计算机科学
COMPUTER SCIENCE

# 众包中基于CIDA和PI-Cosine的双向质量控制策略

刘庆菊, 潘庆先, 童向荣, 于嵩, 潘亚楠

引用本文

刘庆菊, 潘庆先, 童向荣, 于嵩, 潘亚楠. 众包中基于CIDA和PI-Cosine的双向质量控制策略[J]. 计算机科学, 2023, 50(10): 282-290.

LIU Qingju, PAN Qingxian, TONG Xiangrong, YU Song, PAN Yanan. Bidirectional Quality Control Strategies Based on CIDA and PI-cosine in Crowdsourcing [J]. Computer Science, 2023, 50(10): 282-290.

---

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221000133

# 众包中基于 CIDA 和 PI-Cosine 的双向质量控制策略

刘庆菊　潘庆先　童向荣　于　嵩　潘亚楠
烟台大学计算机与控制工程学院　山东 烟台 264005
(liuqingju9763@163.com)

**摘　要**　随着移动智能终端的普及,众包采集大规模感知数据变得越来越容易。众包工人的自私性使得他们想通过最少的努力获得最多的报酬,甚至互相勾结、随意提交众包数据,导致众包任务完成质量不高。文中提出了一种基于陪审团的质量控制策略,该机制解决了数据验证问题。针对降低众包质量的行为,在判断是否存在垃圾邮件员工和共谋组织后,使用社区影响力检测算法(CIDA)来检测出共谋团伙领导者及其所在组织,最后使用改进的相似性检测算法(PI-Cosine)筛查垃圾邮件员工。从这两个方面来提高众包数据质量。实验结果表明,所提方法在 accuracy 和 F1-score 衡量指标上相比 Cosine 相似度检测算法提高了 12.3%。

**关键词**:众包;质量控制;CIDA 算法;PI-Cosine 相似性检测;垃圾邮件
**中图法分类号**　TP391

## Bidirectional Quality Control Strategies Based on CIDA and PI-cosine in Crowdsourcing

LIU Qingju, PAN Qingxian, TONG Xiangrong, YU Song and PAN Yanan
School of Computer and Control Engineering, Yantai University, Yantai, Shandong 264005, China

**Abstract**　With the popularity of mobile smart terminals, crowdsourcing to collect large-scale perceptual data becomes easier and easier. The selfishness of crowdworkers makes them want to get the most pay with the least effort, and even collude with each other and submit crowdsourced data arbitrarily, resulting in poor quality of crowdsourced task completion. This paper proposes a jury-based quality control strategy, a mechanism that solves the data validation problem. To address the behaviors that degrade the quality of crowdsourcing, this paper uses the proposed community influence detection algorithm(CIDA) to detect conspiracy leaders and their organizations after determining the presence of spam employees and conspiracy organizations, and finally uses an improved similarity detection algorithm(PI-Cosine) to screen out for spam employees. These two aspects are used to improve the quality of crowdsourcing data. Experiments show that the proposed method improves the accuracy of 12.3% over Cosine similarity detection algorithm in accuracy and F1-score measures.

**Keywords**　Crowdsourcing, Quality control, CIDA algorithm, PI-Cosine similarity detection, Spam

## 1　引言

Howe 在 2006 年首次提出了“众包”概念,他把众包定义为“一种采取传统代理人执行任务的模式并将其以开放的形式外包给非特定大众的问题解决方案”[1]。在大规模面向人类智能服务的应用程序的驱动下,众包利用智能设备的普遍性来获取传感信息的方式为传感数据收集提供了一种新范式。目前众包已被应用到生活的方方面面,如对文本进行图像注释[2]、道路和交通状况监测[3-4]、情感分析[5-7]领域。物联网(Internet of Things, IoT)中,感知能力强大的移动智能设备是移动众包感知环境的主要工具,物联网技术的发展提高了众包的便利性和效率。

一个众包感知任务的成功依赖于每个众包工人的贡献。众包工人的能力参差不齐,提交的数据也千差万别,使得众包中的数据质量问题成为了一项挑战。研究表明,一方面,有一些工人想用较少的努力得到奖励,因此提供一些低质量的感知数据,垃圾信息[8]的比例也因为受到奖励的诱惑而不断增加,大大降低了众包数据的可用性和精确性;另一方面,有些

到稿日期:2022-10-17　返修日期:2023-03-21
基金项目:国家自然科学基金(60903098,61502140,61572418,61472095,62072392);黑龙江自然科学基金(LH2020F023);山东省本科教学改革研究重点项目(Z2022327)
This work was supported by the National Natural Science Foundation of China(60903098,61502140,61572418,61472095,62072392), Natural Science Foundation of Heilongjiang, China(LH2020F023) and Key Research Project of Undergraduate Teaching Reform in Shandong Province (Z2022327).
通信作者:潘庆先(pqx@ytu.edu.cn)

人通过勾结[9]来处理众包任务，这些引入嘈杂数据的行为大大降低了众包数据的可靠性，甚至会导致任务失败。Chen 等[10]考虑到共谋对结果推理的负面影响，提出了一种共谋证明结果推理算法，但是该算法无法获得工作者之间的特定类型的协作关系。Niazi 等[11]提出了一种抗共谋的参与者选择方法，引入了 5 个类似于串通集团成员行为的共谋指标来防止在选定的合适参与者中形成共谋群体。但是该方法的局限性在于没有考虑发布者的共谋可能性。

针对以上问题，为了去除低质量的众包数据，找出低质量众包工人，保证任务的顺利进行，本文提出了一种众包应用中基于陪审团的质量控制策略，由陪审团成员来评审数据，担任数据验证工作，这比由众包平台的管理者验证的方式更可信。针对工人共谋问题，本文提出了一种社区影响力检验算(Community Influence Detection Algorithm，CIDA)，以有效地检测出共谋员工。针对只提交随机或重复的答案而没有串通的垃圾邮件员工，本文提出一种相似度检测算法，通过相似性检测找出垃圾邮件，有效地解决了众包数据结果中垃圾邮件信息比例增加造成众包数据质量低下带来的众包平台失败问题。

