

# 基于子组与社会行为的缩小群组推荐列表方法

毛宇佳 刘学军 徐新艳 张欣

(南京工业大学计算机科学与技术学院 南京 211816)

**摘要** 以多个用户为推荐对象的组推荐系统已成为研究热点。目前,组推荐系统大多考虑如何充分挖掘用户偏好来尽可能满足所有用户的需求,但这也造成了推荐列表规模过大的问题,从而导致群组成员无法快速做出决定。针对该问题,文中提出了一种缩小群组推荐列表的方法(Recommendation Method based on Sub-Group and Social Behavior, RMSGSB)。该方法通过划分子组来缩小群组规模并减少群组偏好属性数量,利用成员的社会行为,从容忍度与利他行为两方面为子组分配权重,以保证推荐公平性。在真实数据集上的实验对比结果表明,该算法具有更好的群组推荐效果。

**关键词** 群组推荐,子组,容忍度,利他行为,推荐列表

**中图分类号** TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.02.031

## Method Based on Sub-group and Social Behavior for Narrowing Recommended List for Groups

MAO Yu-jia LIU Xue-jun XU Xin-yan ZHANG Xin

(College of Computer Science and Technology, Nanjing Tech University, Nanjing 211816, China)

**Abstract** Group recommend systems have been drawn a lot of attention for their capability of targeting users' preference to satisfy all the users' requirements, which cause the problem that the recommended list is too large simultaneously, thus making it hard for group members to make their decisions. Regarding current issues, a method for narrowing the recommended list for groups called RMSGSB (A Recommendation Method based on Sub-group and Social Behavior) was proposed. In order to narrow the group scale and the amount of group preferences, the method divides the target group into several sub-groups. For guaranteeing the fairness of recommending, the weight of sub-groups is calculated according to the tolerance and altruism of group members' social behavior. Based on real date set, the experiment results show that the proposed method has better performance than other methods on group recommendation.

**Keywords** Group recommendation, Sub-group, Tolerance, Altruism, Recommended list

## 1 引言

随着互联网时代的兴起,用户在网络上接收信息的同时也在制造信息,这便造成了数据的爆炸式增长。如何从海量数据中快速且准确地获得用户感兴趣的内容,是目前亟待解决的问题。推荐系统旨在在过载搜索空间中为用户提供感兴趣的项目。传统推荐系统也存在一些不足,例如,当推荐对象是一群人时,推荐结果无法准确满足多数人的需求。因此,组推荐系统应运而生<sup>[1]</sup>。

组推荐系统的目的是给一群偏好不同的用户推荐项目,这群用户由于各种原因聚到一起形成群组。推荐项目应该最大化成员间的满意度,并最小化成员间的不公平性。组推荐

系统的应用在日常生活中十分常见,例如,家庭成员共同决定收看哪个电视节目,以及决定去哪家餐馆聚餐等。目前火爆的购物 APP 之一“拼多多”也是通过用户与朋友、家人等发起拼单,从而享受更低的折扣。

根据目标群组划分,可以将群组划分为固定组和临时组。固定组是指长久以来存在的组(如家庭成员形成的组),这些组具有组与项之间的历史交互信息,要给固定组进行推荐时,只需要将组视为一个虚拟用户,并采用个性化的推荐算法;临时组是指不存在历史交互信息的群组,该群组中的成员初次聚在一起,由于该群组缺乏历史信息,因此采用聚合组成员偏好的方法进行推荐。将组成员偏好融合成群组偏好的常见方法可以分为两种:1)推荐融合,首先为每位成员生成推荐,再

到稿日期:2018-06-14 返修日期:2018-09-12 本文受国家重点研发计划子课题(2017YFC0805605),江苏省重点研发计划项目(BE2015697, BE2017617)资助。

毛宇佳(1994-),男,硕士生,主要研究方向为数据挖掘、推荐系统,E-mail:myj.mao@qq.com;刘学军(1971-),男,博士,教授,CCF高级会员,主要研究方向为数据库、数据挖掘、推荐系统等,E-mail:lxj\_njgd@163.com(通信作者);徐新艳(1980-),女,硕士,讲师,主要研究方向为数据挖掘、智能信息处理;张欣(1994-),女,硕士生,主要研究方向为机器学习、推荐系统等。

通过特定的融合策略得到群组推荐;2)模型融合,融合成员的偏好模型,从而得到群组偏好模型,然后再进行推荐。推荐融合可缓解评分稀疏性高的问题,模型融合可缓解成员之间交互性差的问题。在获取群组推荐时,如何同时满足较小的推荐列表规模和较高的准确性,也是一个需要解决的问题。到目前为止,大多数群组推荐主要集中于解决如何发现成员间的潜在关联,以及如何挖掘成员的潜在偏好来提升推荐结果准确度的问题,而研究推荐列表的规模大小对群组成员满意度的影响并不大。直接将组成员推荐列表融合成组推荐列表会导致推荐列表规模过大,不利于群组做出正确的选择,再选取推荐列表中的 top- $k$  个项目也无法保证该 top- $k$  个项目能最大化满足群组成员的偏好。因此,需要一种既能最小化推荐列表规模又能最大化考虑各成员偏好的方法。

本文提出了一种缩小群组推荐列表的方法,首先基于组成员的共同偏好将群组划分为多个子组,缩小群组规模,将共同偏好作为子组偏好以减少群组偏好属性的数量;然后计算子组间的相似度,以此寻找相似子组,得到子组推荐列表,通过汇总子组推荐列表得到初始推荐列表;接着根据组成员的容忍度与利他行为计算子组权重,再通过权重与子组偏好得到群组偏好;最后将初始推荐列表与群组偏好进行属性匹配,过滤相似度低的项目,得到最终的推荐列表。

