推荐系统中典型用户群组的发现和应用*

谭 昶[1] 刘 淇[1] 吴 乐[1] 马海平[2] 龙 柏[3]

[1](中国科学技术大学 计算机科学与技术学院 合肥 230027)
[2](科大讯飞股份有限公司 合肥 230088)
[3](中国电子科技集团公司第三十八研究所 合肥 230088)

摘 要 推荐系统是解决用户的个性化信息需求的一种有效工具. 但随着推荐系统用户规模的扩大,需要合理地从海量用户中筛选出用户子集,并进行持续和深入的分析以改进推荐系统. 因此,文中首先提出典型用户群组的概念,以期发现推荐系统中的典型用户子集,从而可正确地反映全体用户的兴趣偏好. 随后提出一种典型用户群组的发现算法,通过比较候选新增典型用户对典型用户群组的贡献度,逐一扩大典型用户群组规模,最终达到较高的推荐项目覆盖率和评分准确度. 最后在典型用户群组中寻找用户的最近邻,实现一种改进的协同过滤推荐算法. 通过在真实数据集上的实验结果表明,与其他用户群组发现算法以及经典推荐算法相比,验证典型用户群组不仅具有较好的代表性,也能够获得更好的推荐效果.

关键词 推荐系统,典型用户,覆盖率
中图法分类号 TP 391 **DOI** 10.16451/j.cnki.issn1003-6059.201505010

Finding and Applying Typical User Group in Recommender Systems

TAN Chang[1], LIU Qi[1], WU Le[1], MA Hai-Ping[2], LONG Bo[3]
[1](*School of Computer Science and Technology, University of Science and Technology of China, Hefei* 230027)
[2](*iFLYTEK Co., Ltd., Hefei* 230088)
[3](*China Electronic Technology Group Corporation No.* 38 *Research Institute, Hefei* 230088)

ABSTRACT

Recommender system (RS) provides an effective way to solve the personalized information needs of users. However, with the expansion of the user scale, it is necessary to find some subsets of vast amounts of RS users, and the continuous and in-depth analysis for these user subsets can be used to improve the RS. Therefore, the typical user group (TUG) is defined as a representative subset of the entire users in RS to correctly reflect the preferences of all the users. Then, a weighted typical user group finding algorithm (WTFA) is designed to compare the contributions of the candidate typical users and choose the typical users with higher contribution, so that a TUG is built with high item coverage rate and rating

* 国家 863 计划项目(No. 2014AA015203)、中央高校基本科研基金项目(No. WK0110000042)、安徽省自然科学基金青年基金项目(No. 1408085QF110)、安徽省科技专项(No. 13Z02008-5)、安徽省国际科技合作计划项目(No. 1303063008)、安徽省科技攻关计划项目(No. 1301022064)资助
收稿日期:2014-03-25;修回日期:2014-05-29
作者简介 谭昶,男,1986 年生,博士研究生,主要研究方向为个性化推荐系统. E-mail: tanchang@mail.ustc.edu.cn. 刘淇(通讯作者),男,1986 年生,博士,副研究员,主要研究方向为数据挖掘与知识发现、机器学习方法及其应用. E-mail: qiliuql@ustc.edu.cn. 吴乐,女,1988 年生,博士研究生,主要研究方向为个性化推荐系统、社交网络分析. 马海平,女,1986 年生,博士,工程师,主要研究方向为文本挖掘和移动情境挖掘. 龙柏,男,1980 年生,博士,副研究员,主要研究方向为雷达信息处理.

accuracy. A modified TUG-based collaborative filtering(TUG-CF) algorithm is developed to discover the nearest neighbors in TUG. The experimental results on real world dataset show that TUG is better than most rating user group and maximizes diversified user group on item coverage rate and rating accuracy, and TUG-CF has better recommendation results than traditional collaborative filtering methods.

Key Words　Recommender System, Typical User, Coverage Rate

1　引　言

随着信息社会的发展和计算机技术的进步,推荐系统已成为人们避免信息过载以及获得有效信息的重要工具[1]. 推荐系统通过协同过滤、基于内容的推荐等方法,可有效提供与目标用户的兴趣偏好具有较高相关性的个性化推荐结果,从而提升用户对于信息服务的满意度.

推荐系统的相关研究大多集中在如何提升推荐效果方面,而对推荐系统用户群体的深入分析较少. 应当注意到,随着推荐系统用户规模的不断扩大,逐一分析单个用户较为浪费计算资源和时间. 通常在推荐系统的实际应用中,运用用户聚类[2-4]、删除不活跃用户[1,5-6]等手段,降低待分析用户规模,然后进一步实现后续推荐流程. 文献[5]和文献[6]对于旅游推荐中的不活跃用户进行删除后,综合运用多种用户特征提升推荐效果,但由于类似的方法处理过程存在固有缺陷,在保证精度的同时并不能保证充分反映用户兴趣. 因此,有必要实现新的用户子集筛选方法,使得筛选得到的子集既能充分反映全体用户的兴趣偏好,又能尽可能多地覆盖和包含已知推荐项目.

针对这一问题,本文提出典型用户群组(Typical User Group, TUG)的概念,并且考虑到冷门项目的存在,给出一种加权的典型用户群组的发现算法(Weighted Typical User Group Finding Algorithm, WTFA),期望在保证较高的推荐项目覆盖率和较为准确的评分基础上,筛选得到全体用户的一个最优子集. 在真实数据集上的实验结果表明,WTFA 找出的典型用户群组与其它方法相比具有更高的项目覆盖率和较小的评分误差,因而更具有典型性.

为将典型用户群组应用于推荐过程中,本文提出一种基于典型用户群组的协同过滤推荐算法(TUG-Based Collaborative Filtering Recommendation Algorithm, TUG-CF). TUG-CF 与经典协同过滤算法的最大不同在于,在寻找用户的最近邻时,不在全体用户中搜索,仅仅在典型用户群组中搜索. 通过这种方式,既降低了最近邻搜索的时间和计算耗费,又保证了找到最近邻的有效性. TUG-CF 与其它推荐算法的对比也表明,合理地使用典型用户群组可有效优化现有的协同过滤推荐算法.

