



文章编号:1001-9081(2019)06-1595-06

DOI:10.11772/j.issn.1001-9081.2018122611

基于带权评论图的水军群组检测及特征分析

张琪¹, 纪淑娟^{1*}, 傅强², 张纯金³

(1. 山东省智慧矿山信息技术重点实验室(山东科技大学), 山东 青岛 266590; 2. 秦皇岛市公安局技术侦察支队, 河北 秦皇岛 066000;
3. 山东科技大学 网络信息中心, 山东 青岛 266590)
(*通信作者电子邮箱 jane_ji2003@aliyun.com)

摘要:针对在电子商务平台上检测编写虚假评论的水军群组的问题,提出了基于带权评论图的水军群组检测算法(WGSA)。首先,利用共评论特征构建带权评论图,权重由一系列群组造假指标计算得到;然后,为边权重设置阈值筛选可疑子图;最后,从图的社区结构出发,利用社区发现算法生成最终的水军群组。在 Yelp 大型数据集上的实验结果表明,与 K 均值聚类算法(KMeans)、基于密度的噪声应用空间聚类算法(DBscan)以及层次聚类算法相比 WGSA 算法的准确度更高,同时对检测到水军群组的特征与差异作了分析,发现水军群组的活跃度不同,危害也不同。其中,高活跃度群组危害最大,应重点关注。

关键词:电子商务;水军群组;带权评论图;社区发现;聚类

中图分类号: TP391.4 **文献标志码:**A

Weighted reviewer graph based spammer group detection and characteristic analysis

ZHANG Qi¹, JI Shujuan^{1*}, FU Qiang², ZHANG Chunjin³

(1. Shandong Key Laboratory of Wisdom Mine Information Technology
(Shandong University of Science and Technology), Qingdao Shandong 266590, China;
2. Technical Reconnaissance Detachment, Qinhuangdao Public Security Bureau, Qinhuangdao Hebei 066100, China;
3. Network Information Center, Shandong University of Science and Technology, Qingdao Shandong 266590, China)

Abstract: Concerning the problem that how to detect spammer groups writing fake reviews on the e-commerce platforms, a Weighted reviewer Graph based Spammer group detection Algorithm (WGSA) was proposed. Firstly, a weighted reviewer graph was built based on the co-reviewing feature with the weight calculated by a series of group spam indicators. Then, a threshold was set for the edge weight to filter the suspicious subgraphs. Finally, considering the community structure of the graph, the community discovery algorithm was used to generate the spammer groups. Compared with K-Means clustering algorithm (KMeans), Density-Based spatial clustering of applications with noise (DBscan) and hierarchical clustering algorithm on the large dataset Yelp, the accuracy of WGSA is higher. The characteristics and distinction of the detected spammer groups were also analyzed, which show that spammer groups with different activeness have different harm. The high-active group is more harmful and should be concerned more.

Key words: e-commerce; spammer group; weighted reviewer graph; community discovery; clustering

0 引言

在电子商务平台上,在线商品评论在用户的决策中起着重要作用。用户倾向于购买正面评论较多的产品,而不是负面评论较多的产品。为了抬高或降低某产品的信誉,赚取更多利益,很多商家往往会雇佣虚假评论者发布大量赞美自家商品或诋毁竞争对手商品的不实评论,误导消费者,影响电商平台的公平竞争环境。这些虚假评论者称为水军。近年来,随着电子商务的迅猛发展,水军的规模也越发壮大,甚至结成水军群组协同作案。水军群组即指那些有组织地协同发布虚假评论的一群人。相比水军个体,水军群组影响力更大(甚至能控制产品的舆论走势、造成用户逆向选择)、隐秘性更强,因此对检测算法的设计要求更高。

在水军群组检测方面,研究者也提出了一些有针对性的

检测方法。文献[1]首次进行了电商平台水军群组的检测工作,指出水军群组的一个重要特征——“共评论”,即水军成员通常共同评论相同的产品。为了检测共评论的水军群组,他们利用频繁项挖掘的方法寻找共评论过多个产品的评论者集作为候选水军群组,然后提出一种排序模型来定位最可疑的水军群组。

继文献[1]之后,文献[2]也使用频繁项挖掘的方法来确定候选水军群组,他们还评价了已有的用于识别评论者造假个体的特征与造假群组的特征的有效性;但是他们工作的目的是设计算法实现共谋者个体和非共谋者个体的检测,而不是水军群组的检测。文献[3]提出了一种水军群组检测算法。该算法分两步实现:第一步,量化某产品为水军目标产品的概率,定位目标产品;第二步,利用层次聚类算法得到水军群组。文献[4]提出基于评论-产品构建二部图,然后利用一

收稿日期:2018-12-12;修回日期:2019-03-25;录用日期:2019-03-28。 基金项目:国家自然科学基金资助项目(71772107, 61502281)。

