DOI:10.19651/j.cnki.emt.1902677

基于项目类型的群组推荐方法

宣鹏程 唐 彦 王汪送
(河海大学 计算机与信息学院 南京 211100)

摘 要:近年来,群组推荐吸引了大量研究人员的关注。针对群组推荐中融合策略的不足而导致群组推荐结果准确率低的问题,提出了一种新的基于项目类型的方法来改进偏好融合策略,以此提高推荐结果的准确性。通过引入项目类型占比因子并计算群组类型偏好和用户类型偏好之间的类型相似性,同时提出了评分融合公式来预测群组的最终项目分数,从而改进偏好融合策略。最后,在 Movielens 数据集上进行实验并将本文方法与几种经典的群组推荐方法进行比较。结果表明,本文方法比传统基线方法具有更高的准确率。

关键词:群组推荐;协同过滤;项目类型占比因子;偏好融合策略
中图分类号:TP393;TN941.3 **文献标识码**:A **国家标准学科分类代码**:520.6099

Group recommendation method based on Item Type

Xuan Pengcheng Tang Yan Wang Wangsong
(College of Computer and Information, Hohai University, Nanjing 211100, China)

Abstract: Group recommendation has attracted significant research attention in recent years. In view of the problem of insufficiency of fusion strategy in group recommendation which cause the low accuracy, we propose an novel method to improve preference fusion strategy to increaandse the accuracy. We introduce the concept of item type proportion factor and calculate the type similarity between group type preference and user type preference. Meanwhile we design a score fusion formula to predict item score for the group. Finally, we carry out experiments and compare our method with several classic group recommendation methods using Movielens dataset. The results show that our method achieves higher recommendation accuracy than all the baseline methods.

Keywords: group recommendation; collaborative filtering; item type proportion factor; preference fusion strategy

0 引 言

随着互联网技术的飞速发展,网上的海量信息越来越多,信息过载问题[1]也日益突出,传统的搜索引擎已经很难为用户提供个性化的偏好信息。因此,推荐系统[2-3]应运而生。推荐系统可以帮助用户从大量数据中发现他们可能感兴趣的项目,解决用户面对海量数据选择迷茫的问题。然而,实际生活中很多活动是由多个用户以群组的形式进行的,如旅游、看电影等。换句话说,推荐系统还需要考虑群组中每个用户的偏好做推荐,通常这种形式的推荐系统被称为群组推荐系统[4-5]。它将推荐的对象由单一用户拓展到一个用户组成的组,这也给推荐系统带来了新的问题。其中最为重要的就是如何获取群组成员的共同偏好,缓解群组成员之间的偏好冲突,使群组成员都比较满意[6]。

目前,关于群组推荐的研究主要从两个方面进行。一个方面是在领域专家的引导下通过小组讨论获得群组的共同偏好;另一方面是通过分析群组用户的历史评分数据来预测未被群组评估的项目的评分。该方法通常基于所有用户的评分,但是存在以下问题:两个完全不同类型的项目可能具有高相似性,并且群组兴趣偏好通常仅仅限于一个或多个特定的字段。这可能会带来了许多与群组偏好无关的推荐结果,因而在一定程度上降低群组推荐的准确性。文献[7]对融合方法进行了汇总,总结出群组推荐领域中常用的融合方法有评分融合、推荐融合以及群组的偏好建模。

针对以上情况,本文从提高群组推荐的准确率入手,引入项目类型占比因子的概念来改进偏好融合策略,给群组做出推荐,并在真实数据集上验证了该方法的有效性。

1 基于项目类型的群组推荐方法

1.1 相关定义

定义 1:用户集合 $\boldsymbol{U}=\{\boldsymbol{U}_1,\boldsymbol{U}_2,\cdots,\boldsymbol{U}_i\}$,项目集合 $\boldsymbol{I}=\{\boldsymbol{I}_1,\boldsymbol{I}_2,\cdots,\boldsymbol{I}_n\}$,物品类型集合 $\boldsymbol{T}=\{T_1,T_2,\cdots,T_t\}$。

收稿日期:2019-02-28

定义 2:(用户项目评分矩阵 $\boldsymbol{S}_{m\times n}$)$m$ 代表用户,n 代表项目,$\boldsymbol{S}_{mn}$ 代表用户 m 对项目 n 的评分值。

定义 3:(项目类型关系矩阵 $\boldsymbol{R}_{n\times t}$) n 代表项目,t 代表类型,如果项目 n 包含类型 t,那么 $R_{nt}=1$,否则 $R_{nt}=0$。