本文工作主要如下：

1)对众包工人上传的众包数据进行分析，初步判定众包任务中是否存在共谋行为和垃圾邮件工人；

2)提出了一种基于陪审团的质量控制框架，由陪审团成员来评审数据，担任数据验证工作，这种方式比由众包平台的管理者验证的方式更可信；

3)提出了社区影响力检测算法(CIDA)，利用影响力等级指数(IH)和受牛顿万有引力定律启发改进的吸引力 $f$ 判断社区影响力最大者，进而找到共谋行为，有效地检测出共谋组织；

4)提出了一种引入 Pearson 相关系数的改进相似性检测算法来检测垃圾邮件，对比实验表明，本文提出的相似性检测算法能更好地检测出相似文本，阈值的设定符合垃圾邮件检测算法，与 Cosine 相似度检测算法相比准确率提高了 12.3%。

本文第 2 节介绍相关工作；第 3 节介绍众包网络模型以及现在众包模型中存在的问题；第 4 节介绍基于陪审团的质量控制策略以及众包数据的判断方式；第 5 节介绍算法的具体内容；第 6 节介绍实验内容及结果；最后总结全文并给出未来研究方向。

## 2 相关工作

### 2.1 众包系统

众包系统主要由众包平台(Crowdsourcing Platform)、众包工人(Worker)、任务发布者(Requester)3 部分组成。任务发布者发布任务后，众包工人接收任务，以众包平台为媒介建立联系。任务发布者即为任务请求者，任务请求者以金钱报酬支付、娱乐游戏激励、社交关系激励游戏等激励方式，利用人群的知识、智慧通过互联网无限放大和传播，进而转换成实际收益。这种众包系统模式随着现代互联网的发展体现得淋漓尽致，具有成本低、面对的工人群体大、众包任务多样化等优点，但同时其弊端也逐渐显现出来。随着以通过任务获取经济报酬为目标的众包工人不断增加，不诚实的共谋工人和垃圾邮件发送者也相应增加，他们会为了快速获取任务报酬而提交质量较低的数据或者随意复制其他工人提交的数据，更有甚者会相互勾结，恶意欺骗众包平台。

### 2.2 质量控制

针对提高众包质量的问题，目前的研究工作主要集中在两个方面：1)针对结果质量评估方法进行研究，目的是通过对工作者提交的数据结果进行评估，发现垃圾邮件工作者并对提交的垃圾邮件数据进行筛查；2)对参与众包的工人进行筛查研究，这类工作主要是对工人参与的众包类型进行分析，找到共谋组织者。

针对结果质量评估方法的研究，其中最容易的一种方法就是使用黄金标准数据[12](Golden Standard Data)来衡量数据质。黄金标准数据是用于对比的标准答案，将工人提交的数据与标准答案进行比较可以衡量工人完成任务的质量进而检测出垃圾邮件工作者。Lee 等[13]描述了一种基于签名网络分析的垃圾邮件分类方法。关键是，边缘标志很可能是通过考虑用户的社会关系来确定的，因此垃圾邮件发送者和非垃圾邮件发送者的边缘标志模式之间会有实质性的差异。Madhavan 等[14]讨论了各种机器学习方法(KNN，Naïve Bayes，SVM，Rough Sets Classifiers)对垃圾邮件检测的效率分析，考虑了各种评估指标，如准确度、误差、评估时间、效率等。从机器学习领域的所有分类模型中可以看出，所考虑的每种方法都有其优缺点，混合算法似乎是电子邮件中垃圾邮件检测的最佳可行解决方案，但是其没给出具体的解决方案。

在对参与众包的工人进行筛查研究方面，Xu 等[15]设计了一个新的协议来帮助聚合器收集所有用户的原始数据，同时抵抗共谋攻击。具体来说，他们探索了按位异或同态函数和聚合签名，并设计了一种新的密钥系统来实现抗共谋。Li 等[16]提出一种群智感知应用中基于区块链的激励框架 CrowdBC，设置了应用层、区块链层和存储层。其中应用层利用智能合约进行用户管理和任务管理，区块链层利用矿工验证感知数据，存储层存储任务和解决方案的数据，但是作者没有给出具体的质量估计报酬分配方案。

综上所述，现有的大多数针对结果质量的评估方法需要花费很多时间和金钱，在处理假装拥有良好声誉的垃圾邮件员工时，不能准确地查找到垃圾邮件工人。相比之下，本文提出了一种社区影响力检验算法(CIDA)，能够有效地检测出共谋员工；对于只提交随机或重复的答案而没有串通的垃圾邮件员工，本文提出了一种通过相似性检测找出垃圾邮件的算法，有效地解决了众包工人提交的众包数据中垃圾邮件数据比例高的问题。

## 3 问题描述及众包网络模型

### 3.1 问题定义

**定义 1**(众包任务) 定义为 $T=\langle s_t, e_t, m_t, p_t\rangle$，其中 $s_t$ 为发布任务的起始时间；$e_t$ 为众包任务的终止时间；$m_t$ 为众包任务的约束条件，只有信誉大于一定阈值的众包工人才能接收到该任务；$p_t$ 是工人获得的奖励。

**定义 2**(众包工人) 定义为 $W=\langle s_w, e_w, c_w, q_w\rangle$,其中,$s_w$ 为工人开始执行任务时间;$e_w$ 为工人完成任务时间;$c_w$ 为工人历史信誉值,$p_w$ 为该工人效用。

**定义 3**(任务发布者) 任务发布者可以是个人、组织、政府等,用 requester 表示。任务请求者的效用等于完成任务的真实价值与最终支付给众包工人的总报酬之差,其表达式如式(1)所示:

$$p_r=\begin{cases}p_w-p_t, & \text{任务 Task 被顺利完成}\\ 0, & \text{其他}\end{cases} \tag{1}$$

