荐对象.
2. 善于发现用户新的兴趣点. 协同过滤可以发现内容 上完全不相似的资源, 用户对推荐信息的内容事先是预料不 到的.
3. 不需要专业知识即可进行推荐.
4. 随着用户的增多, 其推荐性能会不断提升.
5. 以用户为中心自动进行推荐.

### 2.3.2 协同过滤的缺点

(1) 存在冷启动问题.

新进入的用户由于得不到他们的兴趣偏好而无法获得推荐, 新的推荐项目由于没有用户评价它就得不到推荐, 这就是冷启动问题. 冷启动问题是推荐系 统研究的难点和重点.

(2) 存在稀疏性问题.

由于用户数目的大量增长, 而且用 户之间选择存在差异性, 使得用户的评分差别非常大. 同时 推荐对象的数量也大量增长, 使得大量的推荐对象没有经过 用户的评价. 这些会导致部分用户无法获得推荐, 部分推荐 对象得不到推荐, 这就是稀疏性问题.

(3) 系统开始时推荐质量差及推荐质量取决于历史数据集.

# 3. 其他推荐系统模块

## 3.1 基于内容的推荐系统

随着机器学习等技术的 完善当前的基于内容的推荐系统可以分别对用户 和产品建立配置文件通过分析已经购买(或浏览)过的内容建立或更新用户的配置文件. 系统可以比较用户与产品配置文件的相似度,并直接向用户推荐与其配置文件最相似的产品.

例如, 在电影推荐中基于内容的系统首先分析用户已经看过的打分比较高的电影的共性(演员、导演、风格等), 再推荐与这些用户感兴趣的电影内容相似度高的其他 电影. 基于内容的推荐算法的根本在于信息获取和信息过滤. 因为在文本信息获取与过滤 方面的研究较为成熟现有很多基于内容的推荐系 统都是通过分析产品的文本信息进行推荐.

## 3.2 混合推荐系统

建立混合推荐系统的方法之一即是独立地应用协同过滤, 基于内容和基于网络结构的算法进行推荐. 然后将两种或多种系统的推荐结果结合起来, 利用预测打分的线性组合进行推荐. 又或者只推荐某一时刻在某一个评价指标下表现更好的算法的结果. 例Daily Learner系统就选择在某一时刻更可信的结果进行推荐[9] . 或者选择一个与用户过去的打分相一致的结果进行推荐.

# 4. 总结和展望

最近几年, 随着互联网的高速发展, 我们身处的信息世界的组织和结构有了很大的变化. 首先信息量的爆炸性增长使得一个普通用户搜寻自己感兴趣内容的难度和成本都提高了很多；其次大量的信息被安静地放在网络的旮角, 因为访问量小, 因此不为人知. 这些 “暗信息” 中或许有一些是用 户感兴趣的, 但是没有外界的帮助, 普通用户根本 无法找到它们. 随着 Web2 0技术的发展, 很多服务型网站可以保留用户的历史记录, 这些记录包括 选择、评价、购买等等. 通过这些记录, 采用合适的推荐算法, 可以在一定的准确程度上猜测用户喜好, 并据此向用户推荐. 这些推荐不仅节省了用户浏览搜索的时间, 更为关键的是, 没有这些推荐, 有些信息用户根本就不可能找到.

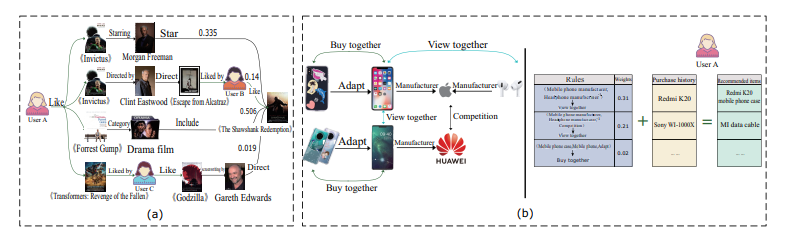


图6 基于知识图谱的可解释性推荐系统示例

推荐系统的技术发展道路还处于探索阶段, 随着深度学习的发展和崛起, 一些新的技术跟新迭代较快[7] [10] [11] , 如图6所示为一个知识图谱引用在推荐系统中, 除了上述讨论之外, 目前所有的推荐系统都面临着一些具有共性的问题. 这些问题的解决可以从根本 上极大地促进推荐系统的研究与应用.

# **参考文献**

* + 1. 黄立威,江碧涛,吕守业,刘艳博,李德毅.基于深度学习的推荐系统研究综述[J].计算机学报,2018,41(07):1619-1647.
    2. 朱郁筱,吕琳媛.推荐系统评价指标综述[J].电子科技大学学报,2012,41(02):163-175.
    3. 王国霞,刘贺平.个性化推荐系统综述[J].计算机工程与应用,2012,48(07):66-76.
    4. 马宏伟,张光卫,李鹏.协同过滤推荐算法综述[J].小型微型计算机系统,2009,30(07):1282-1288.
    5. 许海玲,吴潇,李晓东,阎保平.互联网推荐系统比较研究[J].软件学报,2009,20(02):350-362.
    6. 刘建国,周涛,汪秉宏.个性化推荐系统的研究进展[J].自然科学进展,2009,19(01):1-15.
    7. 张光卫,李德毅,李鹏,康建初,陈桂生.基于云模型的协同过滤推荐算法[J].软件学报,2007(10):2403-2411.
    8. 邢春晓,高凤荣,战思南,周立柱.适应用户兴趣变化的协同过滤推荐算法[J].计算机研究与发展,2007(02):296-301.
    9. 孙小华. 协同过滤系统的稀疏性与冷启动问题研究[D].浙江大学,2005.
    10. 邓爱林,左子叶,朱扬勇.基于项目聚类的协同过滤推荐算法[J].小型微型计算机系统,2004(09):1665-1670.
    11. 赵亮,胡乃静,张守志.个性化推荐算法设计[J].计算机研究与发展,2002(08):986-991.