作者简介:张琪(1995—),男,山东泰安人,硕士研究生,CCF 会员,主要研究方向:人工智能; 纪淑娟(1977—),女,河北唐山人,副教授,博士,CCF 高级会员,主要研究方向:分布式智能、智能信息处理; 傅强(1983—),男,河北唐山人,工程师,主要研究方向:智能信息处理、虚假信息检测; 张纯金(1977—),男,山东菏泽人,工程师,硕士,主要研究方向:网络安全、智能信息处理。



系列群组造假特征作为识别标准,使用图划分方法得到水军群组。文献[5]依据评论者“共评论”的关系特征构建用户关系网络,然后使用一系列特征构建多特征尺度空间模型进行水军群组的识别。

从已有群组检测研究的发展来看,利用基于图的方法来检测水军群组是一个趋势。群组划分多采用聚类算法、图划分算法。然而,上述方法只进行了水军群组的划分,没有对水军群组进行进一步的分析,探究不同水军群组间的联系和差别,以发现水军群组的整体行为特征。

针对上述工作的不足,本文提出了基于带权评论图的水军群组发现算法(Weighted reviewer Graph based Spammer group detection Algorithm, WGS)。本文的主要工作总结如下:

- 1) 本文在基于图的水军群组检测方法基础上,构建了带权评论图,然后利用权重筛选子图。该方法能够去掉大部分不重要的节点,大大降低计算的时空复杂度。

- 2) 本文从图的社区结构出发,认为水军群组的造假行为会形成典型的社区结构,所以本文采用社区发现算法生成水军群组,实验证明效果较好。

- 3) 基于 Yelp 的大型带标签数据集,本文对发现的水军群组作了全面的可疑度分析以证明本文算法的有效性,同时探究了水军群组的差异和整体行为特征。

1 水军群组检测算法

本章描述了本文提出的水军群组检测算法,算法由四个步骤组成,即水军群组造假行为特征选择、带权评论图的构建、可疑子图的筛选以及基于社区发现算法的水军群组的聚类。下面详细介绍每个步骤细节。

1.1 造假行为特征选择

在已有工作中,研究者提出了很多评估个人或群组的造假指标,例如语言指标^[1,3,6-7]、行为指标^[1-4,8-12]、关系指标^[2-6,8-14]等。与之前提出的指标不同,本文使用行为指标量化两个评论者之间的共谋程度,具体指标如下。

1.1.1 共评论次数

水军群组的成员通常同时针对多个产品发表评论,协同合作完成任务。两两评论者,如果只共同评论过一件或两件产品,有可能只是因为巧合,是正常用户的评论,不能因此判定为水军组织成员;而评论用户作为分散的网络用户,若共同评论的产品数很多,就可视为非正常用户行为。本文利用共评论次数(Co-Reviewing Time, CRT)^[1]来捕捉两两评论者的共评论特征。

$$CRT(n_1, n_2) = |P_1 \cap P_2| \quad (1)$$

其中: P_1, P_2 分别是评论者 n_1, n_2 评论的产品集。 $P \in (P_1 \cap P_2)$ 是 P_1, P_2 的交集。 CRT 的值越高,说明两评论者协同作案的可能性越高。

在文献[1~2]中,研究者们用频繁项挖掘(Frequent Itemset Mining, FIM)方法^[15]来生成候选群组。受其启发,本文使用频繁2项集挖掘方法计算 CRT 值。将评论者对作为二项集 $\{n_1, n_2\}$,产品集作为事务 T 。基于频繁2项集挖掘方法, CRT 的计算公式如下:

$$CRT(n_1, n_2) = support_{(n_1, n_2)} \cdot |T| \quad (2)$$

其中, $support_{(n_1, n_2)}$ 是项集 $\{n_1, n_2\}$ 的支持度。

1.1.2 评分相似度

水军群组通常协同发布虚假评论来抬高或贬低目标产品

的评分。因此,水军群组成员往往发布相似评分来控制目标产品的评分趋势。本文定义了评分相似度(Similarity of Rating, SR)^[5]来捕捉这种行为。

$$SR(n_1, n_2) = \frac{\sum_{p \in (P_1 \cap P_2)} (R_{p1} - \beta) \cdot (R_{p2} - \beta)}{\sqrt{\sum_{p \in (P_1 \cap P_2)} (R_{p1} - \beta)^2} \cdot \sqrt{\sum_{p \in (P_1 \cap P_2)} (R_{p2} - \beta)^2}} \quad (3)$$

其中: R_{pi} 是评论者 n_i 对产品 p 的评分,评分 $R \in [1, 5]$;本文引入了一个参数 β 以减少误差, β 取值为 2.5。 $SR(n_1, n_2) \in [1, 5]$, SR 值越趋近于 -1, 表示两两评论者在同一维度上的评分值偏差越大;越趋近于 1, 表示两两评论者观点一致性越强。