1.2 项目类型占比因子的概念

由于项目类型远远少于项目数量,而仅仅利用项目的评分信息来做推荐是非常单一的,这种方式通常会将项目中隐含的其他信息忽略掉。因此,本文利用项目中蕴含的丰富的类型信息来辅助群组做推荐。提出了项目类型占比因子的概念,利用它更好地反映团队成员在该领域的共同偏好,同时提高团队成员对推荐项目的满意度。项目类型比例因子(proportion factor based on item type,ITPF)表示用户 i 在拥有 t 类型的项目上的行为记录占该用户所有行为项目数的比例,形式化表示如式(1)所示:

$$ITPF(i,t)=\frac{\sum_{n=1}^{|N_i|}R_{nt}}{|N_i|} \tag{1}$$

式中:$|N_i|$ 表示已由用户 i 评分过的项目集合。

1.3 基于项目类型的群组推荐方法流程

不同于以往的群组推荐主要针对用户的评分信息进行相关处理进而给用户做推荐,本文针对项目的类型提出了新的群组推荐方法。通过融入项目的类型信息构建群组偏好,然后为群组推荐喜爱的物品。本文首先对项目类型数据进行建模,分别挖掘出用户和群组的类型偏好信息,然后在新的预测评分公式基础上按 Top-N 给用户做出推荐。最后,从准确率和召回率两方面验证所提方法的可行性。基于项目类型的群组推荐方法的流程如图 1 所示。

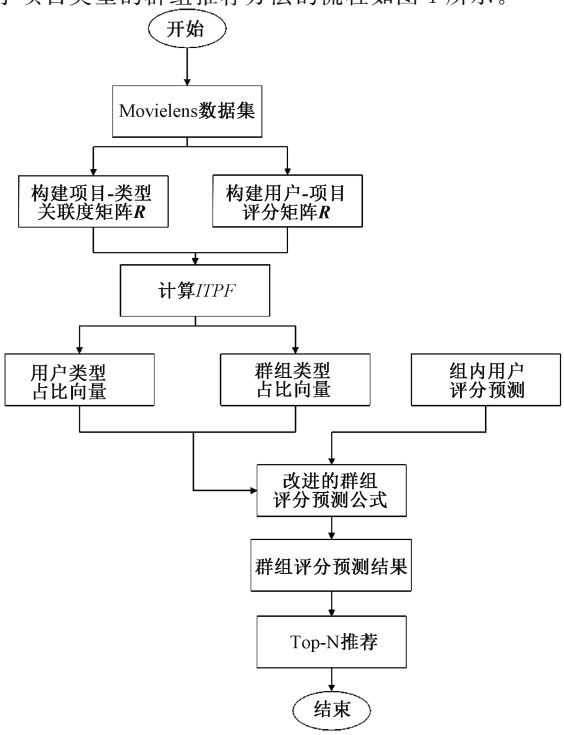

图 1 基于项目类型的群组推荐方法流程

1.4 方法描述

1)数据预处理

首先根据一定的规则将文本表示的项目类型信息转换为相关的项目类型向量,得到用户项目评分矩阵和项目类型关系矩阵,以便于后续计算。

2)群组用户的个性化偏好预测

群组推荐是在个性化推荐的基础上进行的。首先获取个人偏好并采用预测方法计算用户的个性化评分,然后在这一步的基础上融合用户评分。首先采用传统的基于物品的协同过滤算法[8-9]预测用户对某项目的评分。预测的评分如式(2)所示:

$$p\_rate(i,n)=\bar{r}_i+\frac{\sum_{j\in N_{NEI}}sim(i,j)\times(r_{jn}-\bar{r}_j)}{\sum_{j\in N_{NEI}}sim(i,j)} \tag{2}$$

式中:$p\_rate(i,n)$ 表示用户 i 对项目 n 的预测评分;$sim(i,j)$ 表示物品 i 和物品 j 的相似度;$\bar{r}_i$ 表示用户 i 对评分的所有项目的平均评分值;N_{NEI} 表示和物品 n 相似的物品集合。

3)偏好融合

在得到了用户的个性化评分之后,就可以通过一定的融合策略融合这些分数得到最终的群组评分。在通常的应用场景中,群组推荐系统需要满足群组用户的满意度、组内公平性、用户可理解性等要求。因此,不同的融合策略被提出。文献[10]讨论了以下几种典型的群组偏好融合策略:公平策略、最受尊敬者策略、均值策略、痛苦避免均值策略、最小痛苦策略、最开心策略等。通常情况下,需要根据特定的群组类型特征使用不同的融合策略,然而具体使用哪种融合策略能够达到最佳的推荐效果,仍然是群组推荐系统研究领域中的热点问题之一[11]。