本文首先介绍相关工作,然后给出典型用户群组以及相关的项目覆盖率的定义和相关性质,随后给出典型用户群组发现算法(WTFA)的实现,用于发现典型用户群组,同时给出基于典型用户群组的协同过滤推荐算法 TUG-CF 的实现,最后通过在真实数据集上的实验验证典型用户群组的代表性以及 TUG-CF 对于推荐效果优化的有效性.

2　相关工作

本文的相关研究工作主要包括推荐系统中用户的处理和筛选,推荐系统中项目覆盖率的提高以及推荐结果的多样性问题(对不同类别项目的推荐情况).

大部分推荐系统相关的研究工作都会对原始数据的用户集合进行必要的筛选和处理,典型的处理方法包括用户聚类[2-4]、删除不活跃用户[1,5-6]等. 在用户数量较大时,将相似用户聚类并对聚类簇进行推荐,可有效减少计算量并保证精度[2-4]. 文献[4]考察多种聚类方法并指出恰当的聚类可以避免数据稀疏性,并且有效提高精确度. 文献[3]运用聚类方法提出可缩放的最近邻范围,使得协同过滤算法可适应大规模数据集. 文献[2]利用遗传算法优化用户聚类时初始点的选择. 在文献[5]和文献[6]中,对于旅游推荐中的不活跃用户进行删除后,综合运用多种用户特征提升推荐效果. 相关工作的缺陷是聚类过程使部分不活跃的用户或项目淹没在较大的类别中,无法得到合适的推荐. 删除不活跃用户等方法也存在类似的问题.

近年来,研究者也注意到项目覆盖率的问题,并提出许多有价值的方法. 项目覆盖率的研究集中在文档摘要领域,即如何使用一个较低开销的文档子集包含(覆盖)尽可能多的有用信息[7-8]. 由于新闻和一般文档的相似性,这一概念被扩展并应用到新

闻推荐领域[9-11]. 文献[12]进一步指出在电子商务网站中提高推荐覆盖率能够有效提升购买效果. 相关工作为本文研究提供了有益参考,但还没有研究由项目覆盖率寻求用户的最优子集.

推荐结果的多样性目前被认为是和推荐结果的准确性同等重要的推荐系统评价指标[13]. 文献[14]系统地整理和总结推荐结果多样性的概念并指出如何验证多样性. 文献[15]在协同过滤中使用最远邻居并取得较好多样性的推荐结果. 文献[16]将社会网络的信任机制加入到推荐多样性的实现中,通过选择主题多样性好的信任邻居平衡推荐结果的准确性和多样性. 文献[17]在图像标签的推荐中,利用视觉距离模型很好地平衡推荐准确性和多样性. 本文虽然没有直接考察推荐结果的多样性,但通过对电影流派覆盖率的实验和分析,验证了典型用户群组可覆盖更具有多样性的项目集合,为如何平衡推荐准确性和多样性提供一种新的思路.

3 基本原理和定义

3.1 推荐系统的项目覆盖率

推荐系统是一种利用用户和项目的内容信息以及用户项目之间的互动信息,向合适的用户推荐合适的项目的信息过滤系统. 一般地,定义用户集合

$$U = \{u_i\},\ 0 \leqslant i < |U|$$

和项目集合

$$P = \{p_j\},\ 0 \leqslant j < |P|,$$

用户项目之间的互动信息用评分矩阵 $\boldsymbol{R}$ 量化,则有

$$\boldsymbol{R} = \{r_{ij}\}_{|U|\times|P|},\ r_{ij} \geqslant 0,$$

$r_{ij} = 0$ 代表用户 u_i 和项目 p_j 没有互动,即 u_i 的活动(或兴趣)没有覆盖 p_j. 对于用户集合 U 给定的子集 U',U' 的项目覆盖集 $P_{U'}$ 定义为

$$P_{U'} = \{p_k\},\ \exists r_{ik} \neq 0,\ u_i \in U',$$

U' 的项目覆盖率即为 P 的子集 $P_{U'}$ 在 P 中所占的比例,即

$$Cov(U') = \frac{|P_{U'}|}{|P|} \times 100\%.$$

对于大小相同均为 k 的用户子集,可称其中项目覆盖率最大的为 k-最大覆盖子集,显然项目覆盖率有如下性质.

性质 1 项目覆盖率单调递增,即

$$0 \leqslant Cov(U_1) \leqslant Cov(U_1 \cup U_2) \leqslant 1.$$

3.2 递增项目覆盖率

定义子集 U_1 加入到子集 U' 的递增项目覆盖率 $ICov_{U'}(U_1)$ 为

$$ICov_{U'}(U_1) = Cov(U_1 \cup U') - Cov(U').$$

注意到两个不同子集的项目覆盖集可能有非空交集或者空交集,因此 $Cov(U_1) > Cov(U_2)$ 时,并不一定有 $ICov_{U'}(U_1) > ICov_{U'}(U_2)$.

令 $U_1 = \{u\}$,即每次选择一个用户 u 加入 U',有

$$ICov_{U'}(u) = \frac{|P_{u\cup U'}| - |P_{U'}|}{|P|} \times 100\%.$$

由于 P_u 和 $P_{U'}$ 中相同的项目会互相抵消,则有

$$ICov_{U'}(u) = \frac{\sum\limits_{p_j\in P} f(p_j \in P_u, p_j \notin P_{U'})}{|P|}, \tag{1}$$

其中,函数 $f(\ )$ 的定义和取值为

$$f(p_j \in P_u, p_j \notin P_{U'}) = \begin{cases} 1, & p_j \text{ 存在} \\ 0, & p_j \text{ 不存在} \end{cases}$$

由于用户对项目的选择存在长尾性[18-19],冷门项目极易被掩盖在大量热门项目中. 针对此问题,定义加权的递增项目覆盖率为

$$ICov_{U'}(u) = \frac{\sum\limits_{p_j\in P} w_j \times f(p_j \in P_u, p_j \notin P_{U'})}{|P|}, \tag{2}$$

其中,权重 w_j 可表示为

$$w_j = 1 - \frac{\lg(S_j)}{\lg(|U|)},$$

其中,S_j 表示项目 P_j 一共被 S_j 个用户选择过,显然项目被选择的越多,权重 w_j 越小,为较少出现的冷门项目赋予较高的权重,从而提高较少出现项目被覆盖的可能.