1.2 带权评论图的构建

在电子商务网站中,不同的用户可以通过两种方式建立联系:一种是用户之间的直接交互,例如用户发表评论和其他用户回复其评论。另一种隐含的联系是两个用户对同一产品进行评论,即共评论。一个水军群组的成员通常共同评论相同的产品,这是识别水军群组成员间联系的关键。

本文将评论者个体作为节点,将用户的共评论关系作为边的联系,构建带权评论图 $G = (N, E, W)$ 。 N 是由全体评论者组成的节点集,边 $e = (n_1, n_2) \in E$ 存在当且仅当评论者 n_1, n_2 至少共同评论过一个产品。边的权重 $w \in W$, 对应着每一条边,代表了两两评论者节点间共谋的可疑度。

边的权重 w 由 1.1 节描述的造假行为特征计算得到,计算式如下:

$$w_{e=(n_1, n_2)} = k \cdot \frac{CRT}{\max(|P_i \cap P_j|)} + (1 - k) \cdot SR \quad (4)$$

其中:本文将 CRT 值归一化为 $[0, 1]$; P_i, P_j 分别是 n_i, n_j 评论过的产品集; k 表示 CRT, SR 在计算时所占的比重, k 取 0.5。

1.3 可疑子图的筛选

本文构建的评论图是基于评论者的共评论特性,边的权重代表了两两评论者间共谋的可疑度。因为原始评论图十分庞大,计算难度较高,本文首先进行可疑子图的筛选,既可以保证算法的准确度,也可以降低算法的时间复杂度。详见算法 1。

算法 1 可疑子图的筛选。

输入 评论者、评论、产品数据 B ,边的虚假度阈值 δ ;

输出 可疑子图。

描述:

- 1) 构建原始带权评论图 $G = (N, E, W)$, 将边的权重初始化为 1
- 2) for 边 $e = (n_1, n_2) \in E$ do
 - 3) 计算权重
 - 4) if $w_e < \delta$ then
 - 5) 移除边 e
 - 6) end for
 - 7) 输出筛选得到的可疑子图

在算法 1 中,在第 1) 行,首先构建带权评论图 G ,将边的权重初始化为 1;第 2) ~ 7) 行,计算边的权重,设置边权重的阈值 δ ,移除边权重 $w_e < \delta$ 的边,得到筛选后的子图。边筛选阈值 δ 的确定在实验部分具体说明。

1.4 水军群组的聚类

水军群组的造假行为会在评论图中形成典型的社区结构,基于此,本文利用 Louvain 社区发现算法^[16]来生成水军群



组。Louvain 算法是典型的社区发现算法,它基于最大化模块度进行社区划分,能够有效地发现网络中社区结构,即本文中的水军群组。

2 实验及结果分析

2.1 数据集

与文献[6,10~11]中的实验研究相同,本文也使用来自美国著名商户点评网站 Yelp 自 2006 年起历时 7 年的旅店评论数据。该数据集包含了评论虚假与否的标签,数据集的评论真率为 61.1%。特别的,数据集中没有重复交易的买家和卖家对。每条评论包含以下属性:日期、评论 ID、评论者 ID、评论内容、评分、认为该评论有用的用户个数、认为该评论很酷的用户个数、认为该评论有趣的用户个数、标签、旅店 ID。

在数据被使用之前,本文对数据集进行了如下预处理:

1)删除评论集中匿名的用户及评论数据。因为无法确定匿名是被同一人发表还是被多人发表。

2)删除不活跃的用户和产品。在本文研究中关注的是活跃度较高的用户,以及具有较高关注度的产品,不活跃的用户可疑性小,可以忽略。在数据集中评论用户发表的评论数少于三个,以及产品的评论数少于三个,则首先将其删除。

3)将数据表中未使用的属性去除,以精简数据集。

经过以上三个方面的数据处理之后,数据集的概况如表 1 所示。

表 1 预处理前后的数据集概况

Tab. 1 Overview of data sets before and after preprocessing

数据集	评论者数	评论数	产品数
预处理前	5 123	688 329	283 086
预处理后	4 639	641 768	238 236

2.2 边筛选阈值 δ 的确定

δ 的大小决定了筛选得到的可疑子图的大小与质量:如果 δ 取值过大,删除的边过多,可能严重破坏子图的结构,影响后面社区划分的质量;如果取值太小,又无法保证得到的子图中边和节点的可疑度。由于边的权重是通过特征 CRT 与 SR 计算得到,本文分别探究了 CRT 的阈值,记作 ω_{CRT} 和 SR 的阈值 ω_{SR} 。如果一条边的 $CRT \geq \omega_{CRT}, SR \geq \omega_{SR}$, 则该边是可疑的。在这两个阈值的基础上,本文提出了如下 δ 计算方法:

$$\delta = l \cdot \frac{\omega_{CRT}}{\max(|P_i \cap P_j|)} + (1 - l) \cdot \omega_{SR} \quad (5)$$

其中: l 为平衡 ω_{CRT} 、 ω_{SR} 的权重因子。在后面的实验中 l 取 0.5, 表示、在计算时取相同的权重。下面介绍 ω_{CRT} 、 ω_{SR} 的计算。

2.2.1 ω_{CRT} 的计算

用式(2)计算评论图中边的 CRT 值,频繁 2 项集挖掘的结果如表 2 所示。边的 CRT 值统计数据如图 1 所示,其中 61% 的边的 CRT 值为 3、4 和 5。

接下来的问题就是 ω_{CRT} 的选取,以筛选可疑的边。如果 ω_{CRT} 取值过大,会过滤掉大部分边,严重破坏图的结构;如果 ω_{CRT} 取值过小,过滤效果不明显。为了避免过度破坏图的结构,本文选取了 3 个通用的指标:模块度(Modularity, Q)^[17]、平均聚类系数(Average Clustering Coefficient, ACC)^[18] 和平均路径长度(Average Path Length, APL)^[19] 来评价网络社区结构。 Q 、 ACC 、 APL 值越大,则代表相应的图更紧密,社区结构更明显。本文采用插值法,计算了 ω_{CRT} 取不同值时,筛选得到

的子图的 Q 、 ACC 、 APL 值。计算结果如表 3 所示,当 $\omega_{CRT} = 40$ 时, Q 、 ACC 、 APL 均取得最大值,这说明,此时的网络社区结构达到了最佳,所以, ω_{CRT} 取 40。

表 2 频繁项挖掘结果
Tab. 2 Frequent item mining results

类别	数量	备注
项(I)	4 639	初始评论图的节点数
事务(T)	238 236	产品集
频繁二项集(F)(支持度 > 0.000 01)	347 960	初始评论图的边数

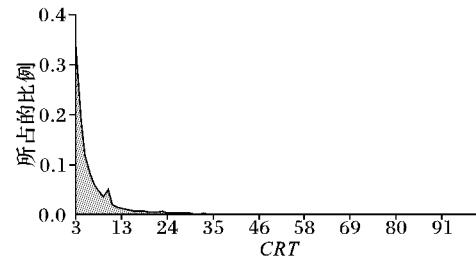


图 1 CRT 的分布

Fig. 1 Distribution of CRT

表 3 ω_{CRT} 取不同值时筛选出的子图的 Q 、 ACC 和 APL 值
Tab. 3 Q , ACC , APL of filtered subgraphs under different ω_{CRT}

ω_{CRT}	Q	ACC	APL	ω_{CRT}	Q	ACC	APL
6	0.442	0.370	2.462	30	0.550	0.395	2.776
7	0.456	0.373	2.520	35	0.558	0.395	2.780
8	0.466	0.376	2.569	38	0.561	0.399	2.765
9	0.475	0.379	2.607	39	0.563	0.397	2.769
10	0.483	0.382	2.636	40	0.566	0.397	2.791
15	0.511	0.392	2.779	41	0.566	0.393	2.779
20	0.528	0.394	2.817	42	0.565	0.389	2.779
25	0.543	0.395	2.813				

2.2.2 ω_{SR} 的计算

用式(3)计算 SR 的值,对 SR 值的分布进行统计,如图 2 所示。从图 2 中可以看出,大部分边的 SR 值都大于 0.5,这说明大部分边所连接的两两评论者之间的评分相似度极高,观点一致性较强。这里取 ω_{SR} 为 0.5。

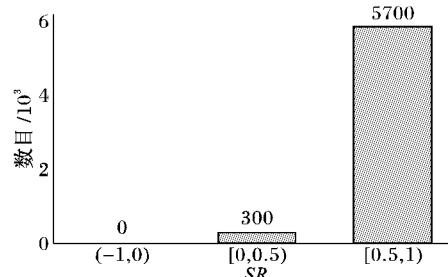


图 2 SR 值的分布

Fig. 2 Distribution of SR

在 ω_{CRT} 与 ω_{SR} 确定之后, δ 随之可以确定。可疑子图的筛选情况如图 3 所示。筛选后的子图包含 840 个节点,5 442 条边,相比初始的评论图,规模大大减小。

2.3 水军群组的聚类

经过 Louvain 算法聚类的可视化结果如图 4(a) 所示。本文移除了聚类结果中成员数较少(小于 10) 的类簇,最终得到了 12 个典型水军群组,如图 4(b) 所示。12 个水军群组的相关信息如表 4 所示。

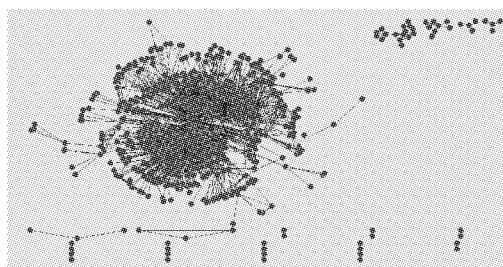


图3 筛选后的子图
Fig. 3 Filtered subgraphs

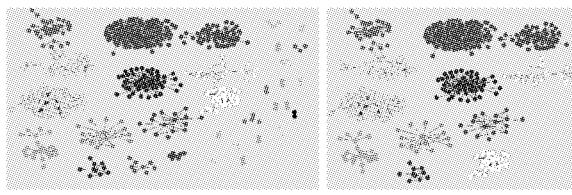


图4 聚类出的水军群组
Fig. 4 Clustered spammer groups

表4 每个群组的成员数
Tab. 4 Number of members per group

群组编号	成员数	群组编号	成员数	群组编号	成员数	群组编号	成员数
1	190	4	52	7	58	10	16
2	283	5	41	8	30	11	16
3	51	6	14	9	20	12	10

2.4 结果分析

鉴于本文使用的 Yelp 数据集只有评论虚假与否的标签,首先从虚假评论比例出发,分析了检测到的水军群组的特征与差异。然后选取 K 均值聚类算法 (K-Means clustering

algorithm, KMeans)、基于密度的噪声应用空间聚类算法 (Density-Based spatial clustering of applications with noise, DBscan) 以及层次聚类算法进行对比,验证本文算法的有效性。

2.4.1 基于虚假评论比例的造假度分析

正如许多研究中所提到的,Yelp、Amazon 和 Dianping 等大型电子商务网站的数据集只能得到虚假/真实的评论标签,很难得到评论者个体的标签,更不用说水军群组了。在文献[6]中,至少发布过一条假(被电商网站过滤掉的)评论的评论者将被视为虚假评论者,没有假评论的评论者将被视为正常评论者。在文献[13]中,如果评论者至少有 10% 的评论被 Dianping 网站检测到是假的,则将其视为虚假评论者。在文献[14]中,一个评论者发布的评论中如果有超过 50% 的评论是假的,即被认为是垃圾邮件用户。为了获取水军群组的标签,文献[1-2,4]中只能采用手动标记的方法。而在文献[14]中通过评估聚类质量来评价得到的水军群组的好坏,这样做说服力明显不足。

结合上述文献对标签的处理,本文进行了有趣的分析,对于每个水军群组,本文计算了在这个群组中,虚假评论超过一定百分比的评论者所占的比例,统计情况如表 5 所示。

表 5 中的值指的是每个群组中至少发布了 10%、20%、…虚假评论的评论者的比例。例如,在第一组中,有 190 个成员。在这一组中,100%(表 5 中的第一组)的评审员发布了超过 10% 的虚假评论,这意味着第一组的所有成员都发布了超过 10% 的虚假评论。注意到,第 6 组第 8 行出现的 0,指的是群组 6 中没有成员的虚假评论比例超过 45%,换言之,群组 6 中的成员发布的虚假评论比例均低于 45%。特别地,第一组的成员中有 69% 的成员至少发布了 50% 的虚假评论。这种群组可疑度极大。

表5 群组中虚假评论超过一定比例的成员占比
Tab. 5 Proportion of members in group whose fake reviews exceed a certain percentage

虚假评论比例/%	群组编号											
	1	2	3	4	5	6	7	8	9	10	11	12
10	1.00	0.99	1.00	0.98	0.90	0.85	1.00	1.00	1.00	0.90	0.75	1.00
15	1.00	0.98	1.00	0.94	0.88	0.64	1.00	1.00	0.95	0.50	0.44	1.00
20	0.99	0.95	0.98	0.86	0.78	0.50	1.00	0.93	0.80	0.30	0.25	1.00
25	0.99	0.90	0.96	0.67	0.70	0.42	1.00	0.83	0.70	0.20	0.25	1.00
30	0.97	0.82	0.92	0.51	0.58	0.28	0.96	0.80	0.55	0.10	0.19	1.00
35	0.94	0.71	0.84	0.34	0.46	0.14	0.84	0.56	0.40	0.10	0.19	0.94
40	0.86	0.59	0.76	0.28	0.24	0.14	0.69	0.47	0.30	0.10	0.06	0.88
45	0.80	0.48	0.64	0.13	0.19	0.00	0.59	0.43	0.25	0.00	0.06	0.56
50	0.69	0.39	0.53	0.05	0.15	0.00	0.50	0.23	0.15	0.00	0.00	0.31

本文还计算了每个水军群组中成员虚假评论比例的平均值,结果如图 5 所示。由图 5 可以看出,不同群组间有极大的差异性,例如群组 1、2、3、7 和 12 中成员的平均虚假评论比例均高于 40%,群组 4、5、8 和 9 为 30% ~ 40%,群组 6、10、11 为 10% ~ 30%。从图 5 中可以看出,不同群组的活跃度是不同的。因此本文将 12 个群组分为 3 类:群组 1、2、3、7 和 12 为高活跃度群组,群组 4、5、8 和 9 为一般活跃群组,群组 6、10、11 为低活跃度群组。

三类群组中成员的虚假评论比例如图 6 所示。从图 6 中可以看出,高活跃度群组,成员数较多,大部分成员的虚假评论比例均超过 30%,危害极大;一般活跃群组,成员规模一般,虚假评论比例也较高,但远低于高活跃度群组;相对来说,

低活跃度群组成员数较少,虚假评论比例也较低。综上所述,高活跃度群组因为人数多、每个人的造假比例高,对整个市场环境的危害也最大,因此应重点关注。

2.4.2 对比实验

为了验证本文算法的性能,本文选举经典的聚类算法 KMeans 算法、基于密度的聚类算法 DBscan 算法作为基准算法进行对比。在现有工作中,文献[3]利用层次聚类算法生成水军群组,所以,本文也与层次聚类算法作了比较。

本文利用 KMeans 算法、DBscan 算法、层次聚类算法以及本文所提出的基于带权评论图的水军群组发现算法(WGSA)对检测出的 top12 个群组的 4 个特征进行评估。具体特征为一天最大评论数 (Maximum One day Review, MOR)^[20]、极端



评分比率(Extreme rating Ratio, EXR)^[20]、评论时间间隔(Review Time Interval, RTI)^[1,20]和评论者比率(Reviewer Ratio, RR)^[4]。之所以选择这些特征作为评估指标,主要因为它们具有很好的通用性,在相应文献采用这4个特征对个体或群体作可疑度的评估和比较,表现较好。

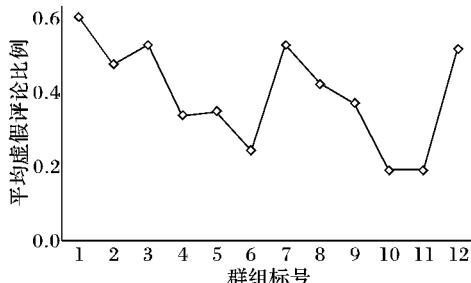


图5 每个水军群组中成员的平均虚假评论比例

Fig. 5 Average fake review proportion of members in each spammer group

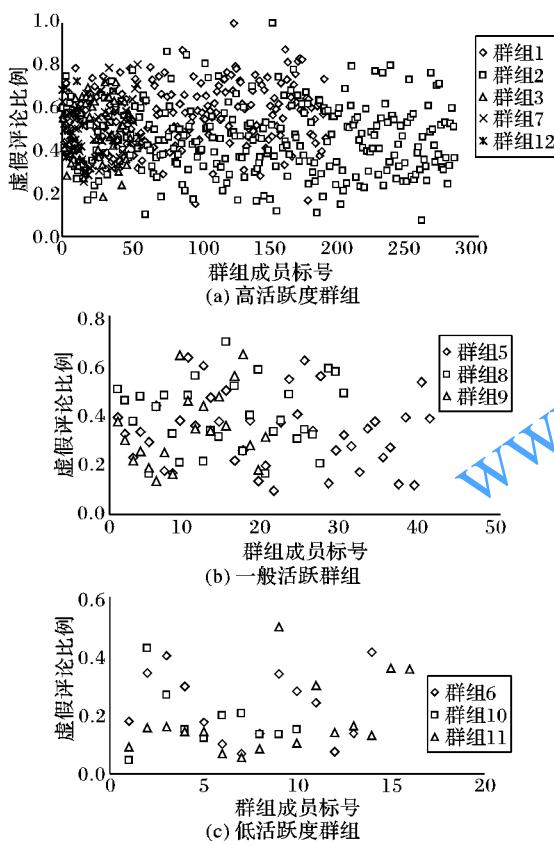


图6 不同活跃度群组中成员的虚假评论比例

Fig. 6 Fake reviews proportion of members in different activity groups

下面详细介绍各种算法对每个群组对这4个特征的评估实验结果。

1)一天最大评论数(MOR)。

一个评论者在一天中发布大量评论是十分可疑的。MOR度量的是一个评论者一天发布评论的最大值。文献[20]的研究结果显示一名水军一天的理论评论数至少为5,而正常评论者一般为2。对每个水军群组的成员计算其MOR值,然后取每个水军群组中成员MOR的平均值,得到如图7所示的结果。从图7可以看出,各算法检测出的水军群组平均一天最大评论数均超过6,有些群组甚至超过20,十分可疑,而本文算法与DBscan算法的表现相对更加突出。

2)极端评分比率(EXR)。

水军往往发布极高或极低的评分来抬高或降低目标产品

的评分。EXR度量的是一个评论者的评分是否极高或极低。由于评分范围为[1,5],本文采用与文献[20]一样的处理方法,即将1,5作为极端评分,然后计算每个评论者极端评分的比例。计算得到的每个水军群组中成员的平均极端评分比率如图8所示。从图8可以看出,本文算法检测出的水军群组中成员的平均极端评分比率均大于0.6,而其他算法只有0.3左右,本文的算法表现较好。

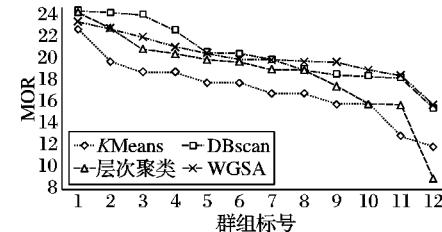


图7 每个水军群组的平均MOR

Fig. 7 Average MOR of each spammer group

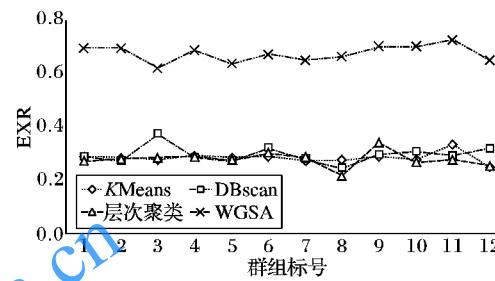


图8 每个水军群组的平均EXR

Fig. 8 Average EXR of each spammer group

3)评论时间间隔(RTI)。

水军通常在较短的时间内连续发布虚假评论,所以相邻评论间较短的时间间隔揭示了疑似水军行为。文献[1,20]指出,如果一个评论者的相邻评论时间间隔小于28天,则是可疑的。本文亦取小于28天的评论时间间隔为可疑时间间隔。RTI计算的是一个评论者的所有相邻评论时间间隔中可疑时间间隔的比例。每个群组中成员的平均RTI值如图9所示。从图9可以看出,本文算法检测出的水军群组的平均RTI值均在0.9左右,而其他算法的表现差一些,在0.7左右。

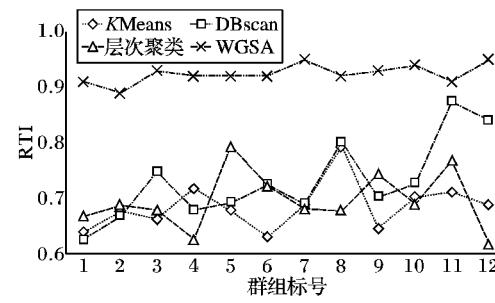


图9 每个水军群组的平均RTI

Fig. 9 Average RTI of each spammer group

4)评论者比率(RR)。

如果目标产品主要由某水军群组的成员所评论,该水军群组就能完全控制该产品的舆论,危害极大。RR度量的是一个产品的评论者中身为某水军群组成员的比例。本文取一个群组中该比例的最大值作为RR的值。每个水军群组的最大RR值如图10所示。从图10可以看出,所有算法中每个水军群组的RR值均为1,这说明这些水军群组完全控制了部分产品的舆论走势,危害极大。



从上述分析可以得到,本文提出的算法 WGS, 在 MOR、RR 指标上表现相对较好, 在 EXR、RTI 指标上比其他算法有较大提升, 总体来看, 本文算法得到的水军群组可疑度更高, 更有效。

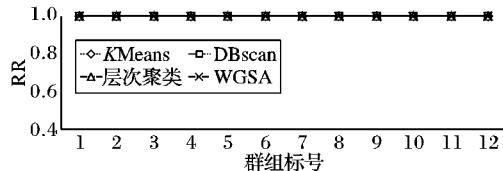


图 10 每个水军群组的 RR 值
Fig. 10 RR of each spammer group

3 结语

本文提出了基于带权评论图的水军群组发现算法 (WRBA)。该算法首先构建带权评论者网络图, 权重由一系列特征计算得到; 然后设置阈值筛选可疑子图; 最后利用社区发现算法生成水军群组。本文从虚假评论比例出发, 发现检测到的水军群组成员的平均虚假评论比例均超过 10%, 表明了本文所提算法的有效性。而且本文研究发现, 水军群组可以分成三类: 高活跃度群组、一般活跃群组以及低活跃度群组。其中, 高活跃度群组发布的评论多, 虚假评论比例高, 危害极大, 应重点关注。为了验证本文算法的性能, 本文选取了多个已有算法在 4 个群组虚假度特征 (MOR、EXT、RTI 和 RR) 上进行比较。实验结果表明, 本文算法检测出的水军群组可疑度更高, 算法性能表现更好。但本文只考虑了两种特征来构建带权评论者网络图, 而且没有考虑时间因素, 在今后的工作中, 将考虑更多的特征, 完善水军群组的检测方法。

参考文献 (References)