文献[12]将传统的模型融合与推荐融合相组合,结合两种方法的优点,给群组做推荐。文献[13]详细分析了几种偏好融合策略。均值策略选择组成员的平均得分作为群组的预测评分。但是,通常会导致群组中的某些成员非常不满意,因而会导致推荐结果的准确性和满意度降低。为避免这个问题,最小痛苦策略选择群组中成员评分最低的项目频分作为群组的预测分数,但是由于群组评分取决于特定的用户,群组偏好容易受到恶意的篡改。加权模型根据群组成员的个性特征,扮演的角色,在群组中的影响度和其他因素为每个群组成员分配不同的权重。总之,常见的融合策略有其缺点和优点,针对什么类型的群组应该选择何种融合策略,依然是个有趣的研究热点。本文引入项目类型占比因子的概念,并利用加权的方法计算项目的预测评分,改进偏好融合策略,进而做出推荐,具体步骤如下:

(1)用户项目类型偏好的获取。在这一步计算群组 g 关于群组所有用户的项目类型占比因子向量。在这之前,已经得到了物品评分矩阵 $\boldsymbol{S}_{m\times n}$,项目类型关系矩阵 $\boldsymbol{R}_{n\times t}$,

然后给用户 i 的项目类型占比因子向量如式(3)所示：

$$\boldsymbol{type(i)} = \{ITPF(i,1), ITPF(i,2), \cdots, ITPF(i,t)\} \tag{3}$$

其中 i 为群组 g 中的某一用户。

(2)群组项目类型偏好的获取。同样地，可以得到群组 g 的项目类型占比因子向量，如式(4)所示：

$$\boldsymbol{type(g)} = \{ITPF(g,1), ITPF(g,2), \cdots, ITPF(g,t)\} \tag{4}$$

(3)项目的群组预测评分。根据每个用户与群组之间的类型偏好相似性，在进行组偏好融合时提出如下公式计算群组 g 给项目 n 的预测评分，如式(5)所示：

$$g\_rate(g,n) = \frac{\sum_{i \in G(g)} p\_rate(i,n) \cdot sim(type(i), type(g))}{\sum sim(type(i), type(g))} \tag{5}$$

其中 $sim(type(i), type(g))$ 表示用户 i 与群组 g 之间的类型相似性。

4)推荐生成

通过上述步骤，可以得到目标群组 g 关于项目 n 的预测评分 $g\_rate(g,n)$，然后，将前 N 项预测分数最高的项目返回给群组 g。至此，推荐流程完成。

2　实验结果与分析

2.1　数据集介绍

为了验证本文提出的方法的有效性，本文采用 Movielens 100k[14]数据集来实施实验。MovieLens 是一个电影评分数据库，它由 Minnesota 大学的 GroupLens 研究小组创建并维护。Movielens 100k 数据集包括了用户对电影的评分信息，评分区间为 1 到 5，同时它包括来自 943 个用户对 1 682 部电影的 100 000 条评分数据。

2.2　群组生成方式

由于在真实的数据集中很难找到大量的群组数据，所以引用文献[15]记录的方法，从传统的推荐系统的数据集中构造群组[15-16]。由于组内相似度对推荐结果具有一定的影响，在分组时考虑两个因素：组内相似度和组规模。根据组内的用户相似度，本文生成高相似度群组和低相似度群组两种类型。组内相似度用 Pearson 相关来计算用户之间的相似性。在组规模的问题上，分别考虑包含 2 个、6 个、10 个、14 个和 18 个用户的组。

2.3　基线方法

针对本文提出的方法，对比方法为：

1)LM-CF(Least Misery-Collaborative filtering)[16]；

2)Avg-CF(Average-Collaborative filtering)[17]；

3)GRIP-CF(Group Recommendation Method Based on Item Type)，即本文提出的方法。

本文的群组推荐的方法为评分融合[16]，即先得到群组用户的个性化评分，再通过选定的融合策略融合成群组的评分。采用的个性化评分预测算法为基于物品的协同过滤算法。使用到的融合策略有均值策略和最小痛苦策略。

