基于双循环图的虚假评论检测算法

陈晋音 黄国瀚 吴洋洋 贾澄钰
(浙江工业大学信息工程学院 杭州 310023)

摘 要 由于对商店的在线评论能给顾客提供许多有价值的信息并极大地影响商店的信誉度,因此,在利益的驱使下出现了大量虚假评论,扰乱了市场秩序。许多商店或个人通过虚假评论故意吹捧或诋毁特定商店,从而达到获利的目的,因此提出有效的虚假评论检测方法至关重要。文中基于大量用户、评论和商店之间的关系构建图过滤器,经过迭代计算获得用户、评论和商店的置信度,从而发现虚假评论。其中包括3个关键问题:获取可靠的用户、评论和商店置信度,有效地辨识真实评论,准确发现虚假评论及虚假用户。针对提高用户、评论和商店置信度的可靠性问题,文中提出了一种循环迭代的方法来获取可靠的用户、评论和商店置信度;为了更加有效地发现虚假评论和虚假用户,设计了一种加权图过滤器,通过与获取的可靠置信度结合,得到了一种双循环图过滤检测算法。将所提检测算法应用到Yelp数据集上展开实验,验证了所虚假检测算法可以有效检测虚假评论。

关键词 虚假检测,双循环图,基于图的过滤器,行为特征,用户影响力

中图法分类号 TP393.1 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.09.034

Double Cycle Graph Based Fraud Review Detection Algorithm

CHEN Jin-yin HUANG Guo-han WU Yang-yang JIA Cheng-yu
(College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract Because online reviews of stores can provide customers with a lot of valuable information and greatly affect the credibility of stores, a large number of spam reviews are emerged to disturb the order of market for pro-fit. Many stores or individuals deliberately flatter or denigrate certain stores through fake reviews to achieve their profit objectives. Thus an efficient fraud review detection algorithm is crucial. This paper built a graph filter based on the relationships among users, comments and stores, and obtained the reliability of users, comments and stores through iterative calculation, so as to find fake reviews. Three key questions are brought up: to get more reliable reliability of users, comments and stores, to identify the real reviews effectively, and to detect fake reviews and spammers effectively. In order to improve the reliability of users, comments and stores, a double cycle graph based detection algorithm was proposed to obtain reliable users, comments and stores. In order to find fake reviews and spammers effectively, this paper designed a novel weighted graph filter, through the combination of reliability and obtain reliable, and put forward double cycle filtering detection algorithm. The proposed detection algorithm is applied to Yelp datasets for experiments and proved efficiently in detection of spammers and identifies real reviews.

Keywords Spam detection, Double cycle graph, Graph-based filter, Behavior characteristic, User influence

1 引言

大多数电子商务允许用户对其服务和质量进行线上评论。线上评论逐渐成为消费者购物的依据,用户对商店的线上评论将极大地影响该商店的口碑和销售量。用户评论在推荐系统方面发挥着重要作用,大量真实有效的评论数据能使推荐系统产生有效的推荐,从而给消费者提供正确的商店或产品推荐。受利益驱使,某些用户会故意撰写与现实不符的评论,即虚假评论,以夸大或诋毁某一家商店,使其获取的利益最大化,甚至出现一些商店雇佣大量网络水军来集中地撰写大量的虚假评论以提高自身评价或诋毁竞争对手的现象。这些虚假评论对商店的点评系统进行了攻击,直接影响到数据的真实性,并且对消费者的消费行为进行了误导。相应地,大量的虚假评论对基于点评系统的推荐系统也产生了影响,被影响的推荐系统将给消费者提供错误的推荐。通过对虚假评论的有效检测,过滤掉虚假评论及虚假用户,能在一定程度

到稿日期:2018-07-27 返修日期:2018-10-29 本文受浙江省自然科学基金项目(LY19F020025),宁波市"科技创新2025"重大专项项目(2018B10063),基于GAN的信号识别项目,深度学习增强识别项目,之江实验室重大科研项目(2019DH0ZX01)资助。

陈晋音 博士,副教授,主要研究方向为深度学习、智能计算、复杂网络和算法安全,E-mail:chenjinyin@163.com;**黄国瀚** 硕士生,主要研究方向为复杂网络和深度学习;**吴洋洋** 硕士生,主要研究方向为数据挖掘和应用、复杂网络和聚类分析;**贾澄钰** 硕士生,主要研究方向为自然语言处理和深度学习。

上还原出相对真实和自然的评论环境，从而对消费者进行正确的消费引导。因此，实现高效的虚假评论过滤至关重要。

当前针对虚假评论的检测方案主要分为两种，即基于评论文本内容的过滤检测系统和基于用户行为特征的过滤检测系统。基于评论文本内容的虚假评论检测又分为 3 种，即基于语法分析的虚假评论检测[1]、基于语义分析的虚假评论检测[2]和基于文本元数据分析的虚假评论检测[3]。基于评论文本内容的虚假检测方案存在局限性，其虽然能在众包平台得到的评论数据集上获得较高的准确率，但是由于真实环境下的虚假评论文本在语义、语法上具有较强的迷惑性，与众包平台得到的评论有较大的区别，因此在真实数据集环境下，基于评论文本内容的检测方式的效果有待提高。其次，对评论文本的理解和分析的不准确性、计算成本高等缺陷，增大了文本检测方法的局限性。

用户自身携带的属性（如用户所拥有的朋友数等）及其行为特征（如最大日评论数、地理位置、个人评价与主流评价之间的评论偏差等）更容易被当成虚假用户检测的评估因素，因为虚假用户在这些行为特征方面与真实用户往往有较大的差异。基于行为特征的检测方法主要分为基于虚假用户群体行为的检测方法和基于虚假用户个体行为的检测方法。基于虚假用户群体行为的检测方法主要通过对虚假用户间的关系进行群组特征提取或聚类，对虚假用户群体进行检测。基于虚假用户个体行为的检测方法有对虚假用户的反常行为特征进行检测的；也有通过构建用户、商店和评论之间关系的图结构，利用迭代计算来进行检测的，本文即是通过这样的方法来进行虚假检测的。然而，由于考虑的因素较少，基于图的过滤算法存在置信度初始值的可信度低、过滤效果欠佳的问题。