- [1] MUKHERJEE A, LIU B, GLANCE N. Spotting fake reviewer groups in consumer reviews [C] // Proceedings of the 21st Annual Conference on World Wide Web. New York: ACM, 2012: 191 – 200.
- [2] XU C, ZHANG J, CHANG K, et al. Uncovering collusive spammers in Chinese review website [C] // Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. New York: ACM, 2013: 979 – 988.
- [3] YE J, AKOGLU L. Discovering opinion spammer groups by network footprints [C] // Proceedings of the 2015 Joint European Conference on Machine Learning and Knowledge Discovery in Databases, LNCS 9284. Cham: Springer, 2015: 267 – 282.
- [4] WANG Z, HOU T, SONG D, et al. Detecting review spammer groups via bipartite graph projection [J]. The Computer Journal, 2016, 59(6): 861 – 874.
- [5] 张慧杰. 基于多特征尺度空间模型的网络水军组织发现技术研究 [D]. 杭州: 浙江工商大学, 2015: 2 – 66. (ZHANG H J. Research technology on found of spammer organizations based on multi-feature scale space model [D]. Hangzhou: Zhejiang Gongshang University , 2015: 2 – 66.)
- [6] RAYANA S, AKOGLU L. Collective opinion spam detection: bridging review networks and metadata [C] // Proceedings of the 2015 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2015: 985 – 994.
- [7] RAYANA S, AKOGLU L. Collective opinion spam detection using active inference [C] // Proceedings of the 2016 16th SIAM International Conference on Data Mining. Philadelphia, PA: SIAM, 2016: 630 – 638.
- [8] JINDAL N, LIU B. Opinion spam and analysis [C] // Proceedings of the 2008 International Conference on Web Search & Data Mining. New York: ACM, 2008: 219 – 230.
- [9] LIM E, NGUYEN V, JINDAL N, et al. Detecting product review spammers using rating behaviors [C] // Proceedings of the 19th ACM Conference on Information and Knowledge Management. New York: ACM, 2010: 939 – 948.
- [10] OTT M, CHOI Y, CARDIE C, et al. Finding deceptive opinion spam by any stretch of the imagination [C] // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: Association for Computational Linguistics, 2011: 309 – 319.
- [11] YU P S, LIU B, XIE S, et al. Review graph based online store review spammer detection [C] // Proceedings of the 11th IEEE International Conference on Data Mining. Piscataway, NJ: IEEE, 2011: 1242 – 1247.
- [12] AKOGLU L, CHANDY R, FALOUTSOS C. Opinion fraud detection in online reviews by network effects [C] // Proceedings of the 2013 7th International Conference on Weblogs and Social Media. Menlo Park, CA: AAAI, 2013: 2 – 11.
- [13] LI H, CHEN Z, MUKHERJEE A, et al. Analyzing and detecting opinion spam on a large-scale dataset via temporal and spatial patterns [C] // Proceedings of the 9th International Conference on Web and Social Media. Menlo Park, CA: AAAI, 2015: 634 – 637.
- [14] LI H Y, FEI G, SHAO W X, et al. Bimodal distribution and co-bursting in review spam detection [C] // Proceedings of the 26th International Conference on World Wide Web. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2017: 1063 – 1072.
- [15] AGRAWAL R, SRIKANT R. Fast algorithms for mining association rules in large databases [C] // Proceedings of the 20th International Conference on Very Large Data Bases. San Francisco, CA: Morgan Kaufmann Publishers Inc., 1994: 487 – 499.
- [16] BLONDEL V D, GUILLAUME J, LAMBIOTTE R, et al. Fast unfolding of communities in large networks [J]. Journal of Statistical Mechanics Theory & Experiment, 2008(10): 155 – 168.
- [17] NEWMAN M E J. The structure and function of complex networks [J]. SIAM Review, 2003, 45(2): 167 – 256.
- [18] WATTS D J, STROGATZ S H. Collective dynamics of ‘small-world’ networks [J]. Nature, 1998(393): 440 – 442.
- [19] FRONCZAK A, FRONCZAK P, HOŁYST J A. Average path length in random networks [J]. Physical Review E, 2004, 70 (5): 056110.
- [20] MUKHERJEE A, KUMAR A, LIU B, et al. Spotting opinion spammers using behavioral footprints [C] // Proceedings of the 2013 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2013: 632 – 640.

This work is partially supported by the National Natural Science Foundation of China (71772107, 61502281).

ZHANG Qi, born in 1995, M. S. candidate. His research interests include artificial intelligence.

JI Shujuan, born in 1977, Ph. D., associate professor. Her research interests include distributed intelligence, intelligent information processing.

FU Qiang, born in 1983, engineer. His research interests include intelligent information processing, false information detection.

ZHANG Chunjin, born in 1977, M. S., engineer. His research interests include network security, intelligent information processing.