基于递增项目覆盖率的定义证明如下.

性质 2 项目覆盖率是一个子模函数.

证明 对于用户集合 A,B,$A \subseteq B \subseteq U$,有 $P_A \subseteq P_B$. 对于任意的单个用户 $u \notin B$ 且 $u \in U$,有

$$|P_{u\cap B}| = |P_u \cap P_B| \geqslant |P_u \cap P_A| = |P_{u\cap A}|.$$

因此,有如下推导成立:

$$\begin{aligned} ICov_A(u) &= \frac{|P_{u\cup A}| - |P_A|}{|P|} \\ &= \frac{|P_u \cup P_A| - |P_A|}{|P|} \\ &= \frac{|P_u| + |P_A| - |P_{u\cap A}| - |P_A|}{|P|} \\ &\geqslant \frac{|P_u| + |P_B| - |P_{u\cap B}| - |P_B|}{|P|} \\ &= ICov_B(u). \end{aligned}$$

随着用户子集大小增加,新加入用户贡献的递增覆盖率越来越小,即项目覆盖率具有子模性. 证毕.

文献[20]和文献[21]已证明,最大化一个子模函数是 NP 难的,且可运用贪婪算法寻求最大化子模函数的近似解. 因此,项目覆盖率 $Cov(U')$ 可用一个贪婪算法寻求最大化的近似解.

3.3　典型用户群组的定义

由于期望典型用户群组能够充分反映全体用户的兴趣偏好,用子集评分误差 $Err(U')$ 表示用户子集和全体用户之间的兴趣偏差,即

$$Err(U') = \sum_{p_j \in P} \frac{|avg(p_j, U') - avg(p_j, U)|}{avg(p_j, U)} \times 100\%,$$

其中,$avg(p_j, U')$ 表示项目 p_j 在用户子集 U' 中的平均分. 显然 $Err(U')$ 越小,子集 U' 越能充分反映全体用户的兴趣偏好. 但如果 $Cov(U')$ 较小,则子集 U' 不能充分覆盖全体项目.

因此,定义典型用户群组(Typical User Group, TUG)是用户集合 U 的一个子集,大小为 k. 在用户集合 U 所有大小为 k 的子集中,它同时满足如下两个条件:

1) $\min Err(TUG)$;

2) $\max Cov(TUG)$.

直接优化 $Err(TUG)$ 较复杂,并且使用协同过滤、矩阵分解等方法可得到较优的结果[1],本文重点研究如何最大化 $Cov(TUG)$. 由于 $Cov(U')$ 具有单调递增性和子模性,可使用一个贪婪算法,在每一步选择最优的 u_i 加入到 U' 中,最终得到 TUG.

4　典型用户群组发现及应用

4.1　典型用户群组发现算法

项目覆盖率的子模性说明,如果在每次选择用户 u 时均选择 $ICov_{U'}(u)$ 最大的 u,那么最后选出的用户子集 U' 一定是 k-最大覆盖子集. 针对冷门项目难以被覆盖的问题,本文提出基于加权递增项目覆盖率的典型用户群组发现算法,表述如下.

算法 1　一种加权的典型用户群组的发现算法(WTFA)

输入　用户集合 U,项目集合 P,评分集合 R,TUG 大小 k

输出　典型用户群组 TUG

step 1　数据预处理: 根据评分集合 R,将从未在 R 中出现的用户或项目从 U 或 P 中删去.

step 2　算法初始化: 构造空用户子集 $U' = \Phi$,候选用户集合 $U_c = U$.

step 3　从 U_c 中逐一找出候选用户 u,并添加到 U' 中:

```
for i = 1 to k
{
    使用式(2)计算 U_c 中所有用户的加权递增项目覆盖率;
    选出具有最大递增项目覆盖率的用户 u;
    将 u 添加到 U' 中,从 U_c 中删去 u.
}
```

此时 U' 即为所求的典型用户群组 TUG.

4.2　基于典型用户群组的协同过滤推荐算法

得到典型用户群组之后,利用典型用户群组实现推荐算法的提升是一个较为直接的应用方向. 由于典型用户群组的代表性和较高的项目覆盖率,可考虑将协同过滤的最近邻搜索范围从全体用户缩小到典型用户群组,从而降低协同过滤算法的时间开销,同时也不会过多地损失推荐结果的准确程度. 基于典型用户群组的协同过滤推荐算法的步骤如下.

算法 2　一种基于典型用户群组的协同过滤推荐算法(TUG-CF)

输入　用户集合 U,项目集合 P,评分集合 R,典型用户群组 TUG,最近邻数量 k

输出　每个用户 u 的最近邻集合 U_N

step 1　相似度计算: 对于每个用户 u,计算 u 和 TUG 中每个典型用户的相似度(不包含 u 本身),即

$$Sim(u, u_2) = \frac{\sum_{p_j \in P} r_{uj} \times r_{2j}}{\sqrt{\sum_{p_j \in P} r_{uj}^2} \sqrt{\sum_{p_j \in P} r_{2j}^2}},\ u_2 \in TUG.$$

step 2　最近邻搜索: 从 TUG 中选出和 u 相似度最高的 k 个典型用户,作为最近邻集合 U_N.

step 3　评分预测: 对于用户 u 未评分的项目 p,得分由其最近邻集合 U_N 对项目 p 评分的加权平均得到.

step 4　对于所有用户的所有未评分项进行预测,并按得分从高到低排序,即得到推荐结果 R'.

基于最近邻搜索的结果,即可使用常用的基于用户的协同过滤算法得到 TUG-CF 的推荐结果. 由于 TUG 的项目覆盖率并不是 100%,对于无法通过 TUG-CF 得到评分的项,可用它在训练集上的平均得分作为替代.