**定义 4**(共谋工人) 以违背社会规则的方式共享信息或通过共同决策来谋取利益的众包工人被称为共谋工人,其所在团队被称为共谋组织。

**定义 5**(垃圾邮件员工) 垃圾邮件员工只提交随机或重复的答案,没有串通。

本文的基本符号及含义如表 1 所列。

表 1 基本符号

Table 1 Basic symbols

| 符号 | 含义 |
|---|---|
| $T$ | 感知任务 task |
| $W$ | 众包工人 worker |
| $R$ | 请求者 requester |
| $s_t$ | 任务发布时间 |
| $e_t$ | 任务截至时间 |
| $m_t$ | 约束条件 |
| $p_t$ | 工人获得的奖励 |
| $s_w$ | 工人任务开始执行时间 |
| $e_w$ | 工人完成任务时间 |
| $l_w$ | 工人历史信誉值 |
| $p_w$ | 工人效用 |
| $p_r$ | 请求者效用效用 |
| $Jn$ | 陪审团成员 |
| $Q$ | 工人提交答案准确率 |
| $G$ | 图 |
| $V$ | 工人合集 |
| $E$ | 边集 |
| $Sim(w_i,w_j)$ | 工人 $i,j$ 间相似性 |
| $K(i)$ | 节点度 |
| $d(w_i,w_j)$ | 节点距离 |
| $\theta$ | 相似性阈值 |

### 3.2 问题描述

#### 3.2.1 共谋问题描述

图 1 给出了有共谋参与者的众包任务,5 个参与者分别表示为 $W1,W2,\cdots,W5$,其中 $W1$ 和 $W2$ 不相互勾结,为正常员工,而 $W3,W4$ 和 $W5$ 构成了一个相互勾结的小团体。互相勾结的团体会为了利益相互复制答案并提交给众包平台,导致众包平台接收到的数据无意义或质量低下。

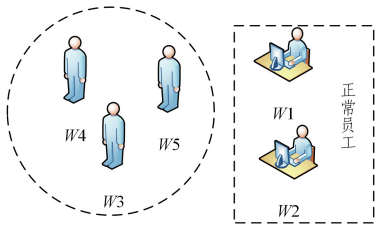



图 1 有共谋参与者的众包任务

Fig. 1 Crowdsourcing missions with collusive participants

由文献[9]可知,为了避免串通行为被发现,共谋者不会提交完全相同的众包数据,然而,仔细检查可能的相关性后发现,对于每项任务,共谋者之间的评级差异很小,共谋者和非共谋者提交的平均评级有差异,共谋导致平均结果偏移了 20%。也就是说,共谋对结果的统计可靠性有很大影响,这也大大影响了众包商的利益。

#### 3.2.2 垃圾邮件员工问题描述

垃圾邮件员工之间虽然没有互相勾结,但是他们会随意提交众包数据。这类众包工人所提交的数据会严重影响众包结果。比如在评级系统中,垃圾邮件员工随意提交的数据会干扰对系统的评判。根据文献[17]得知,产品评论对在线购物者作出购买决定非常有价值。在巨大的利润激励的驱使下,欺诈者故意捏造不真实的评论,歪曲在线产品的声誉。从用户的角度来看,这些评论可能会影响用户对产品的购买决策;从商业角度看,虚假的用户评论可能损失众包商一定的收益。

### 3.3 众包网络模式图

众包网络模型如图 2 所示。众包系统主要由众包平台、众包工人和任务发布者组成,其中任务发布者可以是个人、组织和企业政府。众包任务的完整流程为:1)众包平台发布任务和感知质量要求,众包工人 $l_w$ 利用自己携带的内置传感器(手机、平板、可穿戴设备等)移动设备来完成众包任务;2)当工人的信誉 $l_w$ 大于设定的阈值,那么就允许工人参加众包任务,工人 $w$ 再执行完众包任务并将数据上传至众包平台;3)众包平台分析工人所上传的数据质量,合格的工人将得到相应的报酬 $p_t$,整个过程结束,任务发布者的效用如式(1)所示。众包的这种模式可以将人的知识、智慧通过互联网无限放大和传播,进而转换成实际收益,并创造出巨大的社会财富。

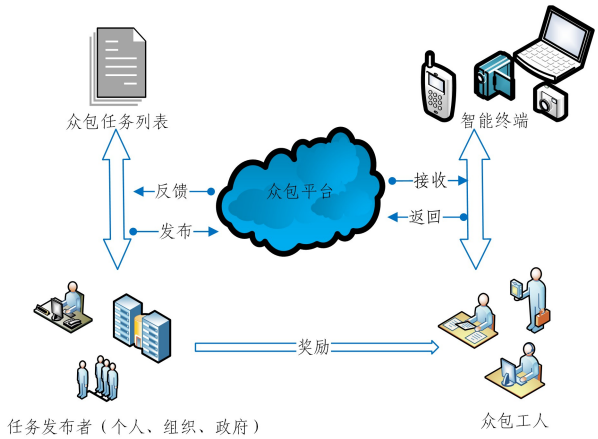



图 2 众包模式图

Fig. 2 Crowdsourcing mode diagram

## 4 基于陪审团的质量控制策略

### 4.1 基于陪审团的质量控制框架

#### 4.1.1 陪审团意义

在美国的法律制度中,陪审团跟原告、被告、法院、警察是完全没有利益冲突的,也几乎无法被收买,近乎 100%的利益中立基本上保证了法律判决中有寻求公平的意愿,这也是陪审团最精妙之处。任何利益相关、利益倾向都会干扰公平的法律。这种法律制度对律师的要求很高,其反过来也促进了