## 2 相关工作

目前,在群组推荐系统中的相关研究主要集中于挖掘用户与项目的相关信息。Feng 等人<sup>[2]</sup>用三元图代表用户、群组、项目之间的关联,通过考虑总的已存在的关系,利用随机游走重启来检测群组与未评分项之间的综合关联。该方法虽然能够较好地挖掘用户、群组、项目三者的关联,但却忽视了每个用户的固有个性;同时,对于某些不活跃用户,由于缺少与群组和项目之间的联系,也无法准确捕捉到这些用户的偏好;另外,群组推荐系统与传统个人推荐系统不同,群组中成员的偏好受自身和群组成员的影响,此问题可利用在个性化推荐中引入社交关系的方式来解决<sup>[3-5]</sup>,此类方法在固定群组中能取得较好的实验性能,而在临时群组中,由于缺乏互动信息,无法判断成员间的社交关系,因此性能较差。有研究致力于挖掘组成员的特征,例如,Hong 等人<sup>[6]</sup>通过组成员间的历史互动信息分析他们的社交吸引力和信任度,但是该方法容易受数据稀疏性的影响,如果群组为临时群组,那么组成员之间的互动信息可能较少,不利于提取社交吸引力与信任度;Yuan 等人<sup>[7]</sup>基于组成员的朋友数量与相似用户的数量引入了群集度(Gregariousness Degree)的概念,并结合成员在群组中的活跃度来分析成员在群组中所占的权重,以此得到群组偏好,同样地,对于临时群组,该方法的准确性不佳。由于群组成员在群组中的影响力可能不同,因此可以适当增加专家用户的权重,但同时也不能忽略不活跃用户的喜好。Lin 等人<sup>[8]</sup>指出群组中的推荐问题为一个多目标优化问题,并引入了两种概念,即社会福利(Social Welfare)和公平性。社会福利评估组的总体效用,公平性反映组内成员对推荐满意度的

不平衡性,从每个用户的角度考虑用户满意度,并在保证公平性的前提下最大化群体的满意度。提取项目特征与用户特征时,有研究提取了评论中的信息<sup>[5]</sup>(如图片信息<sup>[9-10]</sup>)或语料库中的信息(如词频<sup>[11]</sup>),这些方法对于最终推荐结果的准确性都有一定程度的提升。

目前,群组推荐研究主要都是针对结果的准确性与公平性进行提升。然而,由于群组成员的偏好存在差异,在保证准确性与公平性的前提下,最终的推荐列表必然会很大。文献<sup>[12]</sup>将基于多数人的偏好策略(Majority, MAJ)和最小痛苦策略(Least Misery, LM)两种融合策略分别与可扩展动态信息增益(EDIG)与随机式(RANDOM)启发法进行结合,提出了 4 种缩小推荐列表的方法,即 EDIG w MAJ, EDIG w LM, RANDOM w MAJ, RANDOM w LM。实验表明:在不同 top- $k$  取值下,EDIG 的性能始终优于 RANDOM;而在冷启动问题或当项目左偏(Left-Skew,即该项目较容易被成员排斥)时,LM 的性能优于 MAJ 的性能,而在其他情况下,MAJ 优于 LM;能获得最好性能的是 EDIG w MAJ。受该文献启发,在得到推荐列表 top- $N$  后,通过一些方法再次筛选出最适合的 top- $k$  个项目。然而,该文献也存在一些不足,如需要用户在每一次请求时都提供评分,这可能会导致用户产生厌烦情绪;同时,文献还指出筛选出的 top- $k$  中存在一个 necessary winner 项,即该项必然满足群组的综合偏好,由于群组推荐是一个多目标优化问题,存在满足群组所有成员偏好的项的概率很小,因此该 necessary winner 项并不一定存在。事实上,对于一个群组产生的推荐列表,规模过大与过小都会影响群组成员最终的选择,列表越大,意味着多样性越大,成员选择就会更困难;列表过小,无法保证推荐列表中的项目尽可能多地满足成员喜好。推荐性能的好坏最终是由用户满意度决定的,如果用户能快速、方便地获得自己想要的内容,那么用户对该推荐系统是满意的。因此,除了推荐准确性与公平性之外,推荐列表的规模大小以及推荐的呈现方式都会对用户满意度产生影响。

## 3 算法设计

本文旨在满足推荐准确性与公平性的前提下,最大程度地缩小群组推荐列表,使群组成员可以更加方便地做出选择,算法的整体框架如图 1 所示。首先,将群组成员分为活跃用户与不活跃用户,从不同的角度分析每个用户的偏好,挖掘用户之间的隐式关联,并将群组划分为若干个子组,得到子组偏好,将每个子组视为一个虚拟用户,缩小群组规模;分析子组间的相似度,通过相似子组得到子组推荐列表;通过成员的社会行为与社会关系,基于 TKI 冲突测试计算子组权重,得到群组偏好  $p_G$ ;综合之前得到的子组推荐列表,得到目标群组的 top- $N$  个推荐项目,将 top- $N$  个推荐项目中项目的主题特征与群组偏好  $p_G$  进行匹配,过滤掉相似度差的项目,得到最终的推荐列表 top- $N'$ ,保证能保留列表中最符合成员偏好的项目。

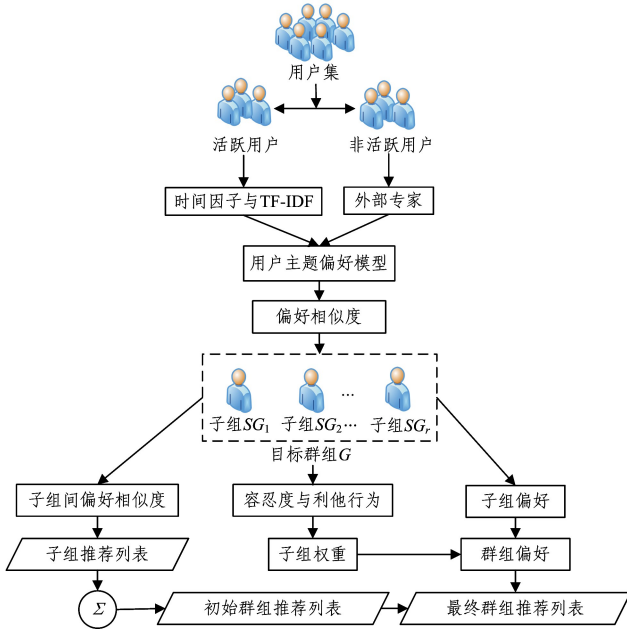


图1 算法的整体框架

Fig. 1 Framework of algorithm

### 3.1 子组的划分与子组推荐列表的获取

为了方便下文叙述,本文中的项目以电影为例。

#### 3.1.1 基于活跃用户与非活跃用户的子组划分

##### (1) 获取项目的主题特征属性

本文中,电影  $m_s$  的主题特征由一个  $k$  维主题向量来表示,这  $k$  个主题特征利用 LDA 主题模型获得。根据每部电影中的主题分布,得到电影  $m_s$  的  $k$  维主题向量模型:

$$R_{m_s} = (f_{m_s, g_1}, f_{m_s, g_2}, \dots, f_{m_s, g_k})$$

其中,  $f_{m_s, g_i}$  表示电影  $m_s$  中是否含主题特征  $g_i$ 。  $f_{m_s, g_i} = 1$  表示电影  $m_s$  中含有主题特征  $g_i$ ;  $f_{m_s, g_i} = 0$  表示电影  $m_s$  中不含主题特征  $g_i$ 。