5　实验与结果分析

本节将通过实验验证 WTFA 和 TUG-CF 算法在真实数据集上的有效性. 首先介绍实验所使用的数

据集和对比方法,然后通过对 WTFA 算法发现的典型用户群组的分析说明其可有效地代表全体用户的兴趣偏好并保持尽可能大的项目覆盖率,随后通过 TUG-CF 和经典协同过滤算法推荐效果的比较,说明 TUG-CF 算法合理使用典型用户群组并有效地优化推荐效果,最后对实验进行总结.

5.1 数据集介绍

实验中主要使用的数据集是 MovieLens-1M 数据集(http://grouplens.org/) . Movielens-1M 数据集是一个公开数据集,包含 100 余万个电影评分,评分范围是从 1 到 5 的整数. 该数据集也被广泛用于推荐系统相关的研究工作[22-23] . 实验中使用所有有评分行为的用户及相关项目,数据集的基本统计情况为电影用户 6 040,项目 3 706,评分 1 000 209,稀疏度 95.53%,流派 18.

5.2 评价标准和对比方法

实验中主要从 3 个角度评价典型用户群组的有效性: 1) 使用子集评分误差 $Err(U')$ 评价项目评分的准确性; 2) 使用项目覆盖率 $Cov(U')$ 评价项目的覆盖率; 3) 使用 Top-K、绝对平均误差(Mean Absolute Error,MAE) 和均方根误差(Root-Mean-Square Error, RMSE) 度量推荐效果的好坏. 它们都是协同过滤算法常用的评估方法[24-26] ,这几种度量方法的计算公式如下:

$$Top-K_u = \frac{\#hit}{K},$$

$$MAE = \frac{1}{|R_{test}|}\sum_{r\in R_{test}} |r_{real} - r_{pred}|,$$

$$RMSE = \sqrt{\frac{\sum_{r\in R_{test}} (r_{real} - r_{pred})^2}{|R_{test}|}},$$

其中,R_{test} 表示测试集,r_{real} 和 r_{pred} 分别表示测试项的真实得分和预测得分.

对比方法分为用户子集选择的对比和推荐效果好坏的对比. 用户子集选择的对比中,本文实现的算法 WTFA 将和 GTFA、MR、MD 这 3 种方法进行对比.

基于流派的典型用户发现(Genre-Based TFA, GTFA) 不是根据项目计算覆盖率,而是根据对不同流派的覆盖率寻找典型用户. 显然找到的典型用户可以覆盖各个流派.

最多评分用户选择(Most Rating, MR) ,即选取在训练集中评分次数最多的用户子集,显然 MR 集合是比较活跃的,能够覆盖较多的项目.

最大差异用户集合(Max Diversity, MD) 是在 MR 的基础上,为评分次数最多的用户逐一选择和他最不相同的另一个用户(即两者之间的用户相似度最小) ,直至 MD 集合达到指定大小 k. 可用如下公式计算用户相似度:

$$Sim(u_1,u_2) = \frac{|P_{u1} \cap P_{u2}|}{|P_{u1} \cup P_{u2}|}.$$

推荐效果好坏的比较,除基于 WTFA 找到的典型用户群组实现 TUG-CF 算法之外,还实现如下 4 种对比算法.

1) 基于流派的 TUG-CF(Genre-Based TUG-CF, GT-CF) ,即在 GTFA 找到的典型用户中寻找最近邻并完成推荐.

2) 基于最多评分用户的协同过滤(Most Rating CF,MR-CF) ,即在 MR 集合中寻找最近邻并完成推荐.

3) 基于最大差异用户的协同过滤(Max Diversity CF, MD-CF) ,即在 MD 集合中寻找最近邻并完成推荐.

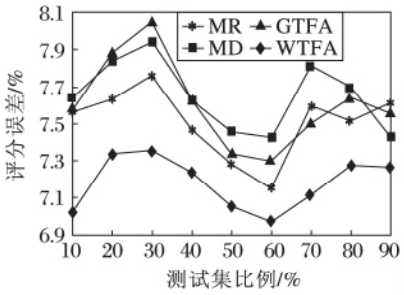
4) 一般的基于用户的协同过滤(User-Based CF,UCF) ,即在所有用户中寻找最近邻并完成推荐,是最基本的基于用户的协同过滤推荐方法.

5.3 用户子集选择的对比实验

5.3.1 不同比例测试集上的对比

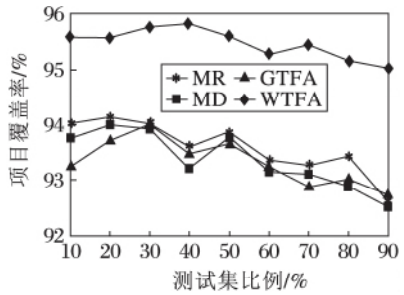
实验中,对于原数据集,随机抽取一定比例的项目及其评分作为测试集,其余的作为训练集,比较使用训练集选出的用户子集是否有效. 比例划分从 10% 到 90%,每 10% 为一档,在每个档次上实验 5 次,并取 5 次的结果平均值为最终结果. 其中,选取的用户子集大小 $k = 50$.

图 1 的实验结果显示,基于加权递增项目覆盖率的典型用户群组发现算法(WTFA) 在评分误差较小的情况下保证最大化的项目覆盖率,选出的用户子集更具有典型性. 而基于流派的典型用户群组发现算法(GTFA) 并没有较为出色的表现,显然对较少出现的项目赋予较高的权重,提高较少出现项目被覆盖的可能. 此外,最多评分(MR) 的用户子集也表现较优的结果,这和通常认识的是一致的. 最大差异(MD) 的用户子集并未表现出较好的结果,可见仅考虑用户之间的差异性,并不能有效提高项目覆盖率和降低评分误差. 另外,随着测试集比例的变化,实验结果也有明显变化. 例如,对于项目覆盖率,当测试集比例增大时,WTFA 的项目覆盖率依然稳定,而其他几种方法都有所下降,也说明 WTFA 选出的典型用户群组更具有典型性和有效性.



(a) 评分误差

(a) Rating error



(b) 覆盖率

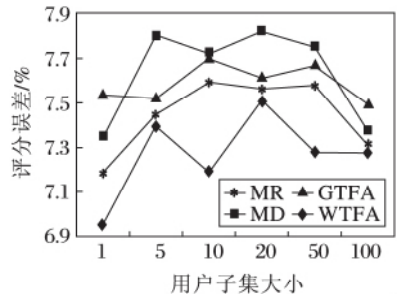
(b) Coverage rate

图 1　不同比例测试集上的实验结果对比

Fig. 1　Results comparison of testing set with different proportions

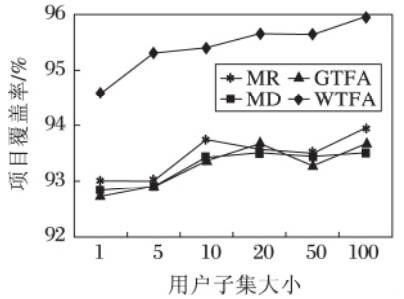
5.3.2　不同用户子集的大小比较

由于用户子集 k 的大小可自行定义，实验也考察了随着用户子集大小 k 的变化，典型用户群组发现算法能否得到较优的结果. 在实验中，对于原数据集，随机抽取50%的项目及其评分作为测试集，其余的作为训练集，比较使用训练集选出的用户子集是否有效. 用户子集大小 k 分别取 1、5、10、20、50 和 100，再取 5 次实验结果平均值作为最终结果，实验结果如图 2 所示.



(a) 评分误差

(a) Rating error



(b) 覆盖率

(b) Coverage rate

图 2　不同大小用户子集上的对比实验结果

Fig. 2　Results comparison of user subset with different sizes

由图 2 可发现，WTFA 依然保持最优的结果，次优结果仍然是最多评分的用户子集. 另外随着 k 的增大，评分误差也逐渐接近，说明在用户子集规模足够大时，子集平均分已越来越接近项目的实际平均分. 但项目覆盖率的提升要缓慢得多，由于通过权重提升较少出现的项目被覆盖的可能性，WTFA 在项目覆盖率上的优势依然明显.

5.3.3　不同流派项目比较

MovieLens-1M 数据集提供的电影流派有 18 个，每部电影可属于多个流派. 电影和流派之间的统计情况见表 1. 由表可知，不同流派的电影数量差别很大，较常见的流派如剧情片和喜剧片分别有1 603 部和1 200 部，而小众的流派如黑色流派和西部流派，分别只有44 部和68 部，其他流派的数量也都有很大区别.

同时，不同流派的电影得到的评分次数也有所不同，例如犯罪流派，平均每部电影仅被评分 70 次，而数量与其接近的音乐剧，平均每部被评分 428.2 次，可看出用户评分覆盖范围的明显差别. 类似地，科幻流派虽然数量较少(276 部)，但是平均评分数量却高达 711.7 次，说明这是一个较受欢迎的小众流派.

实验首先考察给定用户子集规模和测试集比例时，所有流派上的评分误差和项目覆盖率对比. 取用户子集大小 $k = 50$，测试集比例 50%，实验结果如图 3 所示. 可发现，流派包含的电影数量以及流派中电影的平均评分次数和实验结果具有一定的相关性. 比如，电影数量和评分次数都很少的犯罪流派，在评分误差和项目覆盖率上都比较差. 与之类似，电影数量较少的儿童、浪漫、西部等流派，实验结果也比较差. 而恐怖流派虽然电影数量较多，但是平均评

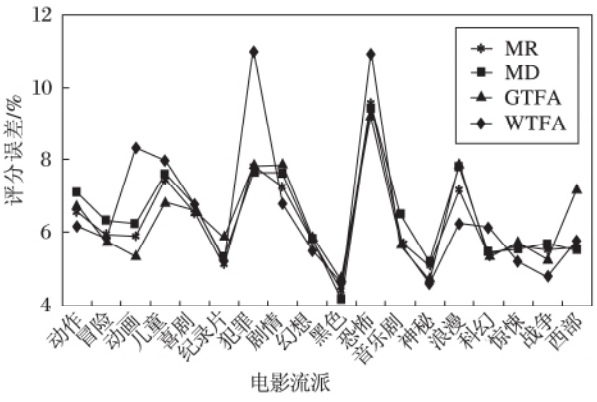
分次数较少，实验结果也不理想. 在实验中，虽然大部分情况下 WTFA 都取得较好的结果，但是在部分小众流派如犯罪、恐怖流派的评分误差上表现略差于其他方法. 在小众流派上，着重考虑流派覆盖率的 GTFA 方法取得较好的结果，说明对于小众流派，不能单纯地从项目覆盖率的角度处理.

表 1 电影流派统计情况

Table 1 General statistics of movie genres

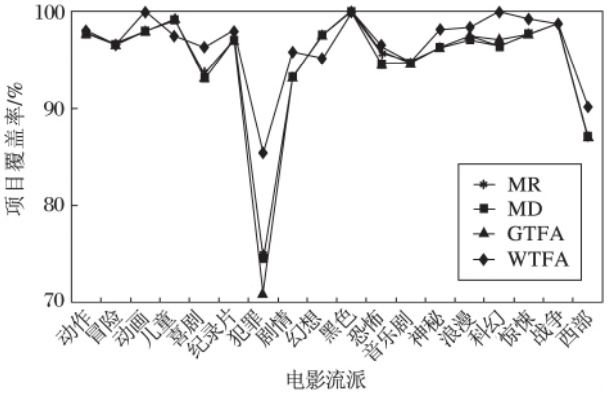
流派	英文名	电影数量	平均评分次数
动作	Action	503	578.6
冒险	Adventure	283	523.3
动画	Animation	105	470.6
儿童	Children´s	251	305.9
喜剧	Comedy	1200	341.2
纪录片	Documentary	211	427.6
犯罪	Crime	127	70
剧情	Drama	1603	250.0
幻想	Fantasy	68	585.5
黑色	Film-Noir	44	434.8
恐怖	Horror	343	285.0
音乐剧	Musical	114	428.2
神秘	Mystery	106	397.8
浪漫	Romance	471	343.9
科幻	Sci-Fi	276	711.7
惊悚	Thriller	492	444.2
战争	War	143	543.9
西部	Western	68	397.8

实验考察给定用户规模时，随着测试集比例变化，各个流派上的实验结果的对比. 取用户子集大小 $k = 50$，测试集比例取 10%、50% 和 90%，实验结果如图 3(50%)、图 4(10%) 和图 5(90%) 所示.



(a) 评分误差

(a) Rating error

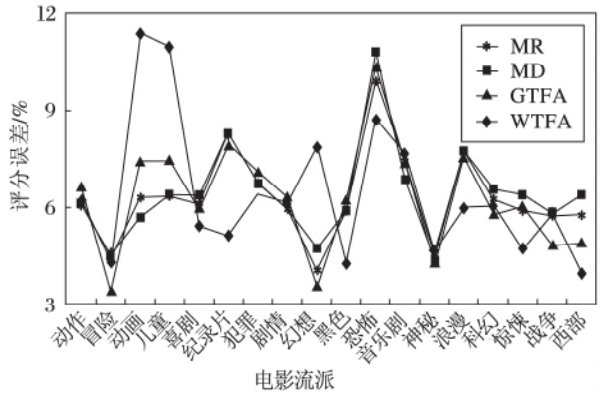


(b) 覆盖率

(b) Coverage rate

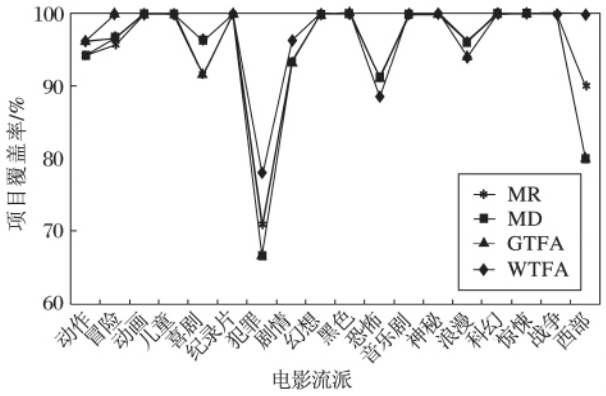
图 3 测试集比例为 50%，$k = 50$ 时不同流派项目上的对比结果

Fig. 3 Results comparison on different genres with 50% test set and $k = 50$



(a) 评分误差

(a) Rating error



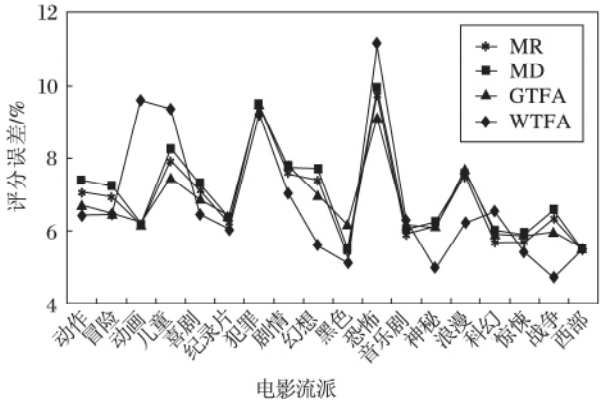
(b) 覆盖率

(b) Coverage rate

图 4 测试集比例为 10%，$k = 50$ 时不同流派项目上的对比结果

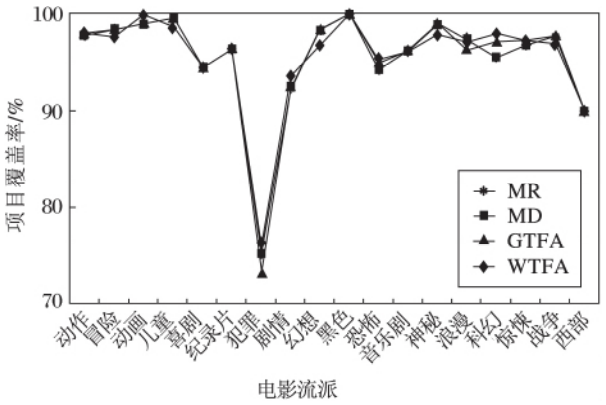
Fig. 4 Results comparison on different genres with 10% test set and $k = 50$

由图3至图5的对比可看出,随着测试集占比减小,所有方法的实验效果都有所上升. 图4(b)中,在测试集比例10%的情况下,几乎所有流派的覆盖率都接近100%. 按流派分别比较实验结果,由于犯罪流派的小众性,该流派的实验结果是最差的. 相较于其它方法,在测试集比例变化时,WTFA依然保持最好的结果. 类似的情况也存在于大部分流派的实验结果中. 进一步的分析可发现,虽然WTFA在部分小众流派上的评分误差较大,但是始终保持较高的项目覆盖率. 另外,随着测试集比例的上升,其依然保持了较小的评分误差和较高的项目覆盖率,这说明WTFA在面临新的从未出现过的项目时,有较强的实用性,其选出的典型用户群组可用于实用推荐系统中的长期跟踪和研究.



