关于电影推荐的 UserCF 算法实证分析

摘要

本文主要采用基于用户的协同过滤算法，通过计算某用户与其他用户的相似度，选取最相似的几个用户后，将这些用户选择的商品推荐给其他用户。通过分析用户的电影评价记录，采用协同过滤算法，预测用户最有可能评价的五部电影。

关键词：协同过滤 UserCF

一、研究内容

分析用户的电影评价记录，着重研究隐反馈数据集中 TopN 推荐问题，即预测用户会不会对某电影评分，给出每个用户最有可能评价的五部电影并对模型经行检评测。

二、研究方法

（一）原理

UserCF 算法是指基于用户的协同过滤算法，核心思想是先计算某用户与其它用户的相似度，然后选取与该用户最相似的几个用户，将这几个用户“买过”的产品推荐给该用户。

（二）选择原因

虽然从多样性、冷启动、电影数量稳定性较强的角度考虑，基于物品的协同过滤算法（ItemCF）更优。然而本实证分析中的物品（电影）数量远超于用户数量，就本数据而言采用 UserCF 更加高效。

三、解释数据

本文采用 GroupLens 提供的 MovieLens 数据集，使用小型 MovieLens 最新数据集，包含 610 个用户对 9724 部电影的 100800 条评分。该数据集是一个评分数据集，用户给电影的评分在 0~5 范围内。本文着重研究隐反馈数据集中 TopN 推荐问题，因此忽略了数据集中的评分记录。

四、建模过程

（一）利用余弦相似度对两两用户计算相似度

1、建立电影-用户倒排表

图 1 左半部分为训练数据格式，A、B、C、D 对应用户，a、b、c 等对应用户评价过的电影。右半部分电影-用户倒排表，如对于电影 a，喜欢它的有用户 A 和 B。

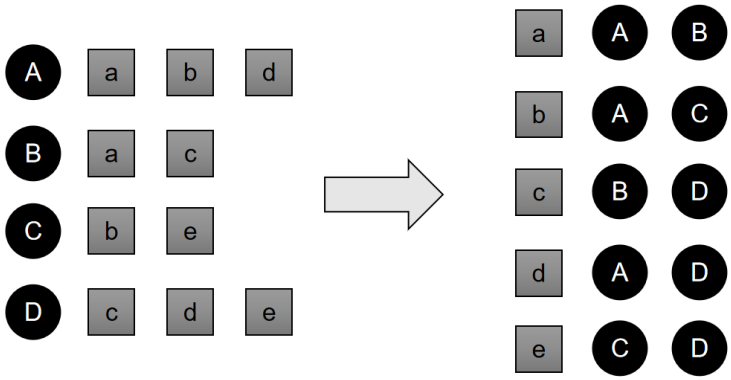

图 1 电影-用户倒排表原理图

2、建立用户相似度矩阵

利用电影-用户倒排表，构建用户相似度矩阵，如图 2，其中的值，如 $matrix[A][B]$ 表示用户 A 和用户 B 共同喜欢的电影的数量。

	A	B	C	D
A	0	1	1	1
B	1	0	0	1
C	1	0	0	1
D	1	1	1	0

图 2 相似度矩阵

3、计算用户相似度

遍历用户相似度矩阵中所有的两两用户，根据两两用户共同喜欢的电影的数量，计算用户相似度。

可以用余弦公式度量用户相似度：

$$w_{uv} = \frac{|N(u) \cap N(v)|}{\sqrt{|N(u)||N(v)|}} \quad (1)$$

其中，w_{uv} 表示用户 u 与 v 的相似度，作为 $matrix[u][v]$ 的值，$N(u)$ 表示用户

u 曾评价过的电影集合，$N(v)$ 表示用户 v 曾评价过的电影集合。

由于该公式惩罚了用户 u 和 v 共同喜欢的电影中热门电影对他们相似度的影响，换句话说，两个用户对冷门电影采取过同样的行为更能说明他们兴趣的相似度。

改进后的相似度计算公式如下：

$$w_{uv}=\frac{\sum_{i\in N(u)\cap N(v)}\frac{1}{\log 1+|N(i)|}}{\sqrt{|N(u)||N(v)|}} \quad (2)$$

其中，i 表示用户 u 和用户 v 都评价过的电影集合，N(i)评价过电影 i 的用户数，即分子部分表示 “用户 u 和 v 均评价过的电影数”。

（二）针对目标用户 u，找到其最相似的 K 个用户，产生 N 个推荐

K 表示与用户 u 兴趣相似的用户个数，N 表示为用户 u 推荐的电影数。首先，对用户 u，在用户相似度中找到与其相似度最高的 K 个用户。利用如下的公式计算用户 u 对电影 i 的感兴趣程度 p(u, i)：

$$p(u,i)=\sum_{v\in S(u,K)\cap N(i)} w_{uv}r_{vi} \quad (3)$$

其中，S(u, k)包含和用户 u 兴趣最接近的 K 个用户，N(i)是对电影 i 有过行为的用户集合，w_{uv} 是用户 u 和用户 v 的相似度，rvi 表示用户 v 对 i 的兴趣，这里使用的是单一行为的隐反馈，即用户 v 对电影 i 有过评价时 rvi=1。

