

在线用户声誉评价算法的鲁棒性分析

李圣楠¹, 杨 凯¹, 刘晓露¹, 刘建国^{1,2}, 郭 强¹

(1. 上海理工大学 复杂系统科学研究中心, 上海 200093; 2. 上海财经大学 科研实验中心, 上海 200433)

摘要: 在线评分系统中的恶意或随机打分为准确评价在线用户声誉带来了极大的挑战. 对3种基于迭代的经典在线用户声誉评价算法的鲁棒性进行了细致研究. 实验先将不同数量用户打分随机化, 再以均方根误差为指标衡量其余用户声誉值受影响程度. 实验共在3个数据集中进行, 在MovieLens和Netflix两个经典实证数据集上的实验结果表明: 系统中1%~60%的用户进行随机打分时, 基于关联分析的CR算法始终保持很好的鲁棒性; 基于打分迭代的IARR算法的均方根误差略有增大, 最大值达到0.22, 但整体波动较小; 而改进的基于打分迭代的IARR2算法的均方根误差最大值达到0.695, 其鲁棒性的较大波动是因算法受高声誉用户的影响较大. 在Douban数据集上的结果表明: 在打分数据稀疏情况下, CR算法也能保持很好的鲁棒性.

关键词: 在线打分系统; 用户声誉; 鲁棒性

中图分类号: N 949 **文献标志码:** A

Robustness Analysis of Online User Reputation Measurements

LI Shengnan¹, YANG Kai¹, LIU Xiaolu¹, LIU Jianguo^{1,2}, GUO Qiang¹

(1. Research Center for Complex Systems Science, University for Shanghai for Science and Technology, Shanghai 200093, China; 2. Laboratory Centre, Shanghai University of Finance and Economics, Shanghai 200433, China)

Abstract: Malicious and spam actions in online rating systems affect the user reputation measurements greatly. By setting different number of spammers and evaluating its effect by the root-mean-square error (RMSE), the robustness of three typical iterative-oriented online user reputation measurements was investigated. The results for MovieLens and Netflix data sets show that when facing with 1%~60% of spammers in the network, the CR algorithm has the best performance of robustness. The largest RMSE value of the iterative algorithm of reputation IARR reaches 0.22, with slight fluctuation of the RMSE. And the RMSE value of the improved iterative algorithm of reputation IARR2 reaches 0.695. The result for Douban data set shows that the CR algorithm still maintains great robustness even when the users rate few common items.

Keywords: online rating system; user reputation; robustness

收稿日期: 2015-08-22

基金项目: 国家自然科学基金资助项目(71271126, 71374177, 61361125); 教育部博士点基金资助项目(20120078110002); 上海市东方学者特聘教授, 上海市曙光学者(14SG42)

第一作者: 李圣楠(1995-), 女, 硕士研究生. 研究方向: 社会网络分析. E-mail: lishengnan318@163.com

通信作者: 刘建国(1979-), 男, 教授. 研究方向: 网络科学. E-mail: liujg004@ustc.edu.cn

在互联网大数据的背景下,电子商务、互联网金融、在线产品打分等诸多用户行为越来越多地融入网络数据中^[1].一方面,在线系统向用户提供大量商品以供选择,包括电影、书籍、网页等^[2];另一方面,用户可以在系统中对商品进行在线打分.近年来,在线用户声誉评价算法引起人们越来越多的关注^[3],Zhou等^[4]提出了基于皮尔森相关系数的迭代算法(简称CR算法).该算法引入量化评价两变量相关程度的皮尔森相关系数,使打分和产品质量相似度越高的用户得到的声誉值也越高.在此基础上,考虑到系统中乱打分用户的存在,Liao等^[5]提出了声誉值再分配迭代算法(简称IARR算法)和引入惩罚因子的声誉值再分配迭代算法(简称IARR2算法).两种算法增强了系统中高声誉或打分产品多的用户的影响力,因为这部分用户常被认为是系统中的可信用户.