(a) 评分误差

(a) Rating error



(b) 覆盖率

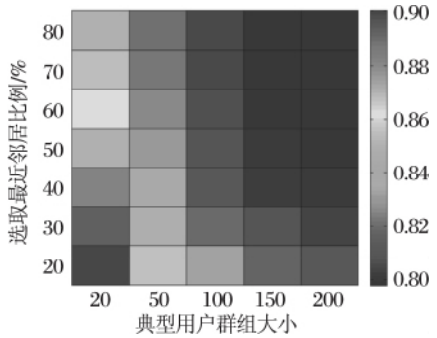
(b) Coverage rate

图5　测试集比例为90%,k = 50时不同流派项目上的对比结果

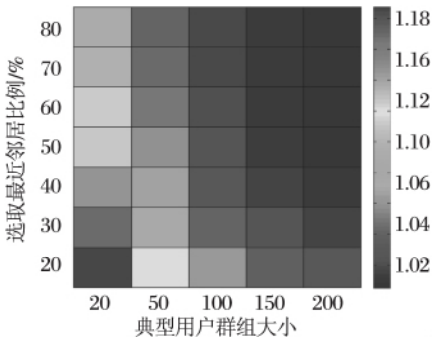
Fig. 5　Results comparison on different genres with 90% test set and k = 50

5.4　推荐效果比较

在进行基于MAE和RMSE的推荐效果的对比试验时,将MovieLens-1M数据集的评分数据进行完全随机划分. 比例划分从10%到90%,每10%为一档,在每个档次上实验5次,并取5次的结果平均值为最终结果. 其中,选取的用户子集大小为50,协同过滤算法的最近邻大小为20. 参数的选择是基于图6中TUG-CF推荐效果随着参数变化的比较得出的,可见随着用户子集和最近邻大小的增加,推荐效果逐渐提升,但同时计算开销也会增大,因此基于综合考虑选择目前的参数.



(a) MAE



(b) RMSE

图6　测试集比例为50%时TUG-CF推荐效果随着参数变化情况

Fig. 6　TUG-CF results versus different parameters with 50% test set

之前所述的5种算法的推荐结果对比如表2所示. 由表可知,在大部分情况下,TUG-CF的评分预测效果都是最优的. 这说明在较小的用户子集中寻找所有用户的最近邻,在充分考虑用户子集对推荐项的覆盖率情况下,可得到较好的结果. GT-CF、MR-CF和MD-CF方法也显示,在用户子集的实现上,WTFA方法找到的典型用户子集对于推荐算法的提升最高,其典型性和代表性也最好. 另外,在测

试集比例较大时，经典的基于用户的协同过滤算法(UCF)取得最好的结果，这说明典型用户群组的代表性也取决于用户数据的完整性与否，原始数据不完整的情况下，典型用户群组的优势并不明显.

表 2 推荐效果比较

Table 2 Comparison of recommendation results

测试集比例/%	MAE					RMSE				
	TUG-CF	GT-CF	MR-CF	MD-CF	UCF	TUG-CF	GT-CF	MR-CF	MD-CF	UCF
10	**0.8047**	0.8184	0.8124	0.8148	0.8215	**1.0204**	1.0322	1.0248	1.0276	1.0519
20	**0.8018**	0.8163	0.8116	0.8106	0.8285	**1.0245**	1.0334	1.0301	1.0303	1.0648
30	**0.8135**	0.8262	0.8230	0.8228	0.8365	**1.0379**	1.0451	1.0411	1.0418	1.0769
40	**0.8207**	0.8326	0.8280	0.8282	0.8463	**1.0462**	1.0546	1.0498	1.0504	1.0895
50	**0.8331**	0.8490	0.8460	0.8423	0.8608	**1.0640**	1.0769	1.0725	1.0709	1.1120
60	**0.8498**	0.8653	0.8654	0.8654	0.8726	**1.0878**	1.1014	1.0996	1.0998	1.1313
70	**0.8726**	0.8823	0.8841	0.8853	0.8735	**1.1238**	1.1303	1.1311	1.1325	1.1360
80	0.9002	0.9136	0.9171	0.9147	**0.8593**	1.1694	1.1795	1.1828	1.1800	**1.1116**
90	0.9311	0.9420	0.9503	0.9523	**0.8347**	1.2233	1.2336	1.2424	1.2451	**1.0694**

Top-K 的实验结果如图 7 所示，实验的测试集比例为 50%. 实验中选取测试集中真实得分大于 3 的项目(假定为用户喜欢的项目)和各个算法预测得分大于 3 的无评分项目(既不在训练集也不在测试集中)构成推荐列表，并检查这些测试集项目的 Top-K 分布，最终得到对比结果. 由图中可看出，基于 TUG-CF 给出的评分得到的排序结果可将较多的用户喜欢的项目排在靠前的位置，其它基于用户子集的方法结果也优于经典的协同过滤方法，可见典型用户群组对于推荐结果的排序也能起到优化作用. 另外，注意到每条曲线最后接近 100% 时都有一个跃升，这是由于部分用户喜欢的项目预测得分小于 3 而被排在推荐列表的最后造成的.

从推荐方法的比较实验可看出，TUG-CF 在推荐评分精确性和推荐结果排序准确性上都优于对比方法，通过合理地运用典型用户群组改进协同过滤算法，可有效提升推荐效果.

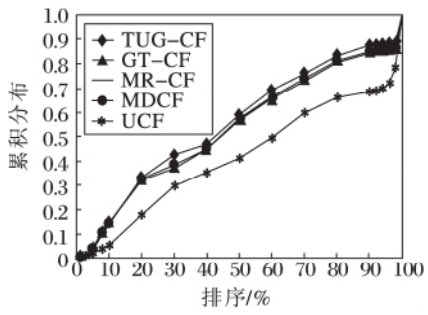


图 7 Top-K 的实验对比结果

Fig. 7 Comparison of Top-K results

5.5 实验总结

本节通过在真实数据集上的实验，验证了 WTFA 算法和 TUG-CF 算法的有效性. 实验结果充分说明 WTFA 算法找到的典型用户群组 TUG 更具有典型性，并可通过 TUG-CF 算法有效地应用于推荐过程的优化之中. 具体而言，相较于最多评分、最大差异的用户子集，WTFA 找到的典型用户群组在平均评分误差和项目覆盖率两个实验指标上都取得更优的结果. 通过在数据集电影项目的不同流派上的细分对比，典型用户群组也可更准确地评估不同规模、不同热门程度流派的平均得分并保持较大的项目覆盖率. 通过推荐效果比较，验证了 TUG-CF 能够合理地使用典型用户群组，并且有效提升协同过滤推荐算法的推荐准确性.