（三）评测指标

将用户行为数据按照均匀分布随机划分为 M 份(如取 M=8)，挑选一份作为测试集，将剩下的 M-1 份作为训练集。为防止评测指标不是过拟合的结果，共进行 M 次实验，每次都使用不同的测试集。然后将 M 次实验测出的评测指标的平均值作为最终的评测指标。

1、召回率

在测试集上，对用户 u 推荐 N 个电影(记为 R(u))，用户实际评价的电影集合为 T(u)，召回率描述有多少比例的推荐电影包含在实际的用户评价列表中。

$$\mathrm{Recall}=\frac{\sum_{u}|R(u)\cap T(u)|}{\sum_{u}|T(u)|} \quad (4)$$

2、准确率

准确率描述最终的推荐列表中有多少比例是发生过的用户-电影评分记录。

$$\Pr ecision = \frac{\sum_{u} |R(u) \cap T(u)|}{\sum_{u} |R(u)|} \qquad (5)$$

3、覆盖率

覆盖率反映了推荐算法发掘长尾的能力，覆盖率越高，说明推荐算法越能够将长尾中的电影推荐给用户。分子部分表示实验中所有被推荐给用户的电影数目(集合去重)，分母表示数据集中所有电影的数目。

$$Coverage = \frac{\left| \bigcup_{u \in U} R(u) \right|}{|I|} \qquad (6)$$

五、结论总结

在初始值 K=80，N=5 的情况下，即针对目标用户，找到其最相似的 80 个用户，产生 5 个最终的推荐电影，部分结果如附录一所示。对模型进行评测得准确率=0.2370、召回率=0.0573、覆盖率=0.0133。因此，该模型最终的推荐列表中有 23.7%是发生过的用户-电影评分记录，有 5.73%的推荐电影包含在实际的用户评价列表中，推荐电影占所有电影的比例为 1.33%。

参考文献

[1] 推荐系统与深度学习[M]. 黄昕等. 清华大学出版社. 2019.
[2] 推荐系统算法实践[M]. 黄美灵. 电子工业出版社. 2019.
[3] 推荐系统算法[M]. 项亮. 人民邮电出版社. 2012.
[4] 美团机器学习实践[M]. 美团算法团队. 人民邮电出版社. 2018.

附录一

userId

1	Godfather, The (1972)	Indiana Jones and the Last Crusade (1989)	Terminator 2: Judgment Day (1991)	Die Hard (1988)	Fargo (1996)
2	Matrix, The (1999)	Dark Knight, The (2008)	Fight Club (1999)	Forrest Gump (1994)	Lord of the Rings: The Return of the King, The...
3	Star Wars: Episode IV - A New Hope (1977)	Star Wars: Episode VI - Return of the Jedi (1983)	Star Wars: Episode V - The Empire Strikes Back...	Forrest Gump (1994)	Pulp Fiction (1994)
4	Godfather, The (1972)	Forrest Gump (1994)	Shawshank Redemption, The (1994)	Schindler's List (1993)	Usual Suspects, The (1995)
5	Forrest Gump (1994)	Jurassic Park (1993)	Speed (1994)	Crimson Tide (1995)	Silence of the Lambs, The (1991)
...	...	...	...	...	...
606	Fight Club (1999)	Fargo (1996)	Groundhog Day (1993)	Lord of the Rings: The Fellowship of the Ring,...	Goodfellas (1990)
607	Raiders of the Lost Ark (Indiana Jones and the...	Forrest Gump (1994)	Aliens (1986)	Terminator 2: Judgment Day (1991)	Twelve Monkeys (a.k.a. 12 Monkeys) (1995)
608	Seven (a.k.a. Se7en) (1995)	Raiders of the Lost Ark (Indiana Jones and the...	Fight Club (1999)	Princess Bride, The (1987)	Shawshank Redemption, The (1994)
609	True Lies (1994)	Shawshank Redemption, The (1994)	Clear and Present Danger (1994)	Silence of the Lambs, The (1991)	Stargate (1994)
610	Matrix, The	Star Wars:	Good Will	Incredibles,	Monsters, Inc.

userId

(1999)	Episode IV - A New Hope (1977)	Hunting (1997)	The (2004)	(2001)