如今,各种在线评价系统中不乏随意打分、恶意打分的用户^[6],因此建立准确的用户声誉评定体系,可以提高产品综合评分的准确性^[7].然而,网络中普遍存在的乱打分现象^[8],对声誉系统的度量造成了一定干扰.因此,对声誉评价模型效果的全面分析就显得尤为重要.传统对模型的评价,多从结果准确性角度展开^[4-5].然而,在准确性的评价中,一方面,真实网络中,产品的实际质量一般未知,同时标准产品的界定也存在一定困难;另一方面,已有指标很难准确反映模型对系统中普遍存在乱打分用户这一现象的抗干扰能力.不同真实网络中,算法应对乱打分干扰的鲁棒性,对系统能否准确评价用户声誉具有重要意义.基于以上思想,本文以均方根误差为指标,分析了系统中不同数量用户的随机打分为对其余用户(以下简称诚实用户)声誉度量的影响,从而研究经典声誉评价算法的鲁棒性.

本文对基于皮尔森相关系数的迭代算法、声誉值再分配迭代算法及引入惩罚因子的改进迭代算法的鲁棒性进行了实证分析.在MovieLens和Netflix两个经典数据集上的实验结果一致表明,在乱打分用户数量不断增加过程中,CR算法始终保持较好的鲁棒性,IARR算法也相对稳定,而IARR2算法受干扰较大.对IARR2算法鲁棒性影响因素的分析结果表明,IARR2算法的鲁棒性受声誉值极大用户的影响较大.在Douban数据集上的实验结果表明,即便在打分数据稀疏情况下,CR算法也能保持很好的鲁棒性.

1 在线用户声誉评价算法

本文考虑的打分系统是二分网络^[9],分为表示用户的集合 U 和表示产品的集合 O . $r_{i\alpha}$ 代表用户 i 对产品 α 的打分. k_i 和 k_α 分别表示用户度和产品度.此外,所有用户 i 打分的产品定义为集合 O_i ,所有给产品 α 打分的用户定义为集合 U_α . Q_α 和 R_i 分别表示产品 α 的质量和用户 i 的声誉.用户 i 的初始声誉值设定为 $R_i = \frac{k_i}{M}$ (M 指产品总数).所以,基于用户打分的产品加权平均质量值表示为

$$Q_\alpha = \frac{\sum_{i \in U_\alpha} R_i r_{i\alpha}}{\sum_{i \in U_\alpha} R_i} \quad (1)$$

1.1 CR 算法

在CR算法中^[4],根据式(1)得到产品 α 的质量 Q_α ,然后计算用户 i 的打分和其对应产品质量间的皮尔森相关系数.其公式表示为

$$\text{Corr}_i = \frac{1}{k_i} \sum_{\alpha \in O_i} \left(\frac{r_{i\alpha} - \bar{r}_i}{\sigma_{r_i}} \right) \left(\frac{Q_\alpha - \bar{Q}_i}{\sigma_{Q_i}} \right) \quad (2)$$

式中: σ_{r_i} 和 σ_{Q_i} 分别表示用户 i 打分变量的标准差和其所有打分产品质量变量的标准差; $\bar{r}_i$ 和 $\bar{Q}_i$ 则分别是用户 i 打分变量和其所有打分产品质量变量的平均值.

用户 i 的声誉值 R_i 为

$$R_i = \begin{cases} \text{Corr}_i, & \text{Corr}_i \geq 0 \\ 0, & \text{Corr}_i < 0 \end{cases} \quad (3)$$

根据更新的用户声誉值,按式(1)重新计算所有产品质量^[10].模型迭代过程停止于所有产品前后两次质量值的均方差(见式(4))小于设定的阈值 Δ (本文MovieLens数据集中 $\Delta = 10^{-5}$;Netflix中 $\Delta = 6 \times 10^{-3}$;Douban中 $\Delta = 0.1$).

$$\Delta = |Q - Q'| = \frac{1}{M} \sum_{l=0}^L (Q_l - Q'_l)^2 \quad (4)$$

1.2 IARR 算法

IARR算法^[5]是基于CR算法建立的,将每次迭代中式(3)得到的用户 i 的声誉值记为 TR_i ,然后按照式(5)对用户 i 声誉值进行再分配.最后根据得到的新声誉值,重新计算产品质量,进而循环迭代.

$$R_i = TR_i^\theta \frac{\sum_{j \in U} TR_j}{\sum_{j \in U} TR_j^\theta} \quad (5)$$

式中, θ 为算法参数,当 $\theta = 1$ 时算法退化为经典的

用户声誉度量方法. IARR 算法加入用户声誉再分配过程, 目的在于放大高声誉用户的声誉值. 因为高声誉的用户常被认为是系统中不存在乱打分为的可信用户, 增加这部分用户的影响力, 可以在一定程度上削减迭代过程中噪声数据的干扰, 提高产品质量评分和用户声誉评定的准确性.

1.3 IARR2 算法

通常, 只给很少量产品打分的用户, 不应获得很高的声誉值. 因为用户偶然猜对产品在大众评分下的质量很容易, 但大量打分都能吻合就相对困难. 因此, 只有用户打分数据较充足时, 其高声誉才可信. 同样地, 从产品角度, 通过很少数用户打分就获得高质量的评定, 这一评价结果是很武断的.

IARR2 算法基于上述对用户度和产品度的考虑^[5], 除了沿用前两种算法的迭代过程, 以及 IARR 算法的用户声誉再分配过程外, 引入了两个惩罚因子分别改进式(1)和式(2). 修改后的公式为

$$Q_a = \max_{i \in U_a} \{R_i\} \frac{\sum_{i \in U_a} R_i r_{ia}}{\sum_{i \in U_a} R_i} \quad (6)$$

$$Corr_i = \frac{\lg(k_i)}{\max\{\lg(k_i)\}} \cdot \frac{1}{k_i} \cdot \sum_{a \in O_i} \left(\frac{r_{ia} - \bar{r}_i}{\sigma_{r_i}} \right) \left(\frac{Q_a - \bar{Q}_i}{\sigma_{Q_i}} \right) \quad (7)$$

2 声誉评价算法的鲁棒性分析

以往对声誉评价算法结果的分析多从准确性角度展开, 并且常采用肯德尔系数(简称 τ 值)^[11] 和 AUC 算法^[12]. 本文则从鲁棒性角度, 以均方根误差值(简称 $RMSE$ 值)为指标, 分析声誉评价算法对网络中随机打分用户情况的抗干扰能力, 算法的参数 θ 设定为 3.

2.1 鲁棒性分析方法

首先, 假定系统中不存在乱打分用户, 并将原数据集通过算法得到的用户 i 声誉值记为原始声誉值 R_i . 然后, 根据不同数据集中用户总数 N 的不同, 按用户编号顺序选择多组不同数量 n 的用户假定为系统中乱打分的用户, 即在每次实验中将选定的编号前 n 个用户的所有打分随机化, 得到新的数据集. 使用之前相应的算法对新数据集重新进行计算, 得到新的用户 i 声誉值 R'_i . 所得 R'_i 中, 前 n 个数值是假定乱打分用户的声誉, 其余 $N - n$ 个数据是系统中打分未改变用户(以下简称诚实用户)的声誉

值. 最后统计出诚实用户声誉值与其相对应的原始声誉值两组变量间的均方根误差, 即 $RMSE =$

$\sqrt{\frac{1}{N-n} \sum_{i=1}^{N-n} (R'_i - R_i)^2}$. $RMSE$ 值越小, 说明在网络中一定数量乱打分用户出现前后算法对诚实用户声誉评价结果变化越小, 评价结果越稳定, 算法的抗干扰能力越强.

2.2 实证分析数据集

本文的研究选择了 MovieLens 和 Netflix 两个包含用户对电影在线打分记录的经典数据集, 以及包括用户对图书在线打分记录的 Douban 数据集. 实验选用的 MovieLens 数据集包含 943 个用户对 1 682 个电影的 100 000 条打分记录. Netflix 数据集规模相对很大, 实验中选出其中打分电影不少于 20 个的 5 000 名用户的数据作为实验数据集. Douban 数据集稀疏性很大, 因此本文也选取打分不少于 20 个的 40 038 个用户的所有记录作为实验数据集. 同时, 用户的打分都是 1~5 的整数. 表 1 描述了数据集的一些基本特性.

表 1 实验数据集的基本特性

Tab.1 Properties of the applied data sets

数据集	用户数	产品数	打分数量	用户平均度	产品平均度
MovieLens	943	1 682	100 000	106	59
Netflix	5 000	17 770	894 482	179	50
Douban	40 038	78 094	1 784 943	45	23

2.3 声誉评价算法的鲁棒性分析

将 3 种在线用户声誉评价算法分别在 3 个数据集上按上述分析方法进行实验. 在不同数据集中逐次设定随机化用户数 n 从 1%~60% 为随机打分用户. 进而, 对比随机打分前后算法得到的均方根误差 $RMSE$ 的值. 在 MovieLens 数据集中, 共实验 50 次 ($n = 10, 20, 30, \dots, 500$); 在 Netflix 数据集中, 共实验 31 次 ($n = 50, 100, 200, 300, \dots, 3\ 000$); 在 Douban 数据集中, 共实验 11 次 ($n = 4\ 000, 6\ 000, 8\ 000, \dots, 24\ 000$). 实验结果如图 1 所示.

图 1 中横坐标 n 为每次实验假定的乱打分用户数量, 纵坐标 $RMSE$ 为网络中 n 个用户乱打分状态下其余诚实用户声誉值与各自原始声誉值两组变量的均方根误差. 图中结果均为 5 次独立实验的平均值.

从图 1(a) 和 (b) 可见, 在 MovieLens 和 Netflix 两个经典数据集中, 随着系统中乱打分用户的数量不断增加, CR 和 IARR 算法的 $RMSE$ 整体上呈现缓慢上升趋势, 即随着乱打分用户数量的增加, 算法对诚实用户声誉计算结果的变化略微增大. CR 算法

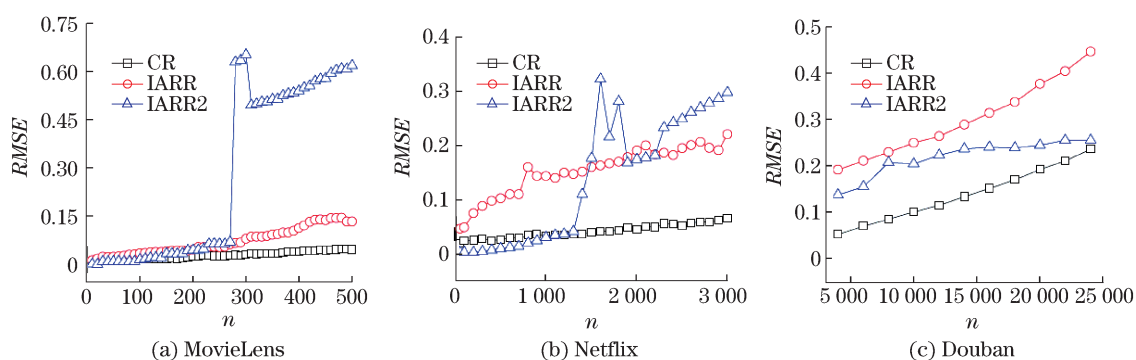


图1 MovieLens, Netflix 和 Douban 数据集中 CR, IARR, IARR2 算法的鲁棒性分析结果

Fig.1 Robustness analysis of CR, IARR and IARR2 algorithms in MovieLens, Netflix and Douban data sets

的 $RMSE$ 值在 MovieLens 数据集中最大值为 0.048 6, 在 Netflix 中最大值也仅为 0.064 3. 因此, 在 CR 算法中, 乱打分用户的出现对系统中诚实用户声誉评价的干扰较小, 即 CR 算法在应对网络中不确定数量用户乱打分情况时, 声誉评价结果相对稳定. 与 CR 算法相比结果变化较大的 IARR 算法, 在两个不同数据集下的 $RMSE$ 最大值也只在 0.2 左右, 并且波动较小. 所以, IARR 算法的鲁棒性也相对较好. 然而, 对于实验结果显著波动的 IARR2 算法, 其 $RMSE$ 最大值在 Netflix 数据集上接近 0.35, 在 MovieLens 数据集上更是超过 0.6. 同时, 随着乱打分用户数量的变化, IARR2 算法结果也出现不同程度的较大波动. 因此, IARR2 算法在网络中存在不确定数量用户乱打分现象时, 算法鲁棒性波动较大.

对算法原理进行分析, CR 和 IARR 模型的声誉值结果是基于用户打分和其相应打分产品平均质量间的皮尔森相关系数产生的. 在线打分系统中, 对同一产品进行打分的用户分布广泛, 因此部分随机用户的乱打分为使产品平均质量产生较大变化的可能性较小, 个体行为也就不会对系统整体评价产生较大影响. 然而, 在产品质量计算中引入最大声誉值作为惩罚因子的 IARR2 算法, 使得高声誉用户在产品质量评价中起关键性作用. 因此, 部分声誉极高用户的行为变动就可能对整个评价系统产生较大干扰.

从图 1(c) 可见, 在 Douban 数据集中, 随着乱打分用户数量的增加, 3 种算法整体都呈现平缓上升趋势, 其中 CR 算法 $RMSE$ 值始终最小, IARR2 算法存在小幅波动. Douban 数据集中打分记录稀疏, 产品平均度相对较小, 因此用户间基于共同打分产品产生的关联性也较小, 所以, 部分用户随机打分为导致诚实用户声誉值的变化就相对平缓, 波动小. Douban 数据集的实验结果表明, CR 算法在稀疏性网络中也能保持很好的鲁棒性.

3 IARR2 算法鲁棒性影响因素的分析

本节对图 1(a) 和 (b) 实验结果中 IARR2 算法出现的突变和不规律现象展开进一步研究分析. 在 MovieLens 和 Netflix 两个经典数据集中, 用户打分彼此关联性较大, 通过分析算法原理发现, IARR2 算法中惩罚因子的引入, 使度大、原始声誉高的这部分特定用户的打分数据对算法评价结果起决定性作用, 所以当这部分特定用户的数据出现变化, 或者其他用户数据的变化对这小部分用户产生较大影响时, 算法对其他用户的声誉评价就可能产生较大幅度变化.

3.1 特定用户的选取和分析方法

为了进一步研究 IARR2 算法受特定用户影响较大这一特性, 本文通过对 IARR2 算法的鲁棒性分析结果和用户度及原始声誉值进行综合分析, 选取算法中原始声誉值极高的用户, 研究其单独的随机打分为在两个经典数据集中对系统中其余用户声誉评价的影响.

在 MovieLens 数据集中, 将其中唯一一个原始声誉值极高用户的打分随机化, 分析系统中其余用户声誉值的变化程度. 该用户具体信息见表 2 (见下页), 同时 IARR2 算法在 MovieLens 数据集中得到的第二大原始声誉值仅为 1.77; 在 Netflix 数据集中, 研究原始声誉值极大的前两位用户同时进行随机打分, 对系统其余用户声誉值评价的影响. 这两位用户具体信息见表 2, 同时位列第三位的原始声誉值仅为 4.82.

3.2 特定用户对 IARR2 算法鲁棒性影响的分析结果

按上述实验方法, 在 MovieLens 和 Netflix 两个

数据集中,分别选定 IARR2 算法中得到极大声誉值的用户为系统中的乱打分用户,然后通过 IARR2 算法重新计算用户声誉值,最后统计除选定用户外的其余所有用户的新声誉值和对应原始声誉值两组变量的 $RMSE$ 值. 10 次独立实验的平均结果见表 2.

表 2 声誉值极大用户对 IARR2 算法鲁棒性影响的分析结果

Tab.2 Effect of users with high reputation on the robustness of IARR2

数据集	用户编号	原始声誉值	$RMSE$
MovieLens	276	14.279 96	0.480 5
	1 546	18.424 81	
Netflix	3 506	17.906 75	0.531 5

两个数据集上的实验结果表明, IARR2 算法的鲁棒性受算法中声誉值极大用户的影响较大. 因此, 如果这小部分特定用户自身打分为发生变化, 或者系统中其他用户乱打分的行为对这部分特定用户声誉值产生较大影响时, 都可能导致系统对其余用户的声誉评价产生较大改变.

4 结束语

本文分析了在线用户声誉评价中 3 种常用算法在不同数据集下的鲁棒性. 实验通过设定不同数量用户随机打分, 以 $RMSE$ 值反映算法对其余用户声誉评价的变化, 从而分析不同算法对用户随机打分行为的鲁棒性. 3 种算法在两个经典数据集下的 $RMSE$ 值表明, CR 算法鲁棒性相对最好, IARR 算法也相对稳定, 而 IARR2 算法的 $RMSE$ 值波动幅度较大. 对 IARR2 算法鲁棒性影响因素的分析结果表明, IARR2 算法的评价结果受其中声誉值极大用户的影响较大, 从而导致算法鲁棒性波动较大. 在较稀疏的 Douban 数据集集中的结果表明, 在打分数据稀疏情况下, CR 算法也能保持很好的鲁棒性. 因此, 对于网络中普遍存在的用户乱打分现象, 就鲁棒性而言, 应优先选择 CR 算法. 该工作对于进一步分析在线社交网络的结构特征对于动力学的影响具有一定参考意义^[13-15].

参考文献:

[1] 郭强, 刘建国. 在线社会系统的用户行为分析研究进展[J]. 复杂系统与复杂性科学, 2015, 12(2):

97-102.

- [2] LÜ L Y, MEDO M, YEUNG C H, et al. Recommender systems[J]. Physics Reports, 2012, 519(1): 1-49.
- [3] 卢志刚. 电子商务声誉——结构与评价研究[M]. 北京: 清华大学出版社, 2014.
- [4] ZHOU Y B, LEI T, ZHOU T. A robust ranking algorithm to spamming [J]. EPL (Europhysics Letters), 2011, 94(4): 48002.
- [5] LIAO H, ZENG A, XIAO R, et al. Ranking reputation and quality in online rating systems[J]. PLoS One, 2014, 9(5): e97146.
- [6] MASUM H, ZHANG Y C. Manifesto for the reputation society[EB/OL]. (2014-07-05)[2015-08-22]. <http://firstmonday.org/ojs/index.php/fm/article/view/1158/1078>.
- [7] LIM E P, NGUYEN V A, JINDAL N, et al. Detecting product review spammers using rating behaviors[C]// Proceedings of the 19th ACM International Conference on Information and Knowledge Management. New York: ACM, 2010: 939-948.
- [8] ZENG A, CIMINI G. Removing spurious interactions in complex networks[J]. Physical Review E, 2012, 85(3): 036101.
- [9] ZENG A, YEUNG C H, SHANG M S, et al. The reinforcing influence of recommendations on global diversification[J]. EPL (Europhysics Letters), 2012, 97(1): 18005.
- [10] LIAO H, CIMINI G, MEDO M. Measuring quality, reputation and trust in online communities [M] // CHEN L, FELSEFERNIG A, LIU J M, et al. Foundations of Intelligent Systems. Heidelberg: Springer, 2012: 405-414.
- [11] KENDALL M G. A new measure of rank correlation [J]. Biometrika, 1938, 30(1/2): 81-93.
- [12] HANLEY J A, MCNEIL B J. The meaning and use of the area under a receiver operating characteristic (ROC) curve[J]. Radiology, 1982, 143(1): 29-36.
- [13] 李旭东, 刘建国. 轨道交通网络的可达性及其网络结构分析[J]. 上海理工大学学报, 2015, 37(3): 279-283.
- [14] 周萍, 张子柯, 章恬, 等. 一种基于社会化媒体和社会网络结构的混合推荐模型[J]. 上海理工大学学报, 2014, 36(3): 267-276.
- [15] 石珂瑞, 刘建国. 二阶有向相似性对协同过滤算法的影响[J]. 上海理工大学学报, 2014, 36(1): 31-33.

(编辑: 丁红艺)