##### (2) 基于时间因子与 TF-IDF 的活跃用户偏好模型

根据用户的历史交互信息,可将用户划分为活跃用户与非活跃用户。对于活跃用户,我们利用用户的评分信息来获取用户偏好。对用户而言,他们对每种项目的主题特征会有不同的偏好程度,例如:用户  $u_j$  共有 8 条评分记录,其中 6 条是关于喜剧类电影的评分,2 条是关于科幻类电影的评分,那么在用户所有的评分记录中,可以认为用户对喜剧类电影的偏好程度更大。我们通过 TF-IDF 方法<sup>[13]</sup>来获取用户  $u_j$  对每种主题特征  $g_i$  的偏好程度。通常采用如下公式计算一个词  $w_i$  的 TF-IDF 值:

$$\begin{aligned} TF-IDF(w_i) &= tf(w_i) \times idf(w_i) \\ &= tf_j(w_i) \times \log(N/df(w_i)) \end{aligned}$$

其中,  $tf_j(w_i)$  代表词  $w_i$  在文本  $j$  中出现的频率,词  $w_i$  出现的频率越高,表示该文本中该词所代表的属性越强烈;  $idf(w_i)$  是指词  $w_i$  在所有文本集中出现的逆文件频率,逆文件频率越高,则该词的主题特征区分力越差;  $N$  代表所有的文本集;  $df(w_i)$  表示包含词  $w_i$  的文本集。

根据 TF-IDF 的定义公式,构建用户  $u_j$  对主题特征  $g_i$  的偏好:

$$p_{u_j, g_i} = \frac{\sum_{s=1}^n r_{u_j, m_s} \times f_{m_s, g_i}}{\sum_{s=1}^n r_{u_j, m_s}} \log \frac{\sum_{s=1}^n \sum_{i=1}^k f_{m_s, g_i}}{\sum_{s=1}^n \sum_{i=1}^k f_{m_s, g_i}}, 1 \leq i \leq k \quad (1)$$

其中,  $r_{u_j, m_s}$  表示用户  $u_j$  对电影  $m_s$  的评分,  $n$  表示用户  $u_j$  评分的电影数量,因此我们能够得到用户  $u_j$  的主题特征偏好模型  $p_{u_j} = (p_{u_j, g_1}, p_{u_j, g_2}, \dots, p_{u_j, g_k}) (1 \leq j \leq m)$ , 其中  $m$  表示目标群组  $G$  中的成员个数。上述公式反映了用户  $u_j$  对主题特征  $g_i$  的偏好程度,但是由于用户偏好会随着时间的改变而改变,因此每个用户存在一些固定偏好和临时偏好。固定偏好是指用户长久以来一直存在的偏好,而临时偏好是指用户近期产生的偏好。例如:某用户共观看了 10 部电影,其中有 7 部喜剧类电影在半年前观看,而剩余 3 部科幻类电影在近一个月观看,我们可以判断,该用户当前的偏好更倾向于科幻类电影,而采用上述公式则会做出用户更倾向于喜剧类电影的判断。因此,此类用户的偏好迁移行为也应该加以考虑。我们利用遗忘函数来表示用户的兴趣随时间变化而衰退的程度,如下所示:

$$f(\Delta t) = \exp\left(-\frac{\log_2(\Delta t + 1)}{T_0}\right) \quad (2)$$

其中,  $\Delta t$  代表从用户发生评分行为到当前时间点的时间差,  $T_0$  是衰退系数,控制兴趣衰减的快慢,  $T_0$  越大,表示兴趣衰减速度越慢。对于临时偏好,即刚产生的偏好,由于时间差  $\Delta t$  较小,因此用户的遗忘速度较快;对于固定偏好,遗忘速度则较慢。融入遗忘函数的用户评分公式为:

$$r'_{u_j, m_s} = r_{u_j, m_s} \times f(\Delta t) \quad (3)$$

改进公式后,得到基于时间因子的用户  $u_j$  对主题特征  $g_i$  的偏好:

$$p'_{u_j, g_i} = \frac{\sum_{s=1}^n r'_{u_j, m_s} \times f_{m_s, g_i}}{\sum_{s=1}^n r'_{u_j, m_s}} \log \frac{\sum_{s=1}^n \sum_{i=1}^k f_{m_s, g_i}}{\sum_{s=1}^n \sum_{i=1}^k f_{m_s, g_i}}, 1 \leq i \leq k \quad (4)$$

因此,用户  $u_j$  的主题特征偏好模型为  $p'_{u_j} = (p'_{u_j, g_1}, p'_{u_j, g_2}, \dots, p'_{u_j, g_k}) (1 \leq j \leq m)$ 。

##### (3) 基于外部专家的非活跃用户偏好模型

针对非活跃用户,由于较难挖掘到其历史交互信息,且对主题偏好的提取较为困难,因此我们考虑是否存在外部专家来反映他们的兴趣,以及应该如何选取外部专家。

首先研究用户的社会关系,通过观察新浪微博中的用户关注列表可以发现:有很大部分互相关注(双向)的关系通常在真实世界中也有联系,如朋友、同事等。尽管用户之间这种互相关注的关系代表了更亲密的社会联系,但对于偏好提取并不适用,因为这种互相关注的连接是基于真实社交关系而非用户自身兴趣的。此外,我们还发现单向的关系(如用户关注)通常是根根据用户自身兴趣而形成的,这种关系更能用来反映用户偏好。例如:用户会关注他喜爱的社会名人,但是在生活中他们并不认识。这种单向连接大致可以分为两种:1) 社会名人,拥有数量庞大的关注者,并且在真实生活中被大众所认识;2) 专家,在某一特定领域被广泛关注。我们将这两类用户统称为外部专家。