综上所述，本文对以下 3 个问题进行研究：

(1)如何获得更为可靠的用户和商店置信度初始值？可靠的用户和商店置信度初始值是对用户进行筛选的关键。是否可以基于已有的条件对用户和商店的初始值进行优化，使图过滤器的构建更加合理？

(2)能否将真实的评论尽可能地保留下来？这是过滤系统发挥作用的关键指标，对后续推荐系统实现精准推荐至关重要。如何有效地辨识真实评论并将其保留下来，从而提高真实评论的保留率？

(3)有效发现虚假用户并将其剔除是判断一个虚假检测系统好坏的重要指标。如何有效地发现虚假用户，提高检测算法对虚假用户的筛选率？

针对以上 3 个研究问题，本文提出了基于双循环图的虚假评论检测算法，主要工作包括：

(1)为了获得更为可靠的用户和商店置信度初始值，本文提出了一种循环利用数据的方法，通过评论置信度和可靠用户分别获得可靠的用户置信度和商店置信度，对用户和商店的置信度初始值进行优化，以构建合理的图过滤器。

(2)设计一种加权图过滤器，考虑用户对商店的个人影响力，设置依据用户对商店的访问记录数的权重函数来表征用户对商店的影响力水平，从而进一步提高图过滤器的合理性。

(3)通过对加权图过滤器的置信度初始值优化，进行数据的二轮过滤，从而提高真实评论的辨识率和虚假用户的筛选率。

2 相关工作

2.1 虚假信息检测算法

自 Liu 等[4]首次提出虚假用户检测[4]以来，虚假用户检测问题就备受关注。Liu 等利用用户、产品和评论的特性构造出了一个分类器，从而预测异常评论。Ott 等[5]通过提取评论文本内容的语言特征对评论进行分类。Mukherjee 等[6]对用户的评论特征和行为特征（如评论相似度、评论偏差、日最大评论数等）进行分析与计算，从而确定可疑的虚假用户。Yoo 等[7]利用 hotel 的 40 条真实评论和 42 条虚假评论来手工比较心理学相关语言之间的差异。Mukherjee 等[8]检测了评论文本、评分和其他元数据中的虚假评论者群体。Li 等利用评论特点和评论者特点，采用协同训练算法对其模型进行训练，从而发现虚假评论者[9]。Fei 等[10]利用信念传播的方法来推断一个评论者是否为虚假评论者。Lim 等[11]定义了 4 种行为模型，这 4 种行为模型分别是针对产品、群体、一般偏差评论和早期偏差评论：通过线性加权的方法，将评论者的评分和评论者的评论行为相结合，以检测出虚假评论。Xu 等[12]认为一个评论人是虚假评论人的概率越大，则其属于虚假评论群组的概率也越大。他们运用了多种成对特征对评论者之间的关系进行挖掘，并利用排序方法对虚假评论群组进行检测。Ye 等[13]构建了评论者网络，运用相邻节点的多样性和节点与网络的自相似性这两类群组特征对评论者的网络足迹进行分析，从而找到评论者的评论反常性，并运用层次聚类的方法检测虚假评论群组。Li 等[14]通过与大众点评合作，首次对大众点评过滤系统过滤的餐厅点评进行了大规模分析，通过评论者发布评论的 IP 地址或所在城市与所评论的餐厅的地理位置进行空间上的分析，从而检测虚假评论者。宋海霞等[15]利用基于自适应聚类的方式进行虚假评论检测。Huang 等[16]根据评论者的写作技能和方式提出了检测专业虚假评论者的方法。Li 等[17]利用虚假用户和真实用户的双模态分布和虚假用户评论的爆发性来鉴别虚假用户。Ye 等[18]认为虚假评论是一种暂时性现象，并设计了一种实时探测器。Wang 等[19]将半监督递归自动编码器应用于垃圾邮件的检测。Narayan 等[20]将监督学习技术应用于垃圾邮件检验，并使用不同的特征集及情绪评分来构建模型。

2.2 基于图的方法

Wang 等[21-22]提出了一种基于评论图的虚假用户检测方法。该方法利用评论、用户和商店之间的相互影响关系建立了一个评论图结构。他们首先提出了评论置信度、商店置信度和用户置信度 3 个评价指标，且每个指标都可通过自身或其他指标计算得出；然后利用节点增强法进行迭代计算，得到各项指标的稳定值；最后根据指标稳定值来对用户进行排名，从而判断虚假评论和虚假用户。实验表明，该方法虽然能适应不同的虚假评论种类，并且能较好地模拟真实的评论环境，但是该方法中的缺陷导致了检测结果存在虚假用户筛选率不高的问题。

王琢等[23]在 Wang 等[22]的基础上，重新定义了评论、评

论人和产品之间的评论图，利用对产品的评论和其他评论间的差异度、其他评论人对该评论的投票数及对该评论有帮助的投票数等，设计了一种逐步淘汰评论人及其评论的 ICE 算法，从而更快收敛得到虚假评论人。另外，由于将商店改为产品，他们引入了更多关于评论作弊模式的特征来刻画评论人及评论的虚假性，从而进一步提高了评论图对虚假评论人的检测精度。

Rayana 等[24]利用用户、产品和评论的图结构，以及马尔科夫随机场模型，对评论者、产品和评论同时进行预测。他们通过这种方法检测虚假用户、虚假评论，并判断产品是否被虚假评论攻击。

Akoglu 等[25]利用用户、产品和评论构建图结构，提出了基于马尔科夫随机场的 FRAUDEAGLE 模型。其中，用户和产品之间利用消极（或积极）的评论作为权重连边进行连接。该算法没有使用评论的文本内容，而仅仅考虑了评论的积极性（或消极性），因此适用于各种类型的评论数据。

其中，由于本文采用了与文献[21,25]中的方法相似的图结构，因此将这两种方法作为对比算法。

3　双循环图过滤检测算法

根据以上问题，本文提出了双循环图过滤检测算法（Double Graph based Review Detection Algorithm，DG-RDA），其基本框架如图 1 所示。通过图过滤算法获得可靠用户和所有用户的初始评论置信度，通过双循环图迭代更新用户和评论置信度；将获得的可靠用户作为数据集，用原始图过滤器更新商店置信度。优化加权图过滤器的商店初始值和用户初始值，使用经过更新初始值的加权图过滤器实现虚假评论的检测。

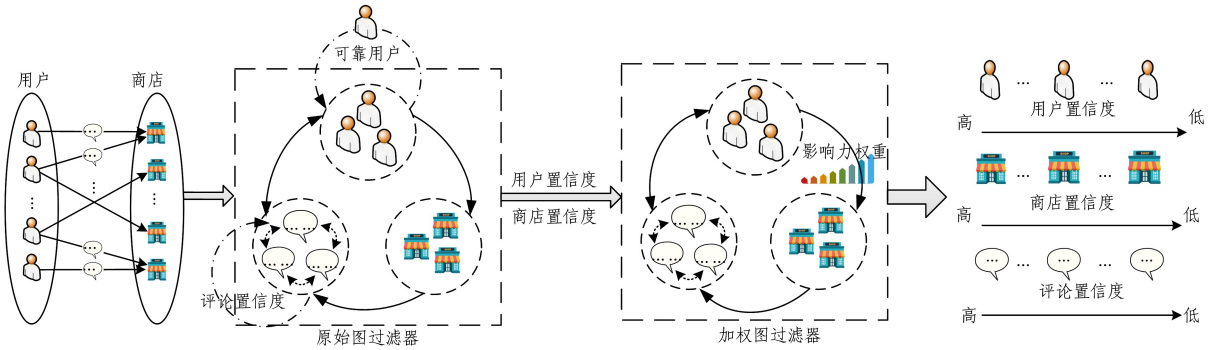


图 1　DG-RDA 算法的框架

Fig. 1　Framework of DG-RDA

3.1　三元加权图的构建

构建的三元加权图，是以评论、用户和商店为节点的有向图。每个用户节点具有指向其撰写的评论节点的边，每个评论节点具有指向其所评论的商店节点的边。其中，每个节点可以附加特征数据，如评论节点具有评论一致性的特征数据；节点与节点之间的边也可以赋予相应的权重，如用户节点和商店节点中可以赋予与用户评论次数相关的权重，用以衡量用户对商店的影响力。同时，该评论图还具有同一用户可以撰写多条对相同或不同商店的评论的特点。评论图的结构如图 2 所示。

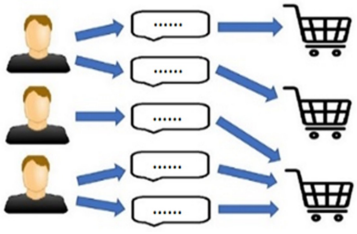


图 2　评论图中各节点间的相互关系

Fig. 2　Relationship between each node in graph

定义 1（评论置信度，Honesty of Reviews）　评论 v 的置信度，表示我们对该评论的置信程度，即该评论的可信度，记为 $H(v)$，取值范围为（−1，1）。其计算方法如下：

$$H_r=\sum_{i=1}^{n_r}H(\alpha_r^i) \tag{1}$$

其中，n_r 表示用户 r 的评论数，$H(\alpha_r^i)$ 表示用户 r 的第 i 条评论的置信度。

$$H(v)=|R(\Gamma_v)|A_n(v,\Delta t) \tag{2}$$

其中，$A_n(v,\Delta t)$ 表示在 Δt 时间内评论的一致度（将在下文详细描述）。Γ_v 为评论 v 对应商店的 ID。$R(\Gamma_v)$ 表示评论 v 所评论商店的置信度。将 $|R(\Gamma_v)|$ 作为 $A_n(v,\Delta t)$ 的放大。当 $|R(\Gamma_v)|$ 很大时，说明该商店要么很好，要么很差，那么当对这家店的评论 v 与其他真实评论的一致度较高，即 $A_n(v,\Delta t)$ 的值较大时，该评论应具有较高的置信度。当 $|R(\Gamma_v)|$ 很小时，则难以判断商店的置信度，评论的真实性也随之下降。

定义 2（用户置信度，Trustiness of Reviewers）　用户 r 的置信度，表示我们对该用户的可信程度，记为 $T(r)$，其取值范围为（−1，1）。$T(r)$ 的计算方法如下：

$$T(r)=\frac{2}{1+e^{-H_r}}-1 \tag{3}$$

其中，H_r 为用户 r 的评论置信度。

定义 3（商店置信度，Reliability of Stores）　商店 s 的置信度，表示商店的可靠程度，即该商店是否值得消费的程度，记为 $R(s)$，其取值范围为（−1，1）。$R(s)$ 的计算方法如下：

$$R(s)=\frac{2}{1+e^{-\theta}}-1 \tag{4}$$

其中：

$$\theta=\sum_{v\in U_s,T(K_v)}Times(\tau,\tau_{\max})T(K_v)(\Psi_v-\mu) \tag{5}$$

其中，$Times(\tau,\tau_{\max})$ 为衡量用户影响力的权重函数，下文将对其进行具体描述；U_s 表示访问过商店 s 的用户集合；Ψ_v 表示用户对商店的具体评分；μ 为阈值参数，用于衡量评分的积极

性或消极性，取值为 3；$T(K_v)$ 表示发表评论 v 的用户的置信度。

一个用户对一家商店评论的影响力与其对该商店的访问次数存在正相关关系，即一个用户对一家商店的访问和评论次数越多，其对商店的影响力就越大，他的评论就更加可信；反之，我们有理由怀疑他是一位虚假用户。因此，考虑到不同用户对同一家商店评论的影响力大小不同，我们设置了权重函数 *Times*，其具体表达式为：

$$Times(\tau_r,\tau_{\max})=\frac{\tau_r}{\tau_{\max}}e^{1-\frac{\tau_r}{\tau_{\max}}} \tag{6}$$

其中，τ_r 为用户 r 对商店 s 的评论次数，$\tau_{\max}$ 为对该商店评论次数最多的用户所发表的评论数。*Times* 函数的图像如图 3 所示。

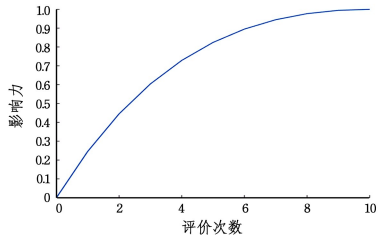


图 3 *Times* 权重函数

Fig. 3 *Times* weighting function

可见，各节点之间具有相互影响的复杂关系：一个用户的置信度取决于其评论的置信度而不是他的评论数量；一个商店的置信度取决于对其评论的用户的打分；一条评论的置信度取决于该评论与周围评论的一致度、其评论的商店的可信度以及撰写评论的用户的可靠程度。其中，在用户置信度和商店置信度之间的连边上赋予与用户对同一家商店的评分次数相关的用以衡量用户对商店的影响力的权重。在给定的合理初始值下，利用节点增强法，经过多次迭代，各节点的置信度将会达到一个稳定值。最后，根据评论/用户置信度排名，可以将置信度较低的评论/用户认定为虚假评论/用户。

3.2 评论一致度的重定义

在原始图过滤算法中，对评论一致性的计算如下所示：

$$A(v,\Delta t)=\sum_{i\in S_{v,a}}T(K_i)-\sum_{j\in S_{v,a}}T(K_j) \tag{7}$$

$$A_n(v,\Delta t)=\frac{2}{1+e^{-A(v,\Delta t)}}-1 \tag{8}$$

其中，$T(K_i)$ 表示相似集 $S_{v,a}$ 内用户的置信度，$T(K_j)$ 表示非相似集 $S_{v,d}$ 内用户的置信度，对相似集 $S_{v,a}$ 与非相似集 $S_{v,d}$ 的划分有如下定义：

$$S_v=\{i\mid \Gamma_i=\Gamma_v,|t_i-t_v|\leqslant\Delta t\},\forall i,j\in S_v \tag{9}$$

$$S_{v,a}=\{i\mid|\Psi_i-\Psi_j|<\delta\} \tag{10}$$

$$S_{v,d}=S_v\backslash S_{v,a} \tag{11}$$

其中，S_v 为在时间 Δt 内，商店的所有评论集合；Γ_i 表示评论 ID。式(10)和式(11)表示一条打分信息 Ψ_i 与周围的打分分值 Ψ_j 相差小于 1 时被划分为相似集 $S_{v,a}$，否则被划分为非相似集 $S_{v,d}$。由此可以得出：当用户打分为 3 分时，它的相似集为 2 分和 4 分，而 1 分和 5 分则被归为同一类，这显然是不合理的。本文通过统计一家商店的评分分布，得出大众对该商店的主流评分，容易发现，在 5 分制打分系统中，大众对商店的主流评分往往存在喜恶的偏向性，而非中肯的 3 分。因此，我们重新对相似集和非相似集进行如下划分：

$$S_{v,a1}=\{i\mid|\Psi_i-\Psi_j|<\delta\} \tag{12}$$

如果评分为 4 分的用户数多于评分为 2 分的用户数，说明用户对该商店有着积极的评价，即可将 5 分纳入相似集：

$$S_{v,a2}=\{i\mid\Psi_i=5\} \tag{13}$$

如果评分为 4 分的用户数少于评分为 2 分的用户数，说明用户对该商店有着消极的评价，即可将 1 分纳入相似集：

$$S_{v,a2}=\{i\mid\Psi_i=1\} \tag{14}$$

则

$$S_{v,a}=S_{v,a1}\bigcup S_{v,a2} \tag{15}$$

$$S_{v,d}=S_v\backslash S_{v,a} \tag{16}$$

当主流评分在 3 分及以上时，将 5 分也归入相似集；当主流评分低于 3 分时，将 1 分也归入相似集。按照上面的方法，在对评论一致性的计算中，获得了更为合理的相似集和非相似集。

3.3 用户置信度的优化

本文通过对实际环境的分析发现，传统的图过滤器中将每一个用户的置信度初始值全部设置为 1 的方法不太合理，该方法忽略了用户的个性特征，即每个用户都应该具有符合他们特征的不同置信度初始值。因此，本文首先通过传统的图过滤器获得用户所有评论的置信度，并将其作为二次迭代的初始值，再通过原始图过滤器重新进行迭代，以获得用户的置信度，从而获得更加接近于真实的用户置信度。

3.4 可靠用户的选择

本文需要对数据集中的用户进行筛选，取出可靠用户用于优化商店的置信度。因此，以 YelpChi 数据集为例，对通过原始图过滤器产生的用户置信度频度分布进行分析，得到图 4 所示的用户置信度频度分布图。

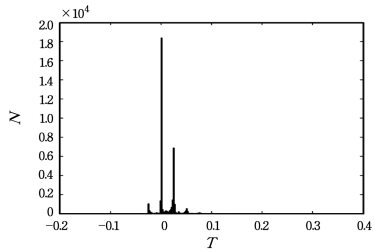


图 4 用户置信度频度分布

Fig. 4 User's trustiness frequency distribution

通过之后的实验分析可知：大量的虚假用户的置信度应当有较低的评分，即在图中表现为 0 及以下的分布；而大量的真实用户节点的置信度普遍高于虚假用户节点的置信度。因此，用户节点置信度的频度分布将呈现如图 4 所示的双峰形态。实验证明，在其他数据集上，用户节点的置信度频度分布也有上述特点。那么，如果将置信度阈值设定在双高峰之间的低谷中，则能有效提取出原始图过滤器筛选出的较为真实的用户群体。同样地，将评分阈值设置在 0 以下的低谷中，则能有效提取出原始图过滤器筛选出的虚假用户群体。通过对用户节点的置信度频度分布进行分析，最终能获得一批置信度较为可靠的用户，即可靠的真实用户和可靠的虚假用户。

3.5 可靠用户优化商店置信度

在自然环境中，人们对商店的评价往往是不同的，这取决于该商店提供的服务好坏和产品的质量高低。在不同用户评

价的环境下，每一家商店也应当具有不同的置信度。因此，在原始图过滤器中将商店置信度初始值设置为 1 的方法也不合理。本文将为各个商店赋予不同的合理置信度。如 3.4 节所述，我们提取了原始图过滤器所筛选的可靠用户，并对这些可靠用户进行预处理。我们将具有高置信度评分的用户的置信度重新设置为 1，将具有低置信度评分的用户的置信度重新设置为 −1，将设置好的可靠用户的置信度作为原始图过滤器的初始值。将所提取的可靠用户作为原始图过滤器的输入，输出得到商店置信度，从而获得真实场景下较为合理的商店置信度。

3.6　加权图过滤器

根据 3.3 节和 3.5 节的分析，获得了较为真实的用户置信度和商店置信度。因此，本文将部分真实用户和商品的置信度作为 3.1 节所述加权图过滤器的用户置信度和商店置信度的初始值。此时，加权图过滤器的输入为整个数据集，输出分别为用户置信度、评论置信度和商店置信度。最后，我们将评论置信度和用户置信度进行排名，具有较低的置信度评分的评论可被视作虚假评论。

4　实验结果与分析

4.1　数据集

本文通过 Yelp 数据集来验证检测算法。Yelp 是美国著名的商户点评网站，点评者会给出多少星级的评价，通常点评者都是亲身体验过该商户服务的消费者。本次实验使用的数据集为文献[24]使用的数据集，其中包含 3 个数据集，分别为 YelpChi，YelpNYC 和 YelpZip。每个数据集包含了用户、商店及用户对商店的打分评价和真假类标属性，表 1 列出了数据集的信息。

表 1　Yelp 数据集的信息统计

Table 1　Basic statistics of Yelp datasets

数据集	评论数	用户数	商店数
YelpChi	67 395	38 063	201
YelpNYC	359 052	160 225	923
YelpZip	608 598	260 277	5 044

由表 1 可知，YelpChi 数据集的虚假用户约占用户总数的 20.33%，虚假评论约占评论总数的 13.23%。YelpNYC 数据集的虚假用户约占用户总数的 17.79%，虚假评论约占评论总数的 10.27%。YelpZip 数据集的虚假用户约占用户总数的 23.91%，虚假评论约占评论总数的 13.22%。

4.2　评价指标

本文使用的评价指标分别为 AUC，F-measure，TopK 下的真实评论的比例及 BottomK 下的虚假用户筛选率。其中 AUC 为 ROC 曲线下的面积，因此，我们需要用到假正类率（False Positive Rate，FPR）和真正类率（True Positive Rate，TPR）。同时，F-measure 可被视为准确率（Precision）和召回率（Recall）的加权调和平均值。这几个评价指标的计算公式如表 2 所列。

表 2　评价指标的计算公式

Table 2　Calculation formula of evaluation indexes

评价指标	计算公式
假正类率（FPR）	$\frac{FP}{FP+TN}$
真正类率（TPR）	$\frac{TP}{TP+FN}$
准确率（Precision）	$\frac{TP}{TP+FP}$
召回率（Recall）	$\frac{TP}{TP+FN}$
F-measure	$\frac{2*Precision*Recall}{Precision+Recall}$

4.3　评论真实性检测实验分析

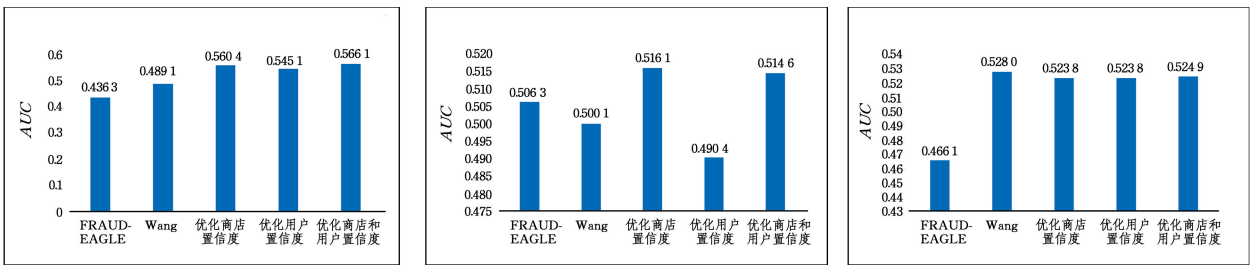
本文使用 Akoglu 等的 FRAUDEAGLE 算法[25]、Wang 等的基于图过滤的方法[21]、优化商店置信度及优化用户置信度的方法作为对比算法。

表 3 和图 5 给出了本文算法与各对比算法间的评论置信度 *AUC* 值的比较。

表 3　不同算法的 *AUC* 值比较

Table 3　Comparison of *AUC* for different methods

	AUC		
	YelpChi	YelpNYC	YelpZip
FRAUDEAGLE[30]	0.436 3	0.506 3	0.466 1
WANG[26]	0.489 1	0.500 1	0.521 5
优化商店置信度	0.560 4	0.516 1	0.523 8
优化用户置信度	0.545 1	0.490 4	0.523 8
优化商店和用户置信度	**0.566 1**	**0.524 6**	**0.524 9**



(a) YelpChi　　(b) YelpNYC　　(c) YelpZip

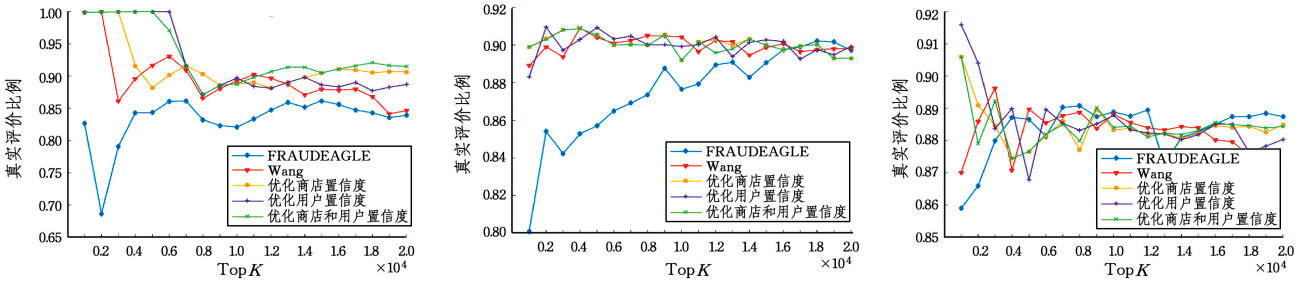
图 5　各方法与评论置信度相关的 *AUC* 对比

Fig. 5　Comparison of *AUC* for different methods related to comment confidence level

对比同一个数据集下各算法的 *AUC* 可以看出，同时优化商店和用户置信度的方法取得了最优效果，尤其在 YelpChi 数据集上，其远远优于其他对比算法。在 YelpNYC 和 YelpZip 数据集上，优化商店和用户置信度的方法与其他对比算法的 *AUC* 接近，但是仍然具有一定的优势。在对比 3 个数据集之后可以发现，仅对商店置信度初值或用户置信度初值进行优化时，其结果的 *AUC* 值存在较大的浮动，如在仅优化用户置信度的情况下，两种方法在 YelpChi 和 YelpNYC 上表现出现了两个不同的极端。也就是说，同时优化商店和用户置信度的方法对不同的数据集具有更强的适应性。同时，从

3 个数据集的结果来看，优化商店置信度的方法的效果接近于优化商店和用户置信度的方法，相比其他对比算法具有优势。而 FRUADEAGEL 算法和 Wang 等人的原始图过滤算法的效果总体上表现不佳。

在推荐系统中，往往是基于评分较高的评论对商品或商店进行推荐，而不是对其进行一一罗列。因此，选取评分较高的评论来检测算法过滤的效果具有较高的参考价值，故本文中采用了 TopK 指标来衡量基于双循环图过滤检测算法的有效性。TopK 指标表示得分靠前的 K 条评论中真实评论的占比。图 6 给出了真实评论在 TopK 比例指标下基于双循环图过滤检测算法处理 3 种不同数据集后所得到的结果。



(a) YelpChi (b) YelpNYC (c) YelpZip

图 6 评论相关的 TopK 中的真实评论比例

Fig. 6 Real review rate in TopK for review ranking

分别对比图 6 中 3 个数据集的实验结果可以看出，在双循环图过滤算法处理 3 个数据集时，优化商店和用户置信度的方法往往能取得较优的结果，即在 Top$K$ 条评论中，真实评论的比例相比其他对比算法来说具有一定的优势。这说明该算法在真实用户的辨别上具有一定的效果。通过分析发现，在对评分更为靠前的评论进行筛选时，FRAUDEAGLE 算法的效果较差，且有较大的波动。相比于 FRAUDEAGLE 算法，双循环图过滤算法得到的结果更为稳定，且能获得较高的真实评论比例。另外，优化商店置信度和优化用户置信度的方法虽然在 3 个数据集上的表现都不佳，但是其效果普遍与原始图过滤器接近，在部分区域上有一定的优势，这也能说明本文提出的双循环图过滤算法能够有效地辨识真实用户。也就是说，优化商店和用户置信度的方法在实际应用中也具有一定意义。

4.4 虚假用户检测实验分析

由于本文数据集中含有评论的类标，因此发表过一次虚假评论的用户即可视为虚假用户，由此我们获得了用户的类标。为了验证双循环图过滤算法的有效性，用 BottomK 中的虚假用户筛选率作为本文算法的评价指标，其定义为：得分靠后的 K 个用户中虚假用户数占总虚假用户数的比例。实验结果如表 4 所列。

表 4 用户相关的 BottomK 中的虚假用户筛选率

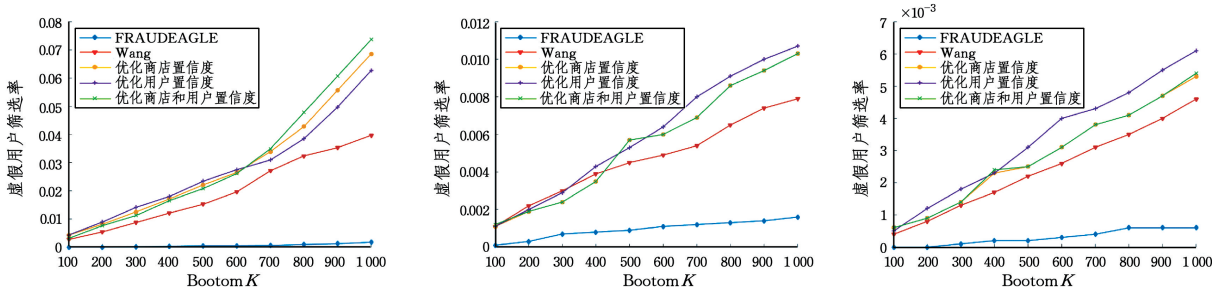
Table 4 Spammer screening rate in BottomK for user ranking

	算法	100	200	300	400	500	600	700	800	900	1000
YelpChi	F	0	0	0.0001	0.0003	0.0005	0.0005	0.0006	0.0009	0.0012	0.0017
	W	0.0027	0.0054	0.0087	0.0121	0.0152	0.0196	0.0271	0.0324	0.0353	0.0397
	R	0.0043	0.0081	0.0125	0.0171	0.0221	0.0265	0.0339	0.0428	0.0557	0.0686
	T	**0.0043**	**0.0089**	**0.0141**	**0.018**	**0.0234**	**0.0275**	0.0309	0.0385	0.0497	0.0627
	T&R	0.0031	0.0076	0.0112	0.0165	0.0208	0.0261	**0.0349**	**0.0478**	**0.0607**	**0.0737**
YelpNYC	F	0.0001	0.0003	0.0007	0.0008	0.0009	0.0011	0.0012	0.0013	0.0014	0.0016
	W	0.0011	0.0022	**0.0030**	0.0039	0.0045	0.0049	0.0054	0.0065	0.0074	0.0079
	R	0.0011	0.0019	0.0024	0.0035	0.0057	0.006	0.0069	0.0086	0.0094	0.0103
	T	0.0011	0.002	0.0029	**0.0043**	0.0053	**0.0064**	**0.008**	**0.0091**	**0.0100**	**0.0107**
	T&R	**0.0012**	0.0029	**0.0030**	0.0035	**0.0057**	0.006	0.0069	0.0086	0.0094	0.0103
YelpZip	F	0	0	0.0001	0.0002	0.0002	0.0003	0.0004	0.0006	0.0006	0.0006
	W	0.0004	0.0008	0.0013	0.0017	0.0022	0.0026	0.0031	0.0035	0.004	0.0046
	R	0.0006	0.0009	0.0014	0.0023	0.0025	0.0031	0.0038	0.0041	0.0047	0.0053
	T	0.0005	**0.0012**	**0.0018**	0.0023	**0.0031**	**0.004**	**0.0043**	**0.0048**	**0.0055**	**0.0061**
	T&R	**0.0006**	0.0009	0.0014	**0.0024**	0.0025	0.0031	0.0038	0.0041	0.0047	0.0054

注：F 表示 FRAUDEAGLE[25]，W 表示 Wang[21]的算法，R 表示优化商店置信度，T 表示优化用户置信度，T&R 表示优化用户和商店置信度

首先将输出的用户置信度由高到低进行排名，则虚假用户应为置信度低的用户。表 4 第一列表示在该排名中位于倒数 K 个的用户。从表中可以发现，随着 K 值的增大，虚假用户的筛选率逐渐提高。在不同数据集上分别比较各算法的虚假用户筛选率可以发现，FRAUDEAGLE 算法所获得的结果始终处于较低的水平，原始图过滤算法所得到的结果有着与双循环图过滤算法相似的增长趋势，但是并没有取得最好的效果。同时我们也发现，K 值越大时，原始图过滤算法与双循环图过滤算法之间的差距也越大，即在对更多的低置信度用户进行筛选时，双循环图过滤算法能得到

更优的筛选结果。在 YelpChi 数据集上，优化商店和用户置信度的方法有着最优的效果，优化商店置信度和优化用户置信度的方法也有着较好的表现，双循环图过滤算法与原始图过滤算法和 FRAUDEAGLE 算法相比具有明显的优势。在 YelpNYC 和 YelpZip 数据集上，由于数据分布的原因，存在部分用户置信度相同的情况，因此优化商店和用户置信度的方法的效果并没有在 YelpChi 数据集上那么明显，但是在整体效果上仍然与优化用户置信度的方法较为接近，并且优于原始图过滤算法和 FRAUDEAGLE 算法。这说明本文提出的双循环图过滤检测算法能对虚假用户进行有效的检测。BottomK 中的虚假用户筛选率的增长趋势如图 7 所示。



(a) YelpChi　　(b) YelpNYC　　(c) YelpZip

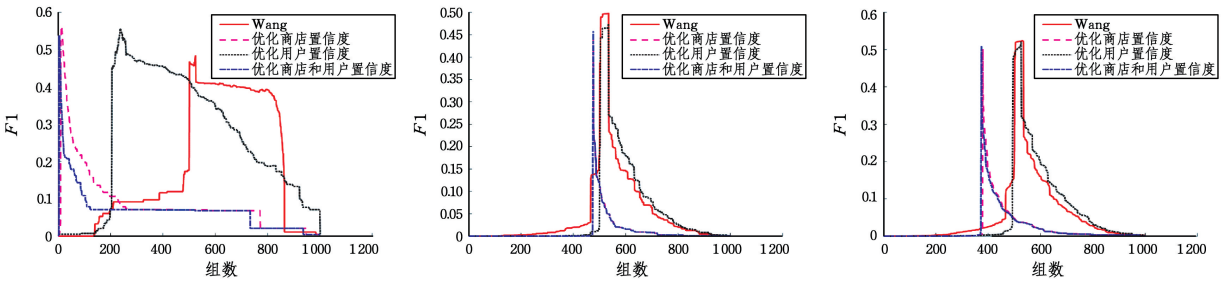
图 7　用户相关的 BottomK 中的虚假用户筛选率

Fig. 7　Spammer screening rate in BottomK for user ranking

4.5　收敛速度分析

本节将对双循环图过滤算法的收敛速度与原始图过滤算法进行比较分析。F-measure 指标可以看作准确率和召回率的加权调和平均值。当同时注重准确率和召回率时，我们可以采用 F1 指标来表征算法的收敛速度，即使准确率和召回率达到最优的平衡值的速度。

图 8 给出了 3 个数据集下的 F1 指标的变化趋势。从图中可以看出，无论是在哪个数据集下，优化商店和用户置信度的算法的 $F1$ 值能更快地达到最大值，这也说明了该算法比其他对比算法能更快地收敛。同时我们也能发现，优化商店置信度的算法也能较快地收敛，但是比优化商店和用户置信度的算法略慢一些。



(a) YelpChi　　(b) YelpNYC　　(c) YelpZip

图 8　收敛速度分析

Fig. 8　Convergence velocity analysis

4.6　参数敏感性分析

将本文提出的算法进行参数敏感性测试和分析。选择的参数为算法计算置信度时的迭代次数和评分相似阈值 δ，实验数据集为 YelpChi。

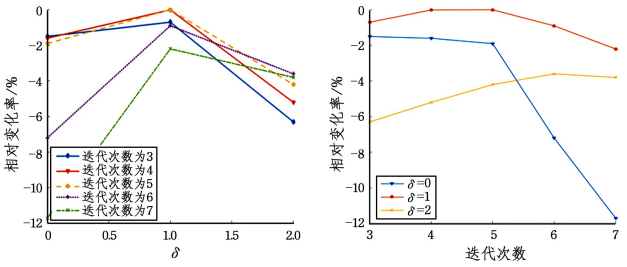
表 5 列出了随迭代次数和评分相似阈值 δ 的变化，AUC 指标与标称值的差值与相对变化率，并以此作为参数敏感性指标。表中 AUC 标称值选择实验采用的迭代次数和评分相似阈值 δ，即迭代次数为 5、评分相似阈值 δ 为 1 时的 AUC 值。

表 5、图 9 和图 10 说明了两种参数单独变化时 AUC 的相对变化率：1）在 δ 值相同的情况下，迭代次数变化时的 AUC 相对变化率；2）迭代次数相同的情况下，δ 值变化时的 AUC 相对变化率。分析可知，前一种情况下的 AUC 相对变化率比后一种的 AUC 相对变化率低，由此我们可以得出本算法对评分相似阈值 δ 的敏感性更高的结论。分析图 9 和图 10 可知，最佳的评分相似阈值 δ 为 1，最佳迭代次数为 4 或 5。

表 5　AUC 指标因迭代次数和评分相似阈值变化而发生的变化

Table 5　Changes of AUC cause by changing number of iteration and the changes of δ

AUC 标称值＝0.5661		迭代次数				
		3	4	5	6	7
$\delta=0$	AUC	0.5577	0.5573	0.5551	0.5252	0.5
	与标称值的差值	−0.0084	−0.0088	−0.011	−0.0409	−0.0661
	相对变化率/%	−1.5	−1.6	−1.9	−7.2	−11.7
$\delta=1$	AUC	0.5619	0.566	0.5661	0.5608	0.5538
	与标称值的差值	−0.0042	−0.0001	0	−0.0053	−0.0123
	相对变化率/%	−0.7	−0.01	0	−0.9	−2.2
$\delta=2$	AUC	0.5306	0.5369	0.5425	0.5456	0.5447
	与标称值的差值	−0.0355	−0.0292	−0.0236	−0.0205	−0.0214
	相对变化率/%	−6.3	−5.2	−4.2	−3.6	−3.8



(a)迭代次数不变而 δ 变化时的 AUC 相对变化率　(b)δ 不变迭而代次数变化时的 AUC 相对变化率

图 9　δ 对相对变化率的影响

Fig. 9　Effect of δ on relative change rate

结束语　本文针对提高用户、评论和商店置信度，提高虚假用户筛选率和真实评论的辨识率等问题，提出了双循环图过滤算法。从实验结果来看，本文提出的算法能够有效地对真实评论进行辨识，并且能有效地对虚假用户进行剔除。但是，本文提出的过滤算法没有考虑更多因素，无法达到更好的虚假用户筛选效果。因此，在接下来的工作中，应增加算法考虑的因素，利用更多自然条件下用户所带有的属性，并进一步提高过滤器的准确率。

参 考 文 献

[1] LI J, OTT M, CARDIE C, et al. Towards a General Rule for Identifying Deceptive Opinion Spam[C]// Meeting of the Association for Computational Linguistics. Baltimere, USA, 2014: 1566-1576.

[2] LAU R Y K, LIAO S Y, KWOK C W, et al. Text mining and probabilistic language modeling for online review spam detection [J]. ACM Transactions on Management Information Systems, 2012, 2(4): 1-30.

[3] LI F, HUANG M, YANG Y, et al. Learning to identify review spam[C]// International Joint Conference on Artificial Intelligence. AAAI Press, 2011: 2488-2493.

[4] JINDAL N, LIU B. Opinion spam and analysis[C]// International Conference on Web Search & Data Mining. ACM, 2008: 219-230.

[5] OTT M, CHOI Y, CARDIE C, et al. Finding deceptive opinion spam by any stretch of the imagination[C]// Meeting of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2011: 309-319.

[6] MUKHERJEE A, VENKATARAMAN V, LIU B, et al. Fake review detection: Classification and analysis of real and pseudo reviews[D]. Chicago: University of Illinois, 2013.

[7] YOO K H, GRETZEL U. Comparison of deceptive and truthful travel reviews[M]// Information and Communication Technologies in Tourism, 2009. Vienna: Springer, 2009: 37-47.

[8] MUKHERJEE A, LIU B, GLANCE N. Spotting fake reviewer groups in consumer reviews[C]// International Conference on World Wide Web. ACM, 2012: 191-200.

[9] LI F, HUANG M, YANG Y, et al. Learning to identify review spam[C]// International Joint Conference on Artificial Intelligence. AAAI Press, 2011: 2488-2493.

[10] FEI G, MUKHERJEE A, LIU B, et al. Exploiting burstiness in reviews for review spammer detection[C]// Seventh International AAAI Conference on Weblogs and Social Media. Menlo Park: AAAI press, 2013.

[11] 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. ACM, 2010: 939-948.

[12] XU C, ZHANG J. Combating product review spam campaigns via multiple heterogeneous pairwise features[C]// Proceedings of the 2015 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2015: 172-180.

[13] YE J, AKOGLU L. Discovering Opinion Spammer Groups by Network Footprints[C]// Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Cham: Springer, 2015: 97-97.

[14] LI H, CHEN Z, MUKHERJEE A, et al. Analyzing and detecting opinion spam on a large-scale dataset via temporal and spatial patterns[C]// Ninth International AAAI Conference on Web and Social Media. AAAI, 2015.

[15] SONG H X, YAN X, YU Z T, et al. Detection of Fake Reviews Based on Adaptive Clustering[J]. Journal of Nanjing University (Natural Science), 2013, 49(4): 433-438. (in Chinese)
宋海霞，严馨，余正涛，等. 基于自适应聚类的虚假评论检测[J]. 南京大学学报(自然科学版)，2013，49(4)：433-438.

[16] HUANG J, QIAN T, HE G, et al. Detecting Professional Spam Reviewers[M]// Advanced Data Mining and Applications. Berlin: Springer, 2013: 288-299.

[17] LI H, FEI G, SHAO W, et al. Bimodal Distribution and Co-Bursting in Review Spam Detection[C]// International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2017: 1063-1072.

[18] YE J, KUMAR S, AKOGLU L. Temporal opinion spam detection by multivariate indicative signals[C]// Tenth International AAAI Conference on Web and Social Media. AAAI, 2016.

[19] WANG B, HUANG J, ZHENG H, et al. Semi-Supervised Recursive Autoencoders for Social Review Spam Detection[C]// International Conference on Computational Intelligence and Security. IEEE, 2017: 116-119.

[20] NARAYAN R, ROUT J K, JENA S K. Review Spam Detection Using Opinion Mining[C]// Progress in Intelligent Computing Techniques: Theory, Practice, and Applications. Singapore: Springer, 2018: 273-279.

[21] WANG G, XIE S, LIU B, et al. Review Graph Based Online Store Review Spammer Detection[C]// IEEE International Conference on Data Mining. IEEE, 2011: 1242-1247.

[22] WANG G, XIE S, LIU B, et al. Identify Online Store Review Spammers via Social Review Graph[J]. ACM Transactions on Intelligent Systems & Technology, 2012, 3(4): 1-21.

[23] WANG Z, LI Z, XU Y, et al. Detecting Product Review Spammers Based on Review Graphs [J]. Computer Science, 2014, 41(10): 295-299. (in Chinese)
王琢，李淮，徐野，等. 基于评论图的虚假产品评论人的检测[J]. 计算机科学，2014，41(10)：295-299.

[24] RAYANA S, AKOGLU L. Collective Opinion Spam Detection: BridgingReview Networks and Metadata[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015: 985-994.

[25] AKOGLU L, CHANDY R, FALOUTSOS C. Opinion fraud detection in online reviews by network effects[C]// Seventh International AAAI Conference on Weblogs and Social Media. 2013.