2.4　评价指标

本文的融合方法为评分融合，因此选择两个指标，分别是准确率和召回率作为算法评价标准。

第 1 个指标是 $Precision$，如式(6)所示：

$$Precision@N = \frac{\sum_i |R(i,N) \cap T(i,N)|}{\sum_i |R(i,N)|} \tag{6}$$

式中：N 是推荐列表的长度；$R(i,N)$ 表示推荐列表 N 中用户 i 喜欢的物品集合；$T(i,N)$ 表示测试集中用户喜欢的物品集合。

第 2 个指标是 $Recall$，如式(7)所示：

$$Recall@N = \frac{\sum_i |R(i,N) \cap T(i,N)|}{\sum_i |T(i,N)|} \tag{7}$$

2.5　实验结果分析

在本文实验中，训练集是总样本的 80%，测试集是剩余的 20%。在实验中使用 5 种交叉验证方法。实验结果取 5 次测试的平均值。定义了两种类型的群组：低相似度群组(组内用户相似度大于 0.3 或等于 0.3)，高相似度群组(组内用户相似度介于 0.3 和 0.5 之间)。此外，实验从两个方面进行：一个是组规模 G，另一个是推荐列表长度 N。

1)组规模 G 对推荐结果的影响

组规模是指群组中的用户数量。文献[11]中提到，群组规模会影响推荐质量[18]，同时它在群组决策中发挥着重要作用。因此，研究了不同群组规模 G 对推荐结果的准确率的影响，推荐列表长度 N 取值为 10。实验结果如图 2 和 3 所示。

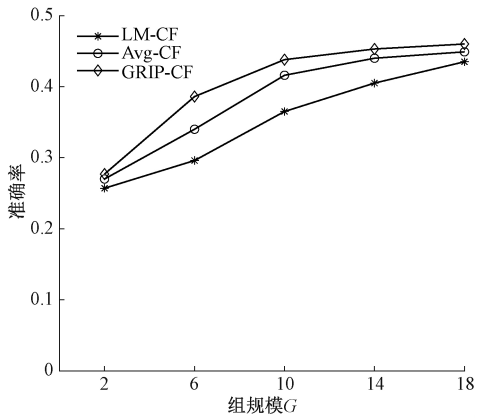

图 2　不同组规模 G 下的准确率实验结果(高相似度群组)

从图 2 和图 3 中可以看出，群组的用户相似度越高，代表群组的偏好越相似，因而推荐结果的准确率也相对比较高。同时，当群组规模 G 为 10 时，推荐效果最好。综上，

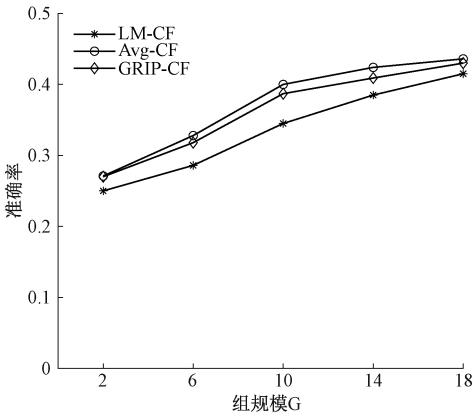

图 3 不同组规模 G 下的准确率实验结果(低相似度群组)

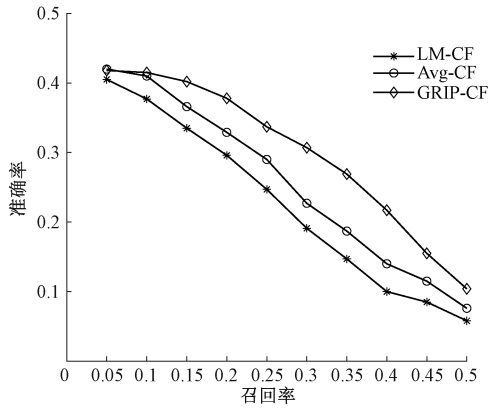

图 5 *Precision-Recall* 曲线(低相似性组,$G=10$)

与其他两种方法相比,本文方法在不同组规模 G 下准确率都优于其他的方法。

2)不同的推荐列表长度 N

本文使用 *Recall-Precision* 曲线来综合显示 3 种方法的性能好坏。简单地说,*Recall-Precision* 曲线是精确度和召回率曲线。它将 *recall* 作为横坐标轴,将 *precision* 作为纵坐标轴。图 4 和 5 所示展示了提出的方法和传统两种方法在推荐列表长度 N 变化时的性能比较。此外,组规模 G 固定为 10,推荐列表长度 N 从左至右为(5,10,15,20,25,30,35,40,45,50)。

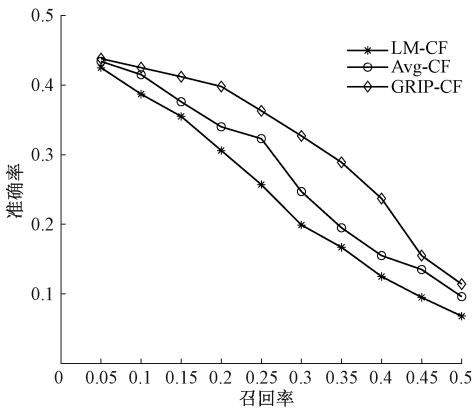

图 4 *Precision-Recall* 曲线(高相似性组,$G=10$)

从图 4 和 5 中可以看出,准确率与召回率之间存在着一些矛盾。随着推荐列表长度 N 的增加,召回率逐渐变高,准确率变低。当召回率低时,准确率比较高。这是因为推荐列表 N 在不断变化。很明显,选择合适的 N 来协调准确率和召回率,才能获得更好的推荐结果。综合两个实验结果,当 $N=10$ 时,本文方法的效果是非常好的,提高了推荐的准确率。

通过以上分析可得,融入项目类型比例因子之后,本文方法在准确率和召回率方面都优于其他两种推荐方法。

3 结 论

本文针对群组推荐领域存在的准确率低的问题,引入了项目类型占比因子的概念,通过计算群组类型偏好和用户类型偏好之间的相似性来改进偏好融合策略,提高推荐的准确率和群组成员的满意度。同时,借助公开的 Movielens 数据集,将本文方法与其他两种经典的传统方法进行比较,实验结果验证了本文提出方法的有效性,而且该方法在推荐准确率和召回率方面优于其他两种基线方法。虽然方法优于其他两种方法,但是也只是考虑了项目的类型这种显性的信息,在未来的研究中,计划将知识图谱中包含的丰富的语义信息融入群组推荐,以达到更好的推荐结果。

参考文献

[1] XU H L, WU X, LI X D, et al. Comparison study of internet recommendation system [J]. Journal of Software, 2009, 20(2):350-362.

[2] RESNICK P, VARIAN H R. Recommender systems [J]. Communications of the Acm, 1997, 40(3):56-58.

[3] 许海玲, 吴潇, 李晓东, 等. 互联网推荐系统比较研究[J]. 软件学报, 2009, 20(2):350-362.

[4] GARCIA I, SEBASTIA L, ONAINDIA E. On the design of individual and group recommender systems for tourism [J]. Expert Systems with Applications, 2011, 38:7683-7692.

[5] 张玉洁,杜雨露,孟祥武.组推荐系统及其应用研究[J].计算机学报,2016,39(4):745-764.

[6] JAMESON A. More than the sum of its members: Challenges for group recommender systems [C]. Working Conference on Advanced Visual Interfaces, ACM, 2004.

[7] JAMESON A, SMYTH B. Recommendation to Groups[M]. The Adaptive Web,2007.

[8] LARA Q S, JUAN A, BELEN D A. Personality and

social trust in group recommendations [C]. IEEE International Conference on Tools with Artificial Intelligence. IEEE Computer Society, 2010.

[9] 陈琦，吕杰，张世超. 一个解决协同过滤推荐系统相关问题的新算法[J]. 电子测量技术，2016，39(5)：66-69.

[10] RICCI F, ROKACH L, SHAPIRA B, et al. Recommender Systems Handbook [M]. Springer US, 2011.

[11] ZHANG Y J, DU Y L, MENG X W. Research on group recommender systems and their applications[J]. Chinese Journal of Computers, 2016, 39(4): 745-764.

[12] 胡川，孟祥武，张玉洁，等. 一种改进的偏好融合群组推荐方法[J]. 软件学报，2018，29(10)：3164-3183.

[13] MASTHOFF J. Group modeling: Selecting a sequence of television items to suit a group of viewer[C]]. User Model, User-Adapt, Interact, 2004.

[14] HARPER F M, KONSTAN J A. The MovieLens Datasets: History and Context[M]. ACM, 2015.

[15] YUAN Q, CONG G, LIN C Y. Com: a generative model for group recommendation [J]. 2014 (8): 163-172.

[16] BALTRUNAS L, MAKCINSKAS T, RICCI F. Group recommendations with rank aggregation and collaborative filtering [C]. ACM Conference on Recommender Systems, 2010.

[17] AMER-YAHIA S, ROY S B, CHAWLAT A, et al. Group recommendation: Semantics and efficiency[J]. Proceedings of the VLDB Endowment, 2009, 2(1): 754-765.

[18] TOON D P, SIMON D, LUC M. Comparison of group recommendation algorithms [J]. Multimedia Tools and Applications, 2014, 72(3):2497-2541.

作者简介

宣鹏程，硕士研究生，主要研究方向为推荐系统。
E-mail：pcxuan.hhu@gmail.com

唐彦，博士、副教授、硕士生导师，主要研究方向为大数据挖掘、分布计算与并行处理、协同计算、水信息学。

王汪送，硕士研究生，主要研究方向为推荐系统。