doi: 10.3969/j.issn.1001 -0505.2013.02.014

社交网络中基于用户投票的推荐机制

刘锡文[1] 蒋俊杰[2]

([1] 东南大学计算机科学与工程学院，南京 211189)
([2] 上海贝尔股份有限公司，上海 201206)

摘要： 为了改善目前社交网络中热点信息推荐与个性化好友推荐的不足，提出基于用户投票的推荐机制. 首先，根据众多用户对某条信息的投票情况评估信息的热度与价值，将用户对信息的浏览、评论、转发等操作以及时间因素与用户主动性投票相结合，提出基于用户投票的热点信息推荐算法. 然后，根据某个用户对众多信息的投票情况评估用户的兴趣，从用户对网络信息的投票以及浏览情况中提取出用户的兴趣度特征，进而提出基于用户投票的个性化好友推荐算法. 最后，针对 2 个算法进行仿真实验，评估各因素对推荐算法的影响和推荐的有效性. 实验结果表明，基于用户投票的推荐机制可以有效地进行热点信息与个性化好友的推荐.

关键词： 社交网络；推荐机制；热点信息；个性化好友
中图分类号： TP393.0 **文献标志码：** A **文章编号：** 1001 -0505(2013)02-0301-06

Recommendation mechanism based on user voting in the social network

Liu Xiwen[1] Jiang Junjie[2]

([1]School of Computer Science and Engineering, Southeast University, Nanjing 211189, China)
([2]Alcatel-Lucent Co., Ltd., Shanghai 201206, China)

Abstract: In order to improve the performance of the hotspot information recommendation and personalized friends recommendation in online social networks, a recommendation mechanism based on user voting is proposed. First, according to a large number of users' voting for a certain message, the heat and value of the message can be evaluated. Then a hotspot information recommendation algorithm is proposed combining users' operation on the information, including browsing, forwarding and commenting, and the time factor. Secondly, according to one user's voting for lots of information, the user's interest feature is extracted. Then a personalized friends recommendation algorithm is proposed. Finally, simulation experiments are performed separately to evaluate the effects of different factors on the validity of the two recommendation algorithms. The results show that the proposed recommendation mechanism based on user voting can work effectively and efficiently.

Key words: social network; recommendation mechanism; hotspot information; personalized friends

社交网络可定义为基于网络的服务. 它允许用户在一个有界限的系统中构建公开或半公开的资料，联系分享连接的用户，观察并详细研究系统中由其他用户所建立的这一系列连接. 不同社交网络中的这些连接的性质和命名方法可能会不同[1]. 交友和维持人际关系是传统社交网络的 2 种作用. 按照传递性理论[2]，社交网络中互为好友关系的用户更容易通过彼此的好友扩展自己的社交圈. 社交网络中的人际关系也是基于这种理论形成并发展起来的，但这种方式使得用户在社交网络中交互的对象大多数是日常生活中的朋友，即人们交互的对象基本上是和他们有强连接关系的人，而忽视了弱连接[3]的作用.

推荐机制是在电子商务领域发展起来的，属于

收稿日期：2012-07-31. 作者简介：刘锡文(1985—)，男，硕士生；蒋俊杰(联系人)，男，博士，高级工程师，Junjie.Jiang@alcatel-sbell.com.cn.
基金项目：教育部科技发展中心网络时代的科技论文快速共享专项研究资助项目(20110092110053).
引文格式：刘锡文，蒋俊杰. 社交网络中基于用户投票的推荐机制[J]. 东南大学学报：自然科学版，2013，43(2)：301-306. [doi: 10.3969/j.issn.1001 -0505.2013.02.014]

信息过滤的一种应用,使用基于物品特点的模型或者基于用户社会环境的模型,来预测用户对那些还没有认真考虑过的物品或社会元素的喜爱程度[1]. 但是目前推荐机制已经不仅仅局限于电子商务领域. 自从首批关于协同过滤的文献[2-3]发表以来推荐机制已经应用到各行各业中,推荐算法是其中最核心和关键的部分,在很大程度上决定了推荐性能的优劣[4].

考虑到目前社交网络中推荐机制的不足,本文将传统的用户对网络信息的浏览、评论、转发等操作以及时间因素与用户主动性投票相结合,研究并设计了基于用户投票的热点信息推荐算法,从而为用户提供热点信息的推荐服务. 另外,将用户对网络信息的投票与浏览情况相结合,确定此用户的兴趣度,进而提出基于用户投票的个性化好友推荐算法,为用户提供个性化的好友推荐服务.

1 基于用户投票的热点信息推荐机制

1.1 基于用户投票的热点信息推荐算法

投票是用户根据其对信息的理解表示赞同或反对的操作. 设 W_{score} 为根据用户投票情况计算出的信息投票得分. 用户可以投2种票,即赞成票和反对票. 信息投票得分最直观的计算方式有2种:① 得分 = 赞成票数 - 反对票数;② 得分 = 赞成票数/总票数. 但是这2种方式都有缺陷.

第1种方式没有考虑赞成票占总票数的比重. 第2种方式在总票数足够大的情况下是正确的,但在总票数很少时会出现误差,因此需要解决小样本情况下好评率准确性的问题. 1927年Wilson提出了一个被称为Wilson score interval的修正公式[5-9],很好地解决了小样本准确性的问题,即

$$\frac{\hat{p}+\frac{1}{2n}z_{1-\alpha/2}^2 \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}+\frac{z_{1-\alpha/2}^2}{4n^2}}}{1+\frac{1}{n}z_{1-\alpha/2}^2} \tag{1}$$

式中,$\hat{p}$ 表示样本的赞成票比例;n 表示样本数量;$z_{1-\alpha/2}$ 表示对应某个置信水平的统计量,是一个常数,这里取在95%的置信水平下,z 的统计量的值为1.96. 然后根据式(1)的下限值来计算信息投票得分,即

$$\frac{\hat{p}+\frac{1}{2n}z_{1-\alpha/2}^2 - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}+\frac{z_{1-\alpha/2}^2}{4n^2}}}{1+\frac{1}{n}z_{1-\alpha/2}^2} \tag{2}$$

当 n 足够大时,式(2)趋于 $\hat{p}$. 由于式(2)计算出的得分是一个(0,1)之间的数,将其值乘以100使之成为一个实际的信息得分后再赋给 W_{score}. 根据式(2),并考虑用户对信息的浏览、评论和转发操作以及时间因素,定义基于用户投票的热点信息推荐算法的计算公式为

$$R=\frac{\log_{10}W_{\text{views}}+W_{\text{score}}(W_{\text{fors}}+1)+\frac{W_{\text{coms}}}{k}}{(W_{\text{age}}+1)^G} \tag{3}$$

式中,R 表示信息整体得分;W_{views} 为信息的浏览次数,$\log_{10}W_{\text{views}}$ 表示信息浏览次数对信息整体得分的影响,采用以10为底的对数是为了使前10个浏览者的权重和后90个浏览者的权重一致. W_{fors} 为信息的转发次数,$W_{\text{score}}(W_{\text{fors}}+1)$ 对信息整体得分有决定性影响,因为转发次数本身就代表信息的传播特性,信息传播得越远越广,在一定程度上反映出信息越有价值,将其次数加1是为了使用户转发次数为零时不对信息整体得分有影响. 信息投票得分是随着转发次数的增多而成倍增长的,但是如果没人投票,那么即使有再多的人转发,这一项的值仍是零,这在一定程度上防止了某些无信息量的信息成为热点信息. W_{coms} 为信息的评论次数,W_{coms}/k 表示评论次数对信息整体得分的影响,这里的评论次数是净评论次数,这样可避免用户围绕一条信息进行聊天从而增加评论次数,进而使整体得分升高这种情况的发生. 参数 k 起到拉低评论次数对整体得分影响的作用,因为评论次数多的信息不一定有价值,有可能存在名人效应或者只是充当朋友们之间的聊天工具等. W_{age} 为距离信息发布的时间,$(W_{\text{age}}+1)^G$ 中的参数 G 为重力因子,即将信息整体得分往下拉的力量. 随着时间增长,信息整体得分下降,加1是考虑到信息的传播需要一定的时间,同时也保证整个公式的分母不为零.

1.2 仿真实验

1.2.1 信息浏览次数的影响

本节只考虑信息浏览次数对信息整体得分的影响. 首先将信息投票得分、转发次数和评论次数均置为零. 然后分析随着时间增长不同的浏览次数对信息整体得分的影响. 信息的浏览次数 W_{views} 分别取10,100,1 000,10 000,实验结果如图1所示.

由图1可知,信息的浏览次数越多,R 值越大,信息的排名就越靠前,但是随着信息浏览次数的增多,信息整体得分并没有很大的变化,这就保证了所有浏览者的权重是一致的,即浏览次数越多对信息整体得分的影响越小. 另外,随着时间的增长,R

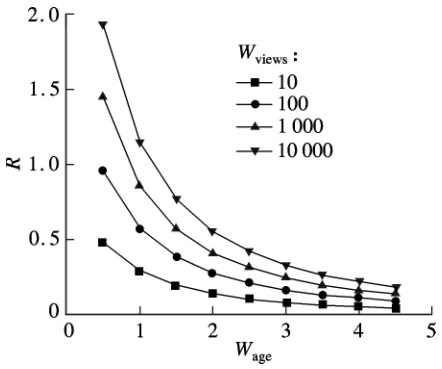


图 1　信息浏览次数对信息整体得分的影响

值在逐渐降低,即热点信息具有时效性.

1.2.2　信息评论次数的影响

对一条信息来说,其浏览次数大于等于评论次数. 如果要分析评论次数对信息排名的影响,就要先确定浏览次数的值,然后评论次数取一系列小于等于浏览次数的值. 这里 W_{views} **取**10 000,W_{coms} **分别取** 10,100,1 000,10 000. **实验结果见图** 2.

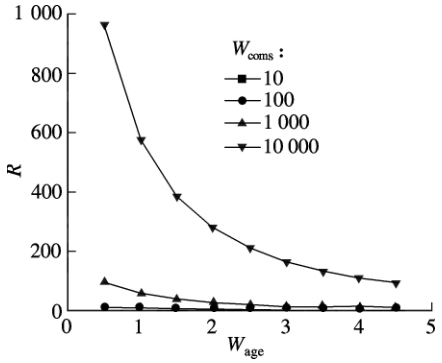


图 2　信息评论次数对信息整体得分的影响

由图 2 **可知,信息整体得分随着评论次数的增多逐渐升高,并且评论次数越多,对信息整体得分的影响越大,当** W_{coms} **为** 1 000 **和** 10 000 **时,**R **值约有** 10 **倍的差距. 另外,随时间增加,**R **值也在逐渐降低,即使** R **值再大,经过一段时间后也会变得很小,这再次说明了热点信息具有时效性. 通过对比图** 1 **和图** 2 **可知,在同等条件下,信息评论次数对信息整体得分的影响远大于信息浏览次数对信息整体得分的影响,因此当用户只对信息进行浏览和评论操作时,信息的评论次数对信息整体得分起主导作用.**

1.2.3　信息投票得分及信息转发次数的影响

本节分析不同信息投票得分和转发次数对信息整体得分的影响. 由于信息的浏览次数大于等于转发次数,所以取 $W_{views}=10\ 000$,$W_{coms}=0$,W_{fors} **分别取** 10,100,1 000 **和** 10 000,**信息投票得分** W_{score} **分别取** 10,25,50 **和** 75,**实验结果如图** 3 **所示.**

由图 3 **可知,在信息投票得分确定的情况下,信息整体得分随着转发次数的增多而升高,并且转发的次数越多对信息整体得分的影响越大. 信息投**

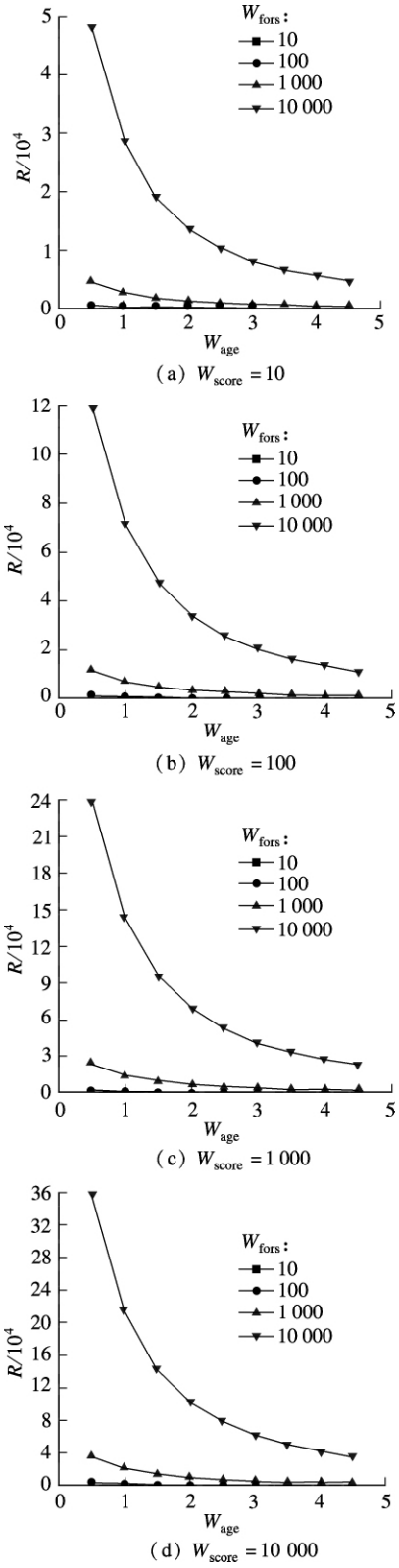


图 3　信息投票得分和转发次数对信息整体得分的影响

票得分越高,整体得分就越高,排名就越靠前,并且信息整体得分几乎是随着投票得分成倍增长的,这就说明当信息投票得分和信息转发次数达到一定程度时,这 2 个因素在算法中起主导作用. 而且通过与图 1 和图 2 作对比,也会发现这 2 个因素对信息整体得分的影响要大于同等条件下浏览次数和评论次数对信息整体得分的影响.

1.2.4 算法有效性验证

本节通过新浪微博提供的 API 提取相关数据来验证算法的有效性. 首先提取用户 ID,然后通过用户 ID 取得其所发布的微博 ID,进而取得相应的评论数和转发数. 所取得的微博 ID 均来自不同用户,以保证数据的有效性. 1.2.3 节已经说明信息的转发次数和投票得分对信息整体得分起主导作用,但也要保证热点信息的质量,因为信息整体得分很高也可能是转发次数很多但投票得分不高引起的,这类信息不应是热点信息. 此算法的目的是给用户推荐有价值的热点信息,所以将提取到的数据代入到式(3)中,并且通过赋予不同的投票得分来验证算法的有效性. 取 1 000 条微博的评论数和转发数,信息的投票得分 W_{score} 取 10,实验结果如图 4(a)所示.

由图 4(a)可知,大部分信息的转发次数很少,整体得分不高,这也符合实际情况. 只有少部分信息的转发次数很多,这些信息才有可能成为热点信息. 当信息的投票得分为 10 时,这些信息的整体得分非常高,系统就会推荐这些信息给用户作为热点信息. 当信息的投票得分为 20 和 50 时,其整体得分也得到提高,如图 4(b)、(c)所示.

信息的转发次数越多表示信息传播得越远越广,用户的参与度越高; 信息投票得分越高表示信息的质量越高,越有价值. 因此,由于这 2 个因素都高而导致的整体得分高的信息将作为热点信息被推荐.

2 基于用户投票的个性化信息与个性化好友推荐机制

2.1 基于用户投票的个性化好友推荐算法

在进行个性化好友推荐之前首先对信息进行简单的分类,并在此基础上给出兴趣度的概念,即用户对某类信息感兴趣的程度. 要想知道用户的兴趣度,就要挖掘用户对信息的一系列操作,而评论和转发操作都是在用户浏览信息的基础上进行的,用户对信息进行浏览、评论、转发就说明其可能对

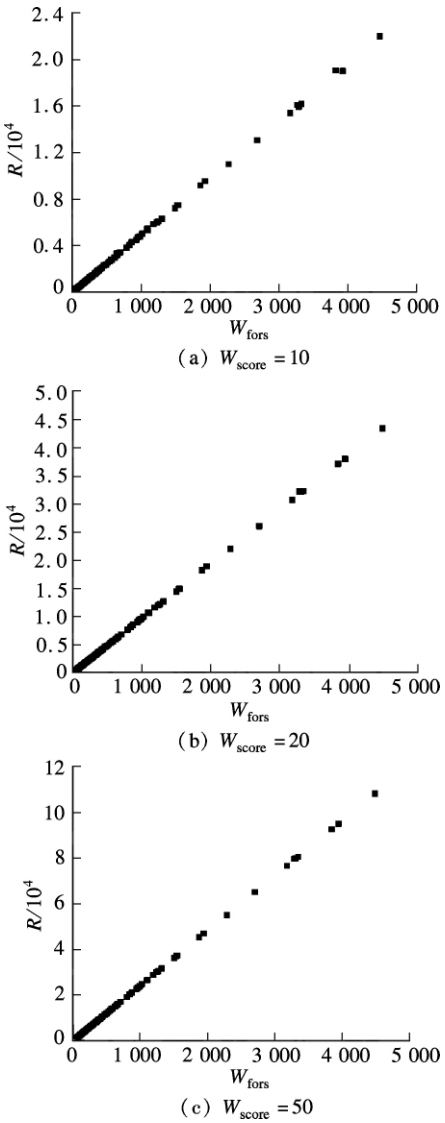


图 4 热点信息分布

某类信息感兴趣. 但是这些操作都不能确定用户是否真的对这些信息感兴趣,即无法确定其可靠性,这时就要结合用户的主动性投票来确定用户到底对哪些信息感兴趣. 兴趣度计算公式如下:

$$I = aC_{\text{views}}(1 + C_{\text{votes}}) \tag{4}$$

式中,C_{views} 为用户对某类信息的浏览次数,aC_{views} 表示用户对某类信息的浏览次数对兴趣度的影响,系数 $a \in (0,1)$,起到降低浏览次数对兴趣度影响的作用,因为用户浏览的信息不一定是感兴趣的信息; C_{votes} 为用户对某类信息投赞成票的次数,$aC_{\text{views}}(1 + C_{\text{votes}})$ 表示用户对某类信息的浏览和投票情况对兴趣度的影响. 如果用户投票说明已经浏览过该信息(即用户对某类信息投赞成票的次数要小于等于浏览次数,$C_{\text{votes}} \leqslant C_{\text{views}}$),所以兴趣度会随着投票次数的增加而成倍增长,而 $C_{\text{votes}} = 0$ 时则只有浏览次数的多少表示兴趣度的大小,因此兴

趣度会较低. 另外,考虑到用户的兴趣度会发生变化,通过式(4)计算出来的是一段时间内的兴趣度. 在用户不活跃时计算其前一段时间内的兴趣度,以适应用户的需求.

基于兴趣度的概念,系统可以提供个性化的好友推荐服务. 首先通过用户的主动性投票确定用户的兴趣度,然后根据用户的兴趣度分析对某类信息感兴趣的其他用户,如果其他用户的兴趣度和当前用户很接近,说明他们感兴趣的信息也相近,因此系统就会推荐这些用户给当前用户.

2.2　仿真实验

2.2.1　用户对某类信息的浏览次数和投票情况对兴趣度的影响

本节根据式(4)分析用户对某类信息的投票以及浏览次数对兴趣度的影响,这里参数 a 取 0.8. 兴趣度是根据用户在一段时间内对各类信息的投票和浏览情况得出的. 当用户不活跃时,再计算前一段时间内的兴趣度,以适应用户兴趣的变化. 用户在一段时间内对某类信息的浏览次数 C_{views} 取 0 ~ 100,赞成票数量小于等于浏览次数,所以 C_{votes} 分别取 0,30,50 和 80,据此来分析这 2 个因素对兴趣度的影响. 实验结果如图 5 所示.

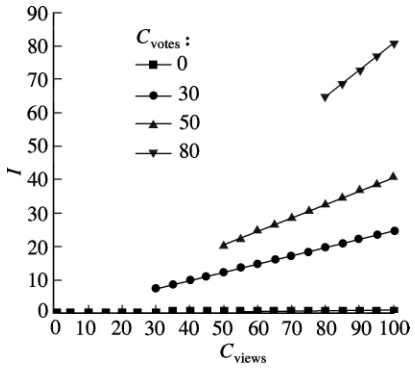


图 5　用户对某类信息浏览次数和投票情况对兴趣度的影响

由图 5 可知,随着用户对某类信息浏览次数的增加,兴趣度逐渐增大,并且投票得分越高,兴趣度增加得越快. 如果用户只是浏览某类信息而没有进行投票,那么兴趣度会非常低,说明用户对这类信息没有什么兴趣. 用户对某类信息的投票情况恰恰反映了其对这类信息感兴趣的程度,投票得分越高说明用户对这类信息越感兴趣,那么相应的兴趣度也就越高. 这说明兴趣度的大小真实地反映了用户对某类信息感兴趣的程度,从而也就证明了以兴趣度为基础的个性化好友推荐的合理性.

2.2.2　算法有效性验证

为了验证算法的有效性,引入 MovieLens[10] 数据集,在此数据集中,用户对其看过的电影进行评分,分值为 1 ~ 5. 本实验使用 MovieLens 中的小规模数据库,其中包含 943 个独立用户对 1 682部电影的 1.0×10^5 次评分,这 1 682 部电影被分为 19 个类别,每部电影属于一个或多个类别. 该库中有 5 组数据: u1. base 和 u1. test, u2. base 和 u2. test, …, u5. base 和 u5. test. 这些数据都是 80% 作为训练集,20% 作为测试集. 首先将式(4)应用于训练集,调整参数 a,以便获得更好的算法推荐值与训练集中用户打分值的拟合度. 因此,必须对数据进行处理,同时为了适应本文提出的个性化好友推荐算法,把得分为 4 分和 5 分的电影认为是用户会投赞成票的信息. 然后分别统计出 u1. base, u2. base, …, u5. base 中各用户看过的电影总数量和得分为4分和5分的电影数量以及在各个类别下用户看过的电影总数量和得分为 4 分和 5 分的电影数量. 根据 5 组数据中用户看过的得分为 4 分和 5 分电影数量占用户看过电影总数量的比例的平均值来确定式(4)中的参数 a,实验数据如表 1 所示. 其中,n_{vote} 表示用户看过的得分为 4 分和 5 分的电影数量,n_{sum} 表示用户看过的电影总数量.

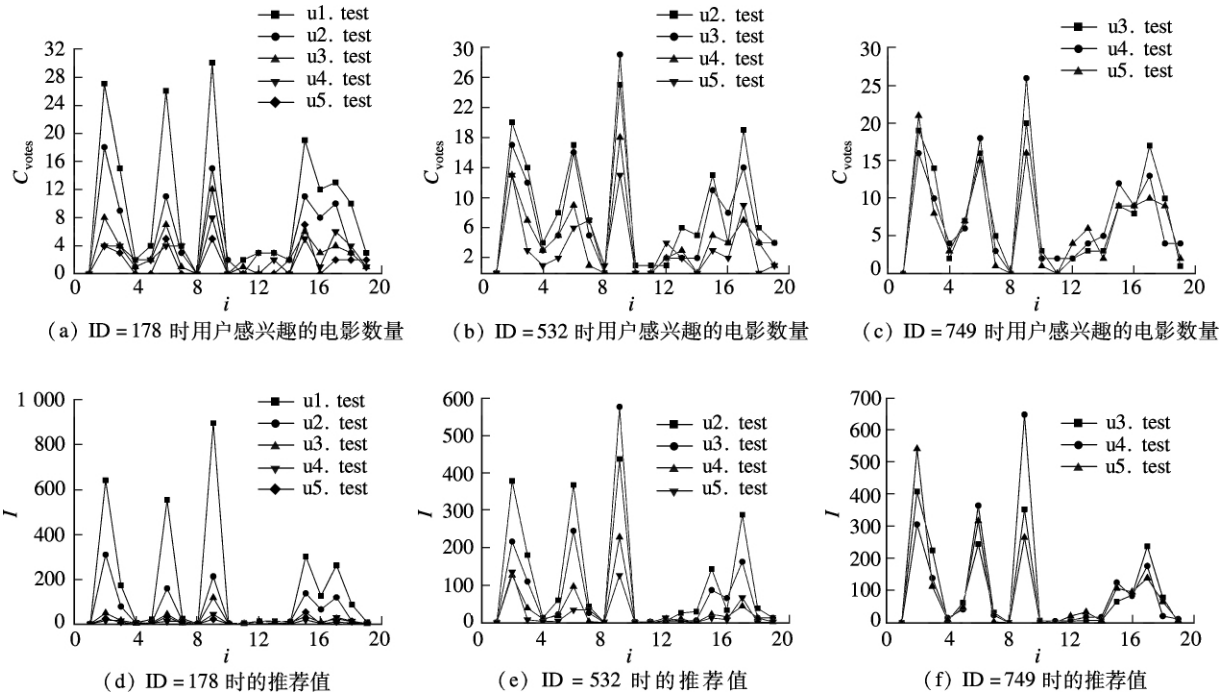
表 1　用户感兴趣的电影数量占看过电影总数量的比例

数据	u1	u2	u3	u4	u5	平均值
$n_{\text{vote}}/n_{\text{sum}}$	0.577	0.578	0.577	0.578	0.579	0.578

由表 1 可知,这个比例的平均值为 0.578,因此把式(4)中的参数 a 取为 0.6.

下面把算法应用于测试集,在 943 个独立用户中随机抽取 3 个,通过比较式(4)计算出来的兴趣度和用户确实感兴趣的电影数量来验证算法的有效性. 实验结果如图 6 所示.

图 6 中的用户 ID 分别为 178,532,749,横坐标 i 为电影类别. 通过对比图 6(a) ~ (c) 与图 6(d) ~ (f) 可知,根据式(4)计算出来的兴趣度的大小和用户感兴趣的电影数量的整体趋势基本一致,这说明了式(4)的有效性,而个性化的好友推荐算法都是基于式(4)的,从而也就说明了基于用户投票的个性化好友推荐算法的有效性.



(a) ID = 178 时用户感兴趣的电影数量 (b) ID = 532 时用户感兴趣的电影数量 (c) ID = 749 时用户感兴趣的电影数量

(d) ID = 178 时的推荐值 (e) ID = 532 时的推荐值 (f) ID = 749 时的推荐值

图 6 用户感兴趣的电影数量与兴趣度的对比

3 结语

本文主要针对目前社交网络中推荐机制的不足,提出了基于用户投票的热点信息推荐机制以及个性化好友推荐机制. 众多用户对某条信息的投票情况反映了此信息的热度与价值,某个用户对众多信息的投票情况反映了此用户的兴趣. 针对 2 个算法进行仿真实验,评估各因素对推荐算法的影响以及推荐的有效性. 实验结果表明,本文提出的基于用户投票的推荐机制可以有效地进行热点性与个性化好友的推荐. 接下来将进一步研究社交网络中推荐机制的合理性和有效性.

参考文献 (References)

[1] Boyd D M, Ellison N B. Social network sites: definition, history, and scholarship [J]. *Journal of Computer-Mediated Communication*, 2008, **13**(1): 210-230.

[2] Snijders T A B, Bunt G G V, Steglich C E G. Introduction to stochastic actor-based models for network dynamics [J]. *Social Networks*, 2010, **32**(1): 44-60.

[3] Granovetter M S. The strength of weak ties [J]. *American Journal of Sociology*, 1973, **78**(6): 1360-1380.

[4] Resnick P, Varian H R. Recommender systems [J]. *Communications of the ACM*, 1997, **40**(3): 56-58.

[5] Balabanović M, Shoham Y. Fab: content-based, collaborative recommendation [J]. *Communications of the ACM*, 1997, **40**(3): 66-72.

[6] Schafer J B, Konstan J, Riedl J. Recommender systems in e-commerce [C] //*Proceedings of the 1st ACM Conference on Electronic Commerce*. New York, 1999: 158-166.

[7] 徐海玲, 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.
Xu Hailing, Wu Xiao, Li Xiaodong, et al. Comparison study of Internet recommendation system [J]. *Journal of Software*, 2009, **20**(2): 350-362. (in Chinese)

[8] Wikipedia. Wilson score interval [EB/OL]. (2011-11-03) [2012-06-31]. http://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval.

[9] Wilson E B. Probable inference, the law of succession, and statistical inference [J]. *Journal of the American Statistical Association*, 1927, **22**(158): 209-212.

[10] GroupLens Research. MovieLens data sets [EB/OL]. (2011-08-24) [2012-06-31]. http://www.grouplens.org/node/73#attachments.