对于用户与外部专家之间的联系,我们通过用户与外部

专家之间的互动信息来获得;用户对他关注的外部专家有不同的偏好程度,具体体现在他们之间的互动频率不同。我们通过共同行为比率(Common Behavior Rate, CBR)来计算用户对外部专家的偏好程度。CBR 是指当前用户(组成员)与外部专家之间发生的共同行为(如共同转发或评论一条微博)与当前用户关注的所有专家的总行为(总微博数)之间的比率。通常来说, CBR 的值越大,代表该外部专家的喜好更能反映当前用户的喜好。用  $\eta_{u_j, e_i}$  来表示用户  $u_j$  与外部专家  $e_i$  之间的 CBR, 计算公式为:

$$\eta_{u_j, e_i} = \frac{\gamma_{u_j, e_i}}{\sum_{e_i \in I_{u_j}} \gamma_{u_j, e_i}} \quad (5)$$

其中,  $\gamma_{u_j, e_i}$  表示用户  $u_j$  转发或评论由  $e_i$  发送的微博,  $I_{u_j}$  表示用户  $u_j$  关注的外部专家集合。

接着计算外部专家的主题偏好。使用  $p_{e_i}$  来代表外部专家的偏好模型, 通过外部专家  $e_i$  发布或分享的内容, 得到外部专家的偏好模型:  $p_{e_i} = (p_{e_i, g_1}, p_{e_i, g_2}, \dots, p_{e_i, g_k})$ 。外部专家对主题特征  $g_i$  的偏好  $p_{e_i, g_i}$  同样采取 TF-IDF 方法获得:

$$p_{e_i, g_i} = \frac{n'_{e_i, g_i}}{n_{e_i}} \log \frac{\sum_{i=1}^{I_{u_j}} n_{e_i}}{\sum_{i=1}^{I_{u_j}} n'_{e_i, g_i}}, 1 \leq i \leq k \quad (6)$$

其中,  $n'_{e_i, g_i} = n_{e_i, g_i} \times f(\Delta t)$ ,  $n_{e_i, g_i}$  表示  $e_i$  的微博中主题特征  $g_i$  的个数,  $n_{e_i}$  表示总微博数。

最后, 基于上述得到的不活跃用户对外部专家的偏好程度以及外部专家的偏好模型, 计算不活跃用户的偏好模型:

$$p'_{U_j} = \Gamma \left( \frac{\sum_{e_i \in I_{u_j}} \eta_{u_j, e_i} p_{e_i}}{\sum_{e_i \in I_{u_j}} \eta_{u_j, e_i}} \right) \quad (7)$$

其中,  $\Gamma(p_M) = \left\{ \frac{p_1}{\sum_i p_i}, \frac{p_2}{\sum_i p_i}, \dots, \frac{p_M}{\sum_i p_i} \right\}$  表示归一化函数。通过建立外部专家与不活跃用户之间的关联, 可以得到不活跃用户的偏好模型。

#### (4) 基于组成员间主题偏好相似度的子组划分

文献[14-16]的研究表明, 群组规模对于推荐结果的准确性具有一定影响, 群组规模越大, 准确性越差。因此, 需要将群组规模尽可能减小。为了减小群组规模, 我们考虑挖掘成员之间的共同偏好, 将有共同偏好的成员聚类形成一个子组, 并将该子组看成一个虚拟用户, 将子组成员的共同偏好作为这个虚拟用户的偏好。成员之间共同偏好的识别根据成员主题特征偏好模型的相似度进行划分。偏好相似度的计算公式为:

$$sim_{g_i}(u_j, u_i) = \frac{p'_{u_j, g_i}}{p'_{u_i, g_i}}, 1 \leq i \leq k \quad (8)$$

通过上述公式, 计算成员在相同主题特征情况下偏好的比值, 比值越接近于 1, 表示当前两成员对该主题特征的偏好相似度越高。划分子组的依据为: 1) 成员间对主题特征  $g_i$  的偏好相似度; 2) 成员对主题特征  $g_i$  的总偏好值。将对某一主题特征的偏好相似度最大且总偏好值高的组成员划分为一组, 最终得到  $r(r \leq m)$  个子组  $SG$ 。由于子组是根据用户个人偏好进行划分的, 在考虑子组偏好时, 直接将子组成员的共同

偏好作为当前子组偏好  $p_{SG_i}$ 。

#### 3.1.2 基于评分与跨项目属性信息的子组推荐列表的获取

在得到子组偏好的基础上, 把子组当作一位虚拟用户, 该虚拟用户的偏好就是子组成员间的共同偏好。接着, 计算子组间的相似度, 并根据相似度来获取子组推荐列表。传统上是基于评分信息来计算组间相似度, 然而组间对相同项目进行评分的稀疏性较大, 仅使用评分信息可能导致推荐结果的准确性欠佳。我们引入跨项目的信息来对子组间的评分信息进行补充。跨项目属性信息是指在两个子组  $SG_x$  和  $SG_y$  的用户评分中选出具有相似主题特征的项目, 利用这些具有相似主题特征的项目来弥补只使用评分的不足。分别从评分与跨项目属性两个方面获得子组间的偏好相似度:

$$sim(SG_x, SG_y) = \lambda sim_R(SG_x, SG_y) + (1 - \lambda) sim_p(SG_x, SG_y) \quad (9)$$

其中,  $sim_R(SG_x, SG_y)$  代表基于评分信息的子组  $SG_x$  与  $SG_y$  间的相似度;  $sim_p(SG_x, SG_y)$  代表基于跨项目属性的子组  $SG_x$  与  $SG_y$  间的相似度;  $\lambda$  为权重因子, 表示为两个子组共同评分项的数量与两个子组所有评分项的数量的比值。我们利用余弦相似度计算公式来计算基于评分与跨项目属性的子组间的相似度:

$$sim_R(SG_x, SG_y) = \frac{\sum_{s \in I_{SG_x} \cap I_{SG_y}} R_{SG_x, m_s} R_{SG_y, m_s}}{\sqrt{\sum_{s \in I_{SG_x} \cap I_{SG_y}} R_{SG_x, m_s}^2} \sqrt{\sum_{s \in I_{SG_x} \cap I_{SG_y}} R_{SG_y, m_s}^2}} \quad (10)$$

$$sim_p(SG_x, SG_y) = \frac{\sum_{i=1}^k p_{SG_x, g_i} p_{SG_y, g_i}}{\sqrt{\sum_{i=1}^k p_{SG_x, g_i}^2} \sqrt{\sum_{i=1}^k p_{SG_y, g_i}^2}} \quad (11)$$

其中,  $I_{SG_x \cap SG_y}$  表示  $SG_x$  和  $SG_y$  共同观看过的电影集合, 通过计算子组成员对电影  $m_s$  的平均评分得到子组  $SG_x$  对电影  $m_s$  的评分  $R_{SG_x, m_s}$ 。

得到子组间的相似度后, 选取相似度高的子组作为目标子组的推荐子组, 并从推荐子组中选取  $k$  部目标子组未观看过的电影作为目标子组的子组推荐列表  $top-k_{SG_i}$ 。

#### 3.2 基于成员容忍度与利他行为的子组权重的获取

为了得到最终的群组推荐列表, 根据组推荐中的偏好融合策略, 采用加权模型来获取群组偏好。首先, 通过子组中成员的容忍度与利他行为来获得每个子组的权重。在群组中, 成员会接收到不是他们喜欢的推荐内容, 不同性格的用户会做出不同反应。为了减小组成员间的冲突与摩擦, 达到较好的推荐满意度, 我们研究当成员接收到这类推荐时会影响他们决定的因素, 并针对每位成员可能会做出的反应做出一定的权重调整。

##### 3.2.1 基于 TKI 冲突测试的用户容忍度模型

当面对冲突情景时, 在能够接受的范围内, 有些用户会选择容忍, 而超出某个阈值时则会提出抗议, 我们将这个阈值称为用户容忍度。为了得到容忍度, 引入了一种 Thomas-Kilman 冲突模式工具(TKI)。

TKI 冲突测试从武断度(Assertiveness)与协作度(Cooperativeness)两个方面来确定一个用户在面对冲突时的表

现。在这两种表现模式下,根据程度的不同,又细分成5种个性模式来具体描述用户个性,每种个性与武断度和协作度的关联如图2所示。

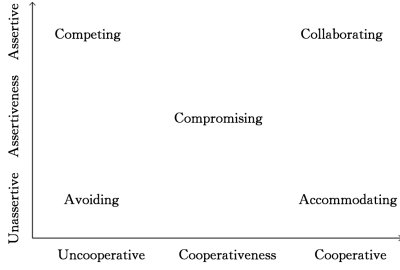


图2 Thomas-Kilmann 冲突模式

Fig. 2 Thomas-Kilmann conflict mode

在利用 TKI 冲突测试计算用户的容忍度时,我们从5种个性模式中选取与容忍度关联最紧密的两种个性:竞争性(C)和迁就性(A)。竞争性用户在面对冲突时,会采取不合作、将自己要求摆在首位、不考虑他人喜好的表现方式;而迁就性用户会采取绝对合作,甚至牺牲自己的喜好来满足群组中其他成员喜好的表现方式。通过主观的 TKI 冲突测试,记录下会选择竞争以及选择迁就的用户数量,并通过它们来计算用户容忍度  $T_{TKI}$  的值:

$$T_{TKI} = 1 - \frac{C - A + 1}{2} \quad (12)$$

其中,  $C \in [0, 1]$ ,  $A \in [0, 1]$ ,  $C$  越大表示用户的竞争性越强,  $A$  越大表示用户的迁就性越强,较小的竞争性和较大的迁就性会导致较高的用户容忍度。

传统的 TKI 冲突测试需要进行大量的主观测试,用户的武断度与协作度无法直接通过在线社交网络获得,因此我们采用一种简化模型来获得子组权重。该模型从用户的容忍度与利他行为两个角度考虑,并且只需要通过用户信息来获取。

### 3.2.2 基于容忍度与利他行为的子组权重的获取

在该简化模型中,我们研究用户社交信息与用户个性之间的关联。首先考虑用户容忍度,分别从用户的社交活跃度与社交影响力两个方面来计算用户容忍指数。利用用户的关注者集合、粉丝集合、注册时间3个因素来得到用户的社交活跃度与社交影响力:

$$X_1 = \frac{|p_{(f_{out}(u_j))}|}{T_{u_j}} \quad (13)$$

$$X_2 = \frac{|f_{r(u_j)}|}{T_{u_j}} \quad (14)$$

其中,  $f_{out}(u_j)$  表示用户  $u_j$  的关注人集合,  $p_{(f_{out}(u_j))}$  表示用户  $u_j$  的关注人的偏好主题特征集合,  $T_{u_j}$  表示用户的注册时间到当前时间的间隔(以天为单位),  $f_{in}(u_j)$  表示用户的粉丝集合,  $f_{r(u_j)} = f_{out}(u_j) \cap f_{in}(u_j)$  表示用户的互粉集合。用户  $u_j$  的关注人的偏好主题特征越少,注册时间越长,代表用户的兴趣越集中,用户的个性也越鲜明;用户的互粉数量越大,代表用户的社交影响力越大,因此,用户的个性也越随和。我们基于这两种变量得到用户  $u_j$  的容忍指数:

$$W_{u_j} = 1 - (\alpha X_1 + \beta X_2 + \gamma) \quad (15)$$

其中,参数  $\alpha < 0, \beta < 0, \gamma > 0$ 。

用户容忍度捕捉的是用户内在个性会表现出的行为,通常,当组成员中有陌生人时,我们考虑用户容忍度。当组成员中有共同朋友时,用户会乐意接受他们朋友喜欢的内容,而不考虑自己的喜好,我们将这种行为称为利他行为。用户  $u_j$  与用户  $u_i$  之间的利他行为指数的计算方式如下:

$$\epsilon_{u_j, u_i} = \frac{|f_{r(u_j)} \cap f_{r(u_i)}|}{|f_{r(u_j)} \cup f_{r(u_i)}|} \quad (16)$$

利他行为指数  $\epsilon_{u_j, u_i}$  越大,表示两个用户更愿意直接接受对方喜爱的内容,因为当两个用户之间共享更多的社会联系(如共同好友)时,他们的社交关系也越紧密。

基于上述得到的容忍指数与利他行为指数,能够获取子组  $SG_x$  的权重,计算方法如下:

$$\lambda_{SG_x} = 1 - \sum_{u_j \in SG_x} \frac{W_{u_j}}{|SG_x|} \sum_{u_i \in SG_x - \{u_j\}} \epsilon_{u_j, u_i} \quad (17)$$

其中,  $|SG_x|$  表示子组  $SG_x$  的规模大小,  $W_{u_j}$  与  $\epsilon_{u_j, u_i}$  都反映了用户能否接受他人喜好的意愿。

下面,我们结合已经获得的子组权重与子组偏好来生成群组偏好。

### 3.3 群组偏好的获取

在获取群组偏好时,首先通过获取群组成员偏好来进行子组划分,减小群组规模,然后基于子组获取群组偏好,这样能够在最大化满足成员偏好的同时有效减少群组偏好的主题特征数量,以便达到缩小群组推荐列表的目的。目标群组  $G$  的偏好计算方法如下:

$$p_G = \sum_{SG_x \in G} \lambda_{SG_x} p_{SG_x} \quad (18)$$

子组偏好  $p_{SG_x}$ , 即为进行划分子组时子组成员的共同偏好。

### 3.4 群组推荐列表的获取

群组推荐列表的获取分为两个步骤:1)综合群组中所有子组的推荐列表  $\sum_{SG_x \in G} \text{top-}k_{SG_x}$ , 得到初始群组推荐列表  $\text{top-}N$ , 该列表虽然能够满足所有成员偏好,但是规模较大,不利于群组做出最后决定,需要对它进行二次筛选;2)通过群组偏好对初始群组推荐列表  $\text{top-}N$  进行筛选,利用上文中跨项目属性相似度的计算公式计算推荐列表中项目属性与群组偏好的相似值,并进行相似度比较,保留相似度高的项目,得到最终的群组推荐列表  $\text{top-}N'$ 。

根据上述算法设计给出算法伪代码,如算法1所示。

#### 算法1 RMSGSB 算法

输入:豆瓣电影评分集 Items, 微博用户数据集 Users

输出:目标群组  $G$  的  $\text{top-}N'$  推荐列表

Begin

1. 将用户随机划分成规模不等的群组;

2. For each  $m_s$  in Items

3. 计算  $R_{m_s}$ ;

4. End for

5. For each  $u_j$  in Users

6. 计算活跃与非活跃用户的主题偏好模型  $p'_{u_j}$ : 活跃用户的主题偏好

模型根据评分信息与式(4)对用户评分进行处理,非活跃用户的主观偏好模型基于外部专家的偏好模型与用户对不同外部专家的偏好程度来获得;

7. End for
8. For each  $u_j$  in Group G
9. 根据式(8)计算组间成员的相似度;
10. If(成员间对主题特征  $g_i$  的相似度  $>$  阈值  $M_1$  & & 成员的总偏好值  $>$  阈值  $M_2$ )
11. 将当前成员划分为一个子组  $SG_x$ ,并将该主题特征  $g_i$  作为当前子组偏好;
12. End for
13. 根据式(10)与式(11)计算子组间的相似度,以此获取相似子组,并得到目标子组的推荐列表  $top-k_{SG_x}$ ;
14. 根据式(15)和式(16)计算组成员的容忍指数和利他行为指数,并通过式(17)计算子组权重;
15. 利用步骤 11 与步骤 14 中获得的子组偏好和子组权重,计算群组偏好;
16. 综合步骤 13 中得到的子组推荐列表与步骤 15 中得到的群组偏好,通过步骤 13 中的跨项目属性相似度计算公式来计算项目的相似度,从而得到最终的群组推荐列表  $top-N'$ ;

End

## 4 实验与分析

### 4.1 数据集描述

本文实验数据集是来自豆瓣用户对电影的评分数据集和新浪微博中的用户集。豆瓣中包含了 2016 年 4 月 23 日至 2018 年 3 月 21 日期间 24120 位用户对 386 部电影的 65525 条评分记录,评分范围为 0~1,每部电影包含了标题、类型、评分、上传时间的相关信息,我们将评分次数大于 10 的用户视为活跃用户,经过数据预处理留下了 1093 位用户的 24563 条评分记录;微博用户集选取了 920 位用户在同时间段内发表的 10825 条微博,其中包含用户转发的微博,每位用户的信息包含了关注人、粉丝以及微博注册时间。

对于外部专家的选取,我们选择通过微博兴趣认证的用户作为外部专家。微博对于兴趣认证用户划分的领域是系统自动进行的,细分领域共有 28 个,而其中有些领域与电影主题特征并无关联,如星座命理、摄影等,经过筛选后,我们留下了 9 个相关领域,分别是科技、历史、军事、萌宠、搞笑幽默、情感、音乐、美食、动漫,并将具有这些领域认证的用户作为外部专家。在实验过程中,分别从豆瓣评分集与微博用户集中随机选取 80% 的用户数据信息作为训练集,剩下的 20% 作为测试集。

### 4.2 评价指标

实验采用精确率(Precision)、召回率(Recall)、F 值(F-measure)来评估算法的性能,具体定义为:

$$Precision = \frac{|E_s|}{|E_p|}, Recall = \frac{|E_s|}{|E_r|} \quad (19)$$

$$F = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (20)$$

其中, $E_s$  代表训练集与测试集的最终推荐列表中同时存在的

项,即预测正确的项; $E_D$  代表测试集中最终推荐列表中的项; $E_r$  代表训练集中最终推荐列表的项。

### 4.3 参数设置与实验对比

在进行实验前,首先对参数  $T_0, \alpha, \beta, \gamma$  进行设置。在设置衰退系数  $T_0$  时,根据艾宾浩斯记忆遗忘曲线,记忆率  $y$  与时间  $x$  (单位为 h) 的关系近似满足  $y = 1 - 0.56x^{0.06}$ ,经过 14 天的反复记忆,我们能够达到长期记忆的目的。当  $\Delta t \leq 14$  时,计算不同的  $T_0$  值,当  $\Delta t > 14$  时,将用户对当前主题特征的偏好视为固定偏好,并保持  $T_0$  不变。 $T_0$  随  $\Delta t$  变化的取值情况如表 1 所列。

表 1  $T_0$  随  $\Delta t$  变化的取值情况  
Table 1 Values of  $T_0$  when  $\Delta t$  changes

| $\Delta t/d$ | $T_0$  | $\Delta t/d$ | $T_0$  |
|--------------|--------|--------------|--------|
| 1            | 0.8833 | 8            | 2.1717 |
| 2            | 1.2932 | 9            | 2.2394 |
| 3            | 1.5544 | 10           | 2.2982 |
| 4            | 1.7414 | 11           | 2.3499 |
| 5            | 1.8844 | 12           | 2.3958 |
| 6            | 1.9986 | 13           | 2.4368 |
| 7            | 2.0926 | 14           | 2.4736 |

下面讨论  $\alpha, \beta, \gamma$  的取值情况。这 3 个变量的取值会影响用户的容忍指数,因此需要在精确率与召回率最优时设置这 3 个变量的取值。通过变化其中单个变量,固定剩余两个变量,来验证实验性能,分别设置  $\alpha = -0.05, \beta = -0.08, \gamma = 0.4$ 。图 3 描述了当 3 个变量的取值各自变化时,对实验结果的影响。图 3(a) 描述了参数  $\alpha$  对实验性能的影响,从中可以发现,当  $\alpha$  取在  $-0.07 \sim -0.04$  之间时,对实验性能的影响较小;而当取值偏小或偏大时,都会对实验性能造成一定的影响。这也说明了参数  $\alpha$  能够影响用户的容忍指数, $\alpha$  取值偏小会导致用户容忍指数偏大,从而导致当前群组中所有子组的权重偏小,群组偏好不全面,出现“欠拟合”现象,因此影响了最终的实验性能;反之  $\alpha$  取值偏大,会导致所有子组的权重都偏大,最终的群组偏好属性过多,导致“噪声”信息过多,从而干扰到最终推荐列表中的项,在这种情况下,实验的性能也受到了一定程度的影响。为了获得最高的 F 值,将  $\alpha$  设置为  $-0.05$ 。

图 3(b) 描述了参数  $\beta$  对实验性能的影响。经过观察发现,曲线的总体趋势与图 3(a) 类似,这是因为参数  $\alpha$  和参数  $\beta$  在整个式子中所占的比重相同, $\alpha$  和  $\beta$  取值偏小会导致子组中用户的容忍度偏大;反之会导致容忍度偏小。具体原因与参数  $\alpha$  相同,此处不再赘述,将参数  $\beta$  设置为  $-0.08$ 。

图 3(c) 描述了参数  $\gamma$  对实验性能的影响,从中可以发现参数  $\gamma$  对实验的性能有较大的影响,主要体现在精确率与召回率的差值上,这也证明了  $\gamma$  的取值对于用户容忍指数会产生较大的偏差。 $\gamma$  过小时,会造成子组权重严重偏小,群组偏好属性单一,从而导致推荐项目的属性单一,这样虽然能保证较高的精确率,但同时也导致了低召回率;而  $\gamma$  过大时,系统将群组中每个成员的所有偏好属性都纳入了群组偏好,导致群组偏好属性的数量过大,推荐列表中的项虽然能较为准确地预测出用户的真实偏好,但同时也包含了很多用户不喜欢的项,这就造成了曲线最后呈现出低精确率、高召回率的状态。

态。根据实验结果,  $\gamma$  的值设置为 0.4。

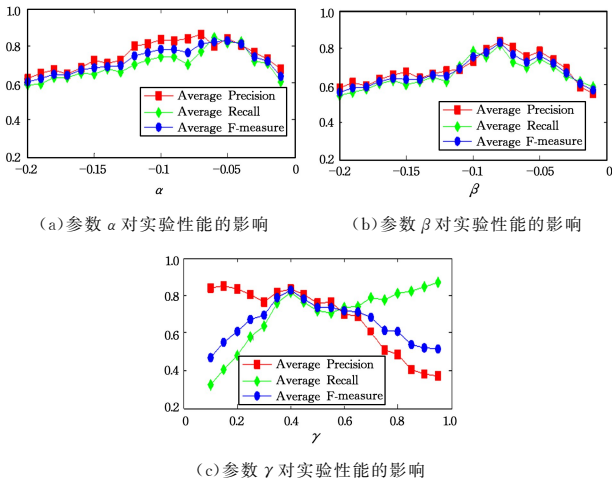


图3 参数对实验性能的影响

Fig. 3 Influences of parameters on experiment performance

为了验证本文算法的有效性,对3种不同算法的结果进行了对比。

1) EDIG w MAJ: 该算法由 Naamani-Dery 等人<sup>[12]</sup> 提出,他们一共提出了4种策略,本文选取了其中获得最好结果的方法,即 EDIG w MAJ。该方法是一种基于多数人偏好的可扩展动态信息增益法。

2) RMSGSB-O: 本文提出的算法中直接将初始群组推荐列表作为最终群组推荐列表,不考虑社会行为对群组偏好的影响。该方法能够验证通过加入社会行为能否有效提高预测的准确性。

3) RMSGSB: 即本文所提算法,将参数设置为  $\alpha = -0.05$ ,  $\beta = -0.08$ ,  $\gamma = 0.4$ 。

在 top-5 情况下,分别计算3种算法的精确率、召回率与 F 值,如图4所示。通过观察可以发现,本文所提算法 RMSGSB 获得的性能优于其他两种算法,值得注意的是, RMSGSB-O 的性能略差于 EDIG w MAJ,这也从侧面证明了引入社会行为中的容忍指数与利他行为指数对预测准确性具有较好的提升效果。

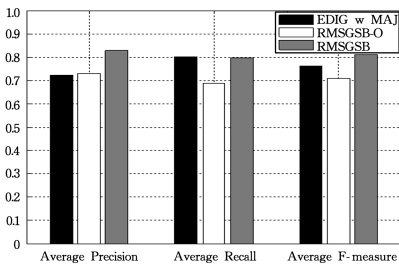


图4 top-5 条件下不同算法的性能对比

Fig. 4 Performance comparison among three algorithms under top-5

图5描述了 top-k 的取值对3种算法中 F 值的影响。从3种算法的整体性能来看,在不同的 top-k 取值下,本文所提方法始终优于其他两种算法;随着 k 的改变, RMSGSB 下的实验性能并未受到太大影响,这是由于群组偏好充分考虑了群组成员的个性,最终推荐列表中的项能最大化满足每位成员的需求,因此 k 的值对本文提出的算法并没有显著影响;此

外,我们可以发现 EDIG w MAJ 的性能较容易受到 k 的影响,过小或过大的 k 都会影响 EDIG w MAJ 最终的准确性,这是由于该算法是基于大部分成员的偏好产生推荐,过小或过大的推荐列表都会使算法优先考虑大部分成员的偏好而忽视少部分成员的偏好,从而导致较差的准确性;而随着 k 的增大, RMSGSB-O 的性能越来越好,与 RMSGSB 的差距也越来越小,这也说明了在不引入社会行为时,推荐列表中项的数量与推荐的性能有很大关联,通过引入社会行为, RMSGSB 减小了群组偏好属性的数量,留下了最具代表性的偏好, RMSGSB-O 中需要 top-15 达到的 F 值在 RMSGSB 中只需要在 top-10 甚至 top-8 就能达到相近的 F 值,这也验证了本文所提的在满足推荐准确性与公平性的前提下最大化缩小群组推荐列表。

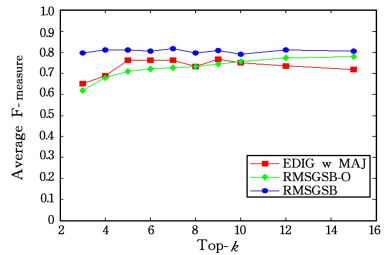


图5 不同 top-k 条件下3种算法的性能对比

Fig. 5 Performance comparison among three algorithms under different top-k

此外,我们还针对群组规模对3种算法的影响进行了对比,如图6所示。从3种算法的总体趋势来看,随着群组规模的增大,3种算法的准确性都越来越差,这与之前的研究<sup>[14-16]</sup> 结论一致。当群组规模较小时,3种算法的性能基本相同,本文算法的优势并不明显。然而随着群组规模的增大, RMSGSB 的优势逐渐凸显,有两个主要原因: 1) RMSGSB 通过分析群组成员的共同偏好来减少群组偏好的属性数量; 2) RMSGSB 从用户的行为出发,分析了用户在面对冲突情境下的容忍指数与利他行为指数,从而进一步过滤多余的偏好,与另外两种算法相比, RMSGSB 算法能够在群组规模较大(如 10)时达到另外两种算法在群组规模较小(如 5)时的准确度。因此,本文所提算法在不同的 top-k、群组规模下的性能均优于另外两种算法。

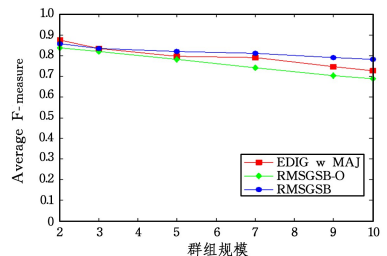


图6 不同群组规模下3种算法的性能对比

Fig. 6 Performance comparison among three algorithms under different group sizes

**结束语** 本文提出了一种缩小群组推荐列表的方法。划分子组旨在最小化群组的偏好属性数量,从而达到缩小群组推荐列表的目的;除了保证尽可能减小推荐列表的规模,我们

还从每个用户的角度考虑了他们的偏好,最大化满足群组成员偏好,从用户固有个性与社会关系方面分析了用户在面对接收到不符合自己偏好的项目时可能会做出的社会行为,根据不同的社会行为为每个子组设置不同的权重,并利用加权模型来获取群组偏好,得到最终的群组推荐列表。基于真实数据集的实验结果表明,本文所提算法能够保证较高的公平性与准确性,当群组规模较大时,优势更为明显。

在计算过程中,由于需要对每位用户进行复杂的分析,从而导致了算法的复杂性高、计算效率低等问题,在今后的工作中,我们将研究如何降低算法的复杂度;在群组推荐中,我们还将研究其他影响用户偏好的因素;另外,组推荐中的用户隐私与安全也是一个需要关注的方面。以上都将是我们下一步的研究内容。

### 参 考 文 献

- [1] O'CONNOR M, DAN C, KONSTAN J A, et al. PolyLens: A Recommender System for Groups of Users[C]// Conference on European Conference on Computer Supported Cooperative Work. Netherlands: Springer, 2001: 199-218.
- [2] FENG S, CAO J. Improving group recommendations via detecting comprehensive correlative information [J]. *Multimedia Tools & Applications*, 2015, 76(1): 1-23.
- [3] SYMEONIDIS P, TIAKAS E, MANOLOPOULOS Y. Product recommendation and rating prediction based on multi-modal social networks[C]// Proceedings of the fifth ACM Conference on Recommender Systems. New York: ACM, 2011: 61-68.
- [4] BAGCI H, KARAGOZ P. Context-Aware Friend Recommendation for Location Based Social Networks using Random Walk [C]// Proceedings of the 25th International Conference Companion on World Wide Web. Canada, 2016: 531-536.
- [5] REN Z, LIANG S, LI P, et al. Social Collaborative Viewpoint Regression with Explainable Recommendations[C]// Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. New York: ACM, 2017: 485-494.
- [6] HONG M, JUNG J J, CAMACHO D. GRSAT: A Novel Method on Group Recommendation by Social Affinity and Trustworthiness[J]. *Cybernetics & Systems*, 2017, 48(3): 140-161.
- [7] YUAN Z, CHEN C. Research on group POIs recommendation fusion of users' gregariousness and activity in LBSN[C]// IEEE International Conference on Cloud Computing and Big Data Analysis. China: IEEE, 2017: 305-310.
- [8] LIN X, ZHANG M, ZHANG Y F, et al. Fairness-Aware Group Recommendation with Pareto-Efficiency [C]// The Eleventh ACM Conference. New York: ACM, 2017: 107-115.
- [9] WANG S, WANG Y, TANG J, et al. What Your Images Reveal: Exploiting Visual Contents for Point-of-Interest Recommendation [C]// Proceedings of the 26th International Conference on World Wide Web. Canada, 2017: 391-400.
- [10] HE R, MCAULEY J. Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering [C]// Proceedings of the 25th International Conference on World Wide Web. Canada, 2016: 507-517.
- [11] LAI C H, HONG P R. Group Recommendation Based on the Analysis of Group Influence and Review Content [C]// Asian Conference on Intelligent Information and Database Systems. Japan: Springer, Cham, 2017: 100-109.
- [12] NAAMANI-DERY L, KALECH M, ROKACH L, et al. Preference elicitation for narrowing the recommended list for groups [C]// Proceedings of the 8th ACM Conference on Recommender systems. New York: ACM, 2014: 333-336.
- [13] HUANG C H, YIN J, HOU F. A text similarity measurement combining word semantic information with TF-IDF method [J]. *Chinese Journal of Computers*, 2011, 34(5): 856-864. (in Chinese)  
黄承慧, 印鉴, 侯昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度度量方法 [J]. *计算机学报*, 2011, 34(5): 856-864.
- [14] ZHANG Y J, DU Y L, MENG X W. Research on group recommendation systems and their application [J]. *Chinese Journal of Computers*, 2016, 39(4): 745-764. (in Chinese)  
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.
- [15] PESSEMIER T, DOOMS S, MARTENS L. Comparison of group recommendation algorithms [M]. Dordrecht: Kluwer Academic Publishers, 2014: 2497-2541.
- [16] KAGITA V R, PUJARI A K, PADMANABHAN V. Virtual user approach for group recommender systems using precedence relations [M]. Tanytown: Elsevier Science Inc, 2015: 15-30.