6 结 束 语

本文提出推荐系统中典型用户群组的概念，并相应提出一种典型用户群组的发现算法，以期发现推荐系统用户的最优子集，用于正确反映全体用户的兴趣偏好并最大化对推荐项目的覆盖率. 活跃用户群组、多样化用户群组等实验验证了本文算法发现的典型用户群组能够获得更好的效果. 如何将典型用户群组应用到推荐系统的优化和改进中是一个值得深入研究的课题. 另一个可能的研究方向是运用典型用户群组处理推荐系统冷启动问题. 此外，基于典型用户群组这种代表性的用户子集的研究，可

以帮助推荐系统研究者更加深入了解和分析用户的兴趣偏好并进一步提升推荐效果,这也是未来的研究方向之一.

参 考 文 献

[1] Adomavicius G, Tuzhilin A. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Trans on Knowledge and Data Engineering, 2005, 17(6): 734 – 749

[2] Kim K J, Ahn H. A Recommender System Using GA *K*-Means Clustering in an Online Shopping Market. Expert Systems with Applications, 2008, 34(2): 1200 – 1209

[3] Sarwar B M, Karypis G, Konstan J, *et al*. Recommender Systems for Large-Scale E-Commerce: Scalable Neighborhood Formation Using Clustering. [EB/OL]. [2014 – 03 – 20]. Http://glaros.dtc.umn.edu/gkhome/fetch/pagers/clusterICCIT02.pdf

[4] Ungar L, Foster D. Clustering Methods for Collaborative Filtering. [EB/OL]. [2014 – 03 – 18]. http://www.aaai.org/Papers/Workshops/1998/WS – 98 – 08/WS98 – 08 – 029.pdf

[5] Liu Q, Chen E H, Xiong H, *et al*. A Cocktail Approach for Travel Package Recommendation. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(2): 278 – 293

[6] Tan C, Liu Q, Chen E H, *et al*. Object-oriented Travel Package Recommendation. ACM Transactions on Intelligent Systems and Technology, 2013, 5(3): 43: 1 – 43:26

[7] Lin H, Bilmes J. A Class of Submodular Functions for Document Summarization // Proc of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, USA, 2011: 510 – 520

[8] Sipos R, Swaminathan A, Shivaswamy P, *et al*. Temporal Corpus Summarization Using Submodular Word Coverage // Proc of the 21st ACM International Conference on Information and Knowledge Management. Maui, USA, 2012: 754 – 763

[9] El-Arini K, Veda G, Shahaf D, *et al*. Turning Down the Noise in the Blogosphere // Proc of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, 2009: 289 – 298

[10] Krause A, Guestrin C. Submodularity and Its Applications in Optimized Information Gathering. ACM Transactions on Intelligent Systems and Technology, 2011, 2(4): 389 – 396

[11] Pennacchiotti M, Silvestri F, Vahabi H, *et al*. Making Your Interests Follow You on Twitter // Proc of the 21st ACM International Conference on Information and Knowledge Management. Maui, USA, 2012: 165 – 174

[12] Hammar M, Karlsson R, Nilsson B J. Using Maximum Coverage to Optimize Recommendation Systems in E-Commerce // Proc of the 7th ACM Conference on Recommender Systems. Hong Kong, China, 2013: 265 – 272

[13] McNee S M, Riedl J, Konstan J A. Being Accurate is Not Enough: How Accuracy Metrics Have Hurt Recommender Systems // Proc of the Extended Abstracts on Human Factors in Computing Systems. Montreal, Canada, 2006: 1097 – 1101

[14] Zhang M, Hurly N. Evaluating the Diversity of Top-N Recommendations // Proc of the 21st International Conference on Tools with Artificial Intelligence. Newark, USA, 2009: 457 – 460

[15] Said A, Fields B, Jain B J, *et al*. User-Centric Evaluation of a K-Furthest Neighbor Collaborative Filtering Recommender Algorithm // Proc of the Conference on Computer Supported Cooperative Work. San Antonio, USA, 2013: 1399 – 1408

[16] Zhang F G, Xu S H. Research on Recommendation Diversification in Trust Based E-Commerce Recommender Systems. Journal of the China Society for Scientific and Technical Information, 2010, 29(2): 350 – 355 (in Chinese)
(张富国,徐升华. 基于信任的电子商务推荐多样性研究. 情报学报, 2010, 29(2): 350 – 355)

[17] Cui C Y, Ma J. An Image Tag Recommendation Approach Combining Relevance with Diversity. Chinese Journal of Computers, 2013, 36(3): 654 – 663 (in Chinese)
(崔超然,马 军. 一种结合相关性和多样性的图像标签推荐方法. 计算机学报, 2013, 36(3): 654 – 663)

[18] Park Y J, Tuzhilin A. The Long Tail of Recommender Systems and How to Leverage It // Proc of the ACM conference on Recommender Systems. Lausanne, Switzerland, 2008: 11 – 18

[19] Anderson C. The Long Tail. New York, USA: Random House Audiobooks, 2007

[20] Nemhauser G L, Wolsey L A, Fisher M L. An Analysis of Approximations for Maximizing Submodular Set Functions – I. Mathematical Programming, 1978, 14(1): 265 – 294

[21] Feige U. A Threshold of in n for Approximating Set Cover. Journal of the ACM, 1998, 45(4): 634 – 652

[22] Liu Q, Xiang B, Chen E H, *et al*. Influential Seed Items Recommendation // Proc of the 6th ACM Conference on Recommender Systems. Dublin, Ireland, 2012: 245 – 248

[23] Adomavicius G, Kwon Y O. Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques. IEEE Trans on Knowledge and Data Engineering, 2012, 24(5): 896 – 911

[24] Herlocker J L, Konstan J A, Terveen L G, *et al*. Evaluating Collaborative Filtering Recommender Systems. ACM Transactions on Information Systems, 2004, 22(1): 5 – 53

[25] Koren Y. Factorization Meets the Neighborhood: A Multifaceted Collaborative Filtering Model // Proc of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, USA, 2008: 624 – 634

[26] Koren Y, Bell R M, Volinsky C. Matrix Factorization Techniques for Recommender Systems. Computer, 2009, 42(8): 30 – 37