法律的不断完善。受此启发，本文提出了基于陪审团的质量控制策略，一方面解决众包中的欺诈行为，另一方面促使众包工人主动提高自己的任务质量。

众包中，设计好的质量控制机制可以影响工人的参与热情进而控制质量，但是目前质量控制机制的设计存在以下缺点：

1）一些众包工人想以最少的努力来获得最多的奖励从而进行共谋。

2）众包平台的管理人员可能会受利益驱使而滥用职权，出售众包任务信息或者与众包工人联合起来欺骗众包，导致诚实的众包工人得不到相应的奖励。对于众包机制来说，并非个别工人得到的利益越高就越能吸引工人，而是得到的奖励和付出的劳动成正比更具有吸引力。

#### 4.1.2 陪审员的选择

陪审团成员是由非众包平台的管理人员组成的，陪审团成员都具有超高的信誉值。陪审团 $J_n=\{j_1,j_2,\cdots,j_n\}\subseteq L$ 是一组拥有大量成员的评审团队，通过初步筛选的陪审团成员具有团内投票权力。最终陪审团成员为：

$$JER(J_n)=\begin{cases}\text{if } \sum j_i\geqslant\dfrac{n+1}{2}, & \text{陪审员}\\ \text{if } \sum j_i\leqslant\dfrac{n-1}{2}, & \text{非陪审员}\end{cases} \tag{2}$$

选择出来的陪审团成员会对众包工人上传的数据进行质量评估，采用工人匿名的方式，由陪审团成员给出最终工人报酬，然后将报酬标准上传至众包平台，这就有效地避免了一些众包平台的管理人员为了利益而提前将答案泄露给众包工人的问题。

#### 4.1.3 陪审团激励框架

本文提出了一种基于陪审团的质量控制模型（见图 3），该模型将陪审团的思想融入众包平台中。基于陪审团的质量控制模型由任务发布者、众包平台、众包工人和陪审团成员 4 个角色组成。与传统的众包质量控制框架不同的是，本文所提框架的验证工人质量以及报酬分配工作是由陪审团成员完成的。其中任务发布、工作人员选择、数据上传阶段与上文一致，最后一个阶段则由陪审团成员验证工人提交的数据质量，通过质量验证的用户会获得报酬 $p_t$。

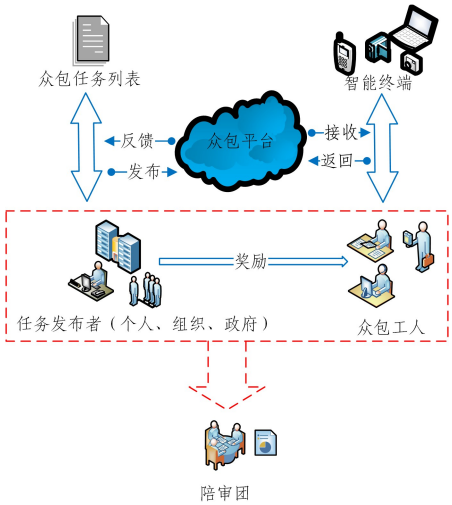



图 3 基于陪审团的质量控制模型

Fig. 3 Quality control model based on jury

众包平台验证数据质量的过程由陪审团成员负责，相应的报酬标准也需经过严格的量化分析。且该质量控制架构不会存在众包平台的管理人员受利益驱使而滥用职权、出售众包任务信息或者与众包工人联合起来欺骗的行为，从而保证了用户数据上传质量的真实性。

### 4.2 感知数据判断

本文用 $Q$ 来表示提交答案的准确率，$Q$ 值的大小由工人答案决定，这是对众包工人中是否有低质量员工参与的初步筛选和考察。计算方法如下：

$$Q(w,r)=\frac{use(w,r)}{total(w,r)},0\leqslant Q(w,r)\leqslant 1 \tag{3}$$

其中，$w$ 和 $r$ 分别表示一个众包工人和一个众包任务请求者；$total(w,r)$ 为 $r$ 发布的任务中 $w$ 提交的答案总数，$use(w,r)$ 为 $r$ 采用 $w$ 的答案个数。我们对文献[18]进行进一步研究分析得知，工人不会选择与降低感知报酬的工人进行合作，本文将所有的工人分为以下 3 类：1）正常员工，正常员工是正常参与众包任务的工人，他们通过自己的努力获得公正的报酬；2）共谋组织，工人选择与可以提高自己感知报酬的高能力者或者事先知道答案的任务请求者共谋，通过共谋以获取更多的报酬；3）垃圾邮件员工，垃圾邮件员工只提交随机或重复的答案，没有共谋。基于产品评论的任务，当恶意的用户相互复制，对评论进行最小的编辑时，就会出现质量下降的问题，给产品的真实质量带来错误的表述。

本文进一步对答案合格率进行研究筛查。参与众包任务的工人可分为 3 类，即正常员工、共谋工人和垃圾邮件员工。其中正常众包工人 $Q$ 值分布如图 4 所示。一个工人在任务上的表现与他的能力密切相关，由于受任务难度和任务类型等因素的影响，员工的 $Q$ 值可能略有波动。

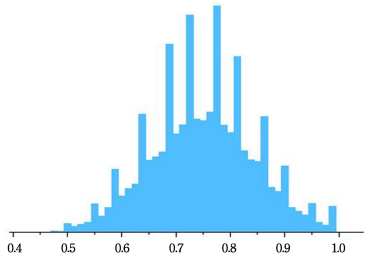



图 4 正常工人

Fig. 4 Normal workers

在共谋任务中，如果工人和任务发布者合谋，由于发布者知道众包数据，因此他们会将数据提前泄露给共谋团伙，那么将会有很大一部分人的准确率很高，接近于 1，如图 5 所示。

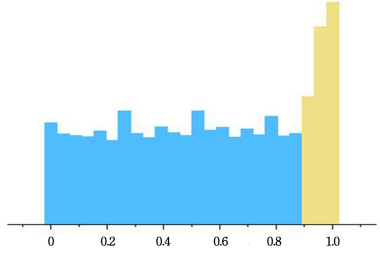



图 5 共谋工人

Fig. 5 Common workers

垃圾邮件员工只提交随机或重复的答案，没有共谋，或者随机复制其他工作者的答案并进行最小的编辑时，垃圾邮件工作者准确率较为平均地分布在各个分值段，如图 6 所示。

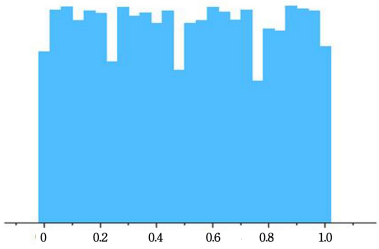



图 6 垃圾邮件员工

Fig. 6 Spam workers

通过对 $Q$ 值的初步分析能得出众包的数据中是否存在共谋工人和垃圾邮件员工。在此基础上，本文有针对性地对众包数据进行质量控制，如果是正常员工，则按照正常的流程给予一定奖励；如果是垃圾邮件员工，则使用垃圾邮件检测算法检测；如果是共谋团伙，那么将使用共谋检测算法来查出共谋组织。本文从这两个方面出发对众包进行质量管控，具体方法将在第 5 章中进行详细讨论。

## 5 算法设计

### 5.1 社会网络

本文定义社会网络 $G=\langle V,E\rangle$ 是工人之间社会联系的无向图。其中，$V=\{w_1,w_2,\cdots,w_m\}$ 是工人的集合（节点）；$E(i,j)=1$（边）表明 $w_i$ 和 $w_j$ 之间存在关系。

**定义 6**（节点度） 节点 $i$ 的度表示如下：

$$k(w_i)=\sum_{w_j}E_{w_iw_j} \tag{4}$$

**定义 7**（结构邻域） 节点的结构邻域定义为节点 $i$ 的最近邻居集合。

$$\Psi(w_i)=\{w_j\}\cup\{w_j\in V\mid(w_i,w_j)\in E\} \tag{5}$$

**定义 8**（改进的相似度） 两个节点之间的公共邻居和所有邻居的比率。其中，$|\Psi(w_i)\cap\Psi(w_j)|$ 表示节点 $i$ 和 $j$ 的共同邻居节点个数，$|\Psi(w_i)\cup\Psi(w_j)|$ 表示 $i$ 和 $j$ 的所有邻居节点总个数。

$$sim(w_i,w_j)=\begin{cases}\dfrac{|\Psi(w_i)\cap\Psi(w_j)|}{|\Psi(w_i)\cup\Psi(w_j)|}, & i,j\text{ 相邻且有公共节点}\\ \dfrac{H}{k(w_i)+k(w_j)}\sum\limits_{c,d\in D_{ij}}S_{cd}, & i,j\text{ 相连但无公共邻居节点}\\ 0, & i,j\text{ 不相连}\end{cases} \tag{6}$$

其中，$k(i)$ 表示节点 $i$ 度；$H$ 表示衰减系数，$H=0.8$；$D_{ij}$ 表示节点 $i$ 和 $j$ 最短路径上的节点集合；$sim(w_i,w_j)$ 为节点 $i$ 和 $j$ 之间的相似度，其值越大，表示节点 $i$ 和 $j$ 的相似度越大，两者共谋的概率就越大。

**定义 9**（节点距离） 两个节点间的距离与它们之间的相似性成反比。随着相似性增加，两节点间的距离减小；随着相似性减小，两节点间距离增大。

$$d(w_i,w_j)=\frac{1}{sim(w_i,w_j)+1} \tag{7}$$

**定义 10**（影响力等级 IH） 该节点与所有邻居的相似性的总和。

$$influence(w_i)=\sum_{w_j\in\Psi(w_i)}sim(w_i,w_j) \tag{8}$$

**定义 11**（吸引力） 牛顿万有引力定律说明任何两个物体之间都存在引力，引力的大小与它们质量的乘积成正比，与它们距离的平方成反比 $F=\dfrac{GMm}{R^2}$。本文定义吸引力就像物理学中的吸引力一样，工人 $i$ 对其他工人 $j$ 的吸引力与两个节点之间距离的平方成反比。

$$f(w_i,w_j)=\frac{k(w_i)}{k(w_j)}*\frac{influence(w_i)}{d(w_i,w_j)^2} \tag{9}$$

**定义 12**（局部共谋领导者） 对于图 $G=(V,E)$ 中的工人 $j$ 来说，查找其附近对自己吸引力最高的节点作为局部共谋领导者。

$$l_{cl}(w_j)=\{w_j\mid\arg\max_{w_i\in N(w_j)}f(w_i,w_j)\} \tag{10}$$

**定义 13**（共谋中心候选集） 共谋中心候选集是相似性较强集合的子集。

$$C(w_j)=\{w_i\mid w_i\in\Psi(w_j),sim(w_j,w_i)\geqslant\overline{sim(w_j)}\} \tag{11}$$

其中

$$\overline{sim(w_j)}=\sum_{w_i\in\Psi(w_j)}sim(w_j,w_i)*\frac{1}{|\Psi(w_j)|} \tag{12}$$

#### 5.1.1 社区共谋检测算法

本文提出的社区共谋检测算法是检测出社区内影响力等级最高的众包工人及其追随者。我们迭代验证社区中的所有节点，并逐个查找工人 $i$ 附近对节点 $i$ 吸引力最高的节点。找到吸引力最高的 $f(i,j)$，则 $j$ 将是 $i$ 的本地领导者。通过距离和领导力计算出吸引力，通过吸引力判定谁是影响力等级最高者，合并之后更新图，找到社区影响力等级最高者即共谋中心。社区共谋检测算法如算法 1 所示。

**算法 1** 社区共谋检测算法（CIDA）

输入：图 G=〈V,E〉

输出：共谋组织者及社区内共谋团伙 C

1. for each edge e=(i,j)∈E do
2. 计算 Newsim($w_i$,$w_j$)和距离 d($w_i$,$w_j$)
3. end for
4. for each node w∈V do

influence($w_i$)=$\sum_{w_j\in\Psi(w_i)}$sim($w_i$,$w_j$)

5. influencers_list={}
6. $f(w_i,w_j)=\frac{k(w_i)}{k(w_j)}*\frac{influence(w_i)}{d(w_i,w_j)^2}$

$l_{cl}(w_j)=\{w_j\mid\arg\max_{w_i\in N(w_j)}f(w_i,w_j)\}$

7. if j exist in influencers_follower do
8. influencers_follower[i].append(j)
9. else
10. influencers_follower[i]=[]
11. influencers_follower[i].push(j)
12. end if
13. end for
14. return C and influencers _follower

找到的局部共谋领导者如果是单独的，我们将在其邻域中找到影响能力等级最高的节点，则该节点将成为新的局部

共谋影响者。否则，我们将检查候选集中的其他项，直至找到候选人集中影响力等级最高的节点，该节点将成为新局部共谋中心。

5.1.2 社区共谋发现结果

用于社区共谋检测的数据集，是 Zachary 对空手道俱乐部的成员关系进行观测得到的数据结果。社区共谋算法对球员之间的关系进行研究。这个网络包含 34 个节点和 78 条边，节点代表所有队员，边代表队员之间的关系。其中，节点 0 和节点 33 分别代表两队的领导者，即教练和校长。在指挥球员的过程中，校长和教练分别带领一队各自进行训练，参加对抗赛。本文认为两支队伍内部为一个共谋团伙，并且其领导者为共谋中心。每一个共谋社区都有一个影响力等级最高的共谋组织者，质量控制机制需要找到这个共谋中心。实验结果如图 7 所示，本文所提方法准确地将成员分为两队，每个队伍中具体包含的节点在表 2 中列出。

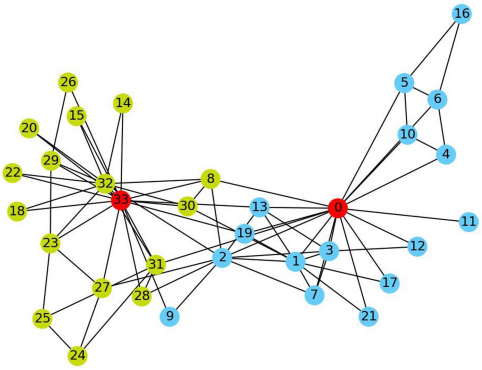



图 7 共谋社区发现

Fig. 7 Collusion community discovery

表 2 共谋检测结果

Table 2 Collusion detection results

| 队伍 | 包含的节点 |
|---|---|
| 0 | 1,2,3,4,5,6,7,9,10,11,12,13,15,16,17,19,21 |
| 1 | 8,18,20,22,23,24,25,26,27,28,29,30,31,32 |

结果显示查到了影响力等级最高的共谋工人，即节点 0 和节点 33。其中，节点 0 代表教练，节点 33 代表校长，符合我们的影响力等级最高标准。以 0 为共谋中心领导的 0,1,2,3,4,5,6,7,9,10,11,12,13,15,16,17,19,21 号球员为一组，以 33 为中心领导的 8,18,20,22,23,24,25,26,27,28,29,30,31,32 号球员为一组。这也符合我们的预期。本文共谋检测算法能准确地检测出共谋领导中心及其跟随者。将本文所提方法与经典的社区检测算法进行对比，对比算法主要包括 GN(Girvan-Newman)算法、SCAN 算法。通过标准化互信息(NMI)、兰德指数(ARI)和聚类纯度(purity)来衡量算法性能，实验结果对比如表 3 所列。

表 3 实验结果对比

Table 3 Comparison of experimental results

| 算法 | Zachary 数据集 | | |
|---|---|---|---|
| | *NMI* | *ARI* | *Purity* |
| Ours | 0.862 | 0.802 | 0.910 |
| SCAN | 0.564 | 0.329 | 0.769 |
| Girvan-Newmann | 0.707 | 0.678 | 0.894 |

从 NMI、ARI 和纯度的结果可以看出，CIDA 在各项指标上均优于另外两种方法。

### 5.2 垃圾邮件筛查算法

文本相似性[19]在自然语言处理[20]问题中普遍存在。本文使用改进的相似度检测算法检测出相似的众包数据，进而检测出垃圾邮件。

基于表面文本的相似性计算筛查方法，其原理比较简单且很容易实现，同时该方法也是其他相似性计算方法的基础。余弦相似性是通过测量两个词向量之间的余弦角来计算的。

$$\cos(w_i,w_j)=\frac{w_1 w_2}{\| w_1 w_2 \|}=\frac{\sum_{i=1}^{n} W_{1i}W_{2i}}{\sqrt{\sum_{i=1}^{n}(W_{1i})^2}\sqrt{\sum_{i=1}^{n}(W_{2i})^2}} \tag{13}$$

基于字符匹配的相似性计算方法是将文本分解为字的集合，以字符间的变化程度作为相似度大小的判定结果，Jaro 距离是两个字符串 $S_A$ 和 $S_B$ 之间的公共字符数 $m$ 和换位数 $t$，其中 l 是字符串开头公共前缀的长度，最多 4 个字符，Winlker 将 $\rho$ 定义为 0.1。

$$d_{\text{jaro}}=\frac{1}{3}\left(\frac{m}{|S_A|}+\frac{m}{|S_B|}+\frac{m-t}{|m|}\right) \tag{14}$$

$$sim_{\text{jaro}}=d_{\text{jaro}}+(lp(1-d_{\text{jaro}})) \tag{15}$$

表 4 相似性算法性能对比

Table 4 Performance comparison of similarity algorithms

| Algorithms | Cosine | Jaro-Winkler | Jaro |
|---|---|---|---|
| Xiao ming/xiao ming | 0.874 | 0.925 | 0.925 |
| xiao ming/ming xiao | 0.999 | 0.555 | 0.555 |
| xiao ming is a boy/ xiao ming | 0.852 | 0.900 | 0.833 |
| xiao ming/xiao ming is a boy | 0.852 | 0.900 | 0.833 |
| aaaaaa/aaaaaa bbb | 0.707 | 0.920 | 0.866 |
| xiao ming is tall /xiao ming is not short | 0.870 | 0.892 | 0.819 |
| She go to school on foot/ She walks to school | 0.750 | 0.821 | 0.701 |

从表 4 所列的 7 个例子中可以看出不同相似性计算方法所得到的相似性百分比。Jaro 方法在文档比较方面的主要影响因素是其不对称性。Jaro-Winkler 虽然考虑了相同的前缀对结果的影响，但是该方法只适用于短文本，不适用于长文本。这些方法仅衡量了文本表面的相似度，与第 6 行的句子表达的是同一个意思，但是可以看出，3 种方法均不能很好地检测出其中的相似性，而在众包任务中，检测工人所提交的众包数据不仅仅限于表面的相似度。

从 3.2.1 节可知，为了避免共谋行为被发现，共谋者不会提交完全相同的众包数据，但是仔细筛查后可以发现，他们提交的数据差异性很小。上文的相似度检测方法忽略了文本结构信息问题，因此本文提出了改进的余弦相似度检测方法。本文通过从文档中提取特定的词性(POS)模式和采用词形还原改进了简单余弦相似性度量，除此之外，还引入 Pearson 相关系数(皮尔逊相关系数)来衡量工人数据之间的相似性关联等级。Pearson 相关系数 $\rho$ 的相似性指数取值在 0.8～1.0 之间，我们认为工人所提供的数据之间为极强相关关系，本文将其分为 5 个等级，0.6～0.8 之间为强相关关系，0.4～0.6 之间为中等程度相关关系，0.2～0.4 之间为弱相关关系，0.0～

0.2 之间为极弱相关或无相关关系。

用 Pearson 相关系数作为权重，考虑单文本内词语间的不相关性和跨文本间词语的语义相关性，并比较它们之间的相对关系，得到文本的相似度。由此，对式(13)进行改进得到新的相似度计算式，如式(16)所示：

$$newsim(w_i,w_j)=\frac{\sum_{x\in w_i}\sum_{y\in w_j}\rho(x,y)}{\sum_{x_1\in w_i}\sum_{x_2\in w_i}\rho(x_1,x_2)*\sum_{y_1\in w_j}\sum_{y_2\in w_j}\rho(y_1,y_2)} \tag{16}$$

其中，$newsim(w_i,w_j)$表示文本 $w_i$ 和 $w_j$ 的相似度，其值越大表明两组数据之间的相似度越大，两者对比内容也越相似。因此，当根据式(16)计算的相似度大于阈值 $\theta$ 时，则表示两者是相似的。垃圾邮件检测算法如算法 2 所示。

**算法 2** 垃圾邮件检测算法

输入：众包工人提交的数据 R 集合

输出：垃圾邮件数据集合 N

```
1. for each review R in dataset do
2.     删除停用词
3.     提取词性
4. end for
5. for each business in R do
6.   for each reviews pair (Ri,Rj)∈R do
7.   Similarity←Newsim(Ri,Rj)
8.   end for
9.     for spam threshold θ=0.8,θ+=0.05 do
10.    if sim(Ri,Rj) > θ then
11.      Mark Ri and Rj as spam
12.    else
13.      为正常众包工人
14.    end if
15.  end for
16. end for
17. return N
```

输入众包工人提交的数据集合，通过相似性检测算法判定数据之间的相似性，当相似性阈值 $\theta$ 大于 0.8 时，我们判定其为垃圾邮件，否则不是垃圾邮件。最后输出为垃圾邮件数据集合 $N$。

## 6 实验与分析

本文算法采用准确率(accuracy)、综合指标 F1 值(F1-score)来评测实验结果，其定义分别如下：

$$accuracy=\frac{TP+TN}{TP+TN+FP+FN}$$

$$F1=2\frac{precision*recall}{precision+recall}$$

$TP$：真正(正确预测非垃圾邮件)；$TN$：真负(正确预测垃圾邮件)；$FP$：假阳性(错误预测非垃圾邮件)；$FN$：假阴性(错误预测垃圾邮件)。

### 6.1 实验环境及数据集

本文实验使用 2 个数据集(Ott 数据集[21]包含 800 条评论，有真实性评论和垃圾评论，并可公开获取；Yelp 数据集[22]有 57000 条评论)，它们被广泛用于相似性垃圾邮件检测。实验环境及配置如表 5 所列。

表 5 实验环境

Table 5 Experimental environment

| 实验环境 | 环境配置 |
|---|---|
| 操作系统 | Microsoft Windows 10 |
| 内存(RAM) | 16GB |
| 编程语言 | Python 3.8 |
| 处理器 | Intel Core i7-1165G7(1.2GHz/L3 12M) |

### 6.2 实验结果

按照以上的实验设置进行文本相似度检测，将不同方法的各项指标以折线图展示，算法执行时间用柱状图表示。图 8 给出了在不同数据集上 3 种算法的执行时间对比，图 9 和图 10 分别给出了 Ott 数据集上 Cosine 和 Cosine-POS-lemmatized 以及本文方法的准确率和 F1 值的比较结果，图 11 和图 12 给出了 Yelp 数据集上的对比结果。

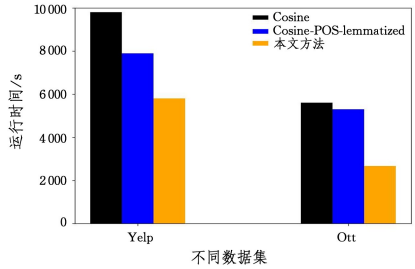



图 8 不同数据集执行时间比较

Fig. 8 Comparison of execution time on different datasets

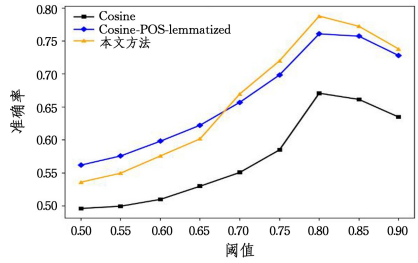



图 9 Ott 数据集上 3 种方法的准确率对比

Fig. 9 Accuracy comparison of three methods on Ott dataset

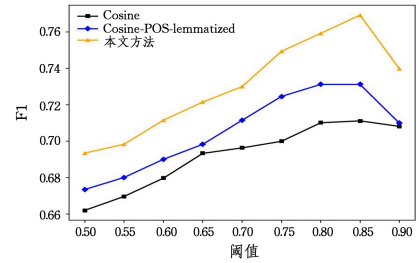



图 10 Ott 数据集上 3 种方法的 F1 值对比

Fig. 10 F1 value comparison of three methods on Ott dataset

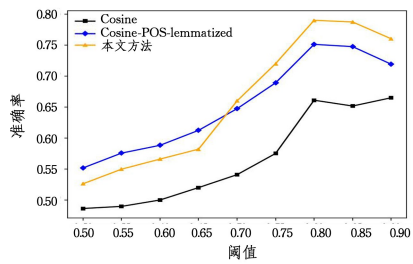



图 11 Yelp 数据集上 3 种方法的准确率对比

Fig. 11 Accuracy comparison of three methods on Yelp dataset

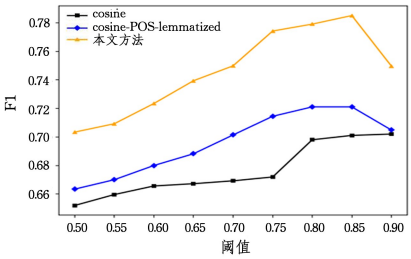



图 12 Yelp 数据集上 3 种方法的 F1 值对比

Fig. 12 F1 value comparison of three methods on Yelp dataset

总体来说，本文方法优于其他两种方法。在 Yelp 数据集上，Cosine 算法的执行时间明显比其他两种算法长，但是在 Ott 数据集上，两种对比算法的执行时间差别并不大，这是因为 Ott 数据集相较于 Yelp 数据集评论数量较少。本文方法在两个数据集上的执行时间都明显缩短了。

由图 9 可知，本文方法的准确率高于其他两种方法，阈值设置为 0.8 时准确率最高，为 78.7%，与 Cosine 和 Cosine-POS-lemmatized 方法相比，分别提高了 11.7%和 2.7%。但是当相似性指数为中等程度相关时，本文方法的准确率不如 Cosine-POS-lemmatized 算法。在垃圾邮件审查过程中，中等程度相关即判定为垃圾邮件会导致更多的数据误判，因此在强相关指数下判定垃圾邮件符合本文的众包质量控制机制的设计。

由图 10 可知，本文方法在阈值为 0.85 时 F1 值达到最大，为 76.9%，与 Cosine 和 Cosine-POS-lemmatized 方法相比分别提高了 3.8%和 5.8%。

在 Yelp 数据集上，评论的数量增加，随着相似度阈值的增大，性能平稳提升。这也表明本文方法更有意义。由图 11 和图 12 的结果可知，与 Ott 数据集结果相比，在同样的阈值下，Yelp 数据集上的 F1 值和准确率更高。相比之下，余弦相似性检测方法受到的影响更大。

如图 11 所示，阈值设置为 0.8 时，本文方法的准确率最高，为 78.9%，与 Cosine 和 Cosine-POS-lemmatized 方法相比分别提高了 12.9%和 3.9%。由图 12 可知，本文方法在阈值为 0.85 时 F1 值达到最大，为 78.5%，与 Cosine 和 Cosine-POS-lemmatized 方法相比分别提高了 6.4%和 8.4%。

实验结果表明，把相似性阈值设定在 0.8～0.85（即相关度等级为极强相关）时本文方法的准确率和 F1 值均达到最优，能够很好地区分垃圾邮件或非垃圾邮。总体而言，本文方法在准确率（accuracy）、综合指标 F1 值（F1-score）方面表现良好。

**结束语** 本文提出了一个基于陪审团的质量控制框架，在初步判断出众包数据中是否存在垃圾邮件员工和共谋团伙的情况下，进一步用本文提出的社区影响力检测算法（CIDA）检测出共谋组织。经过验证，CIDA 算法能准确地检测出共谋中心即影响力最大者以及其所在的共谋团伙。针对只提交随机或重复的答案而没有串通的垃圾邮件员工，本文使用改进的相似性检测算法（PI-cosine）来检测垃圾邮件以提高众包数据质量。通过这两个方面对众包数据质量进行双向质量控制，对众包数据进行检测筛查，大大提高了众包数据的可用性。实验结果表明，本文提出的 PI-cosine 算法优于其他两种检测方法。未来的工作将扩大研究方向并将其应用到面向更大的社交网络的众包共谋检测中。

## 参 考 文 献

[1] HOWE J. The rise of crowdsourcing[J]. Wired Magazine, 2006, 14(6): 1-4.

[2] KOROVINA O, BAEZ M, CASATI F. Reliability of crowdsourcing as a method for collecting emotions labels on pictures [J]. BMC Research Notes, 2019, 12(1): 715-715.

[3] ZHANG C, ZHU L, XU C, et al. A privacy-preserving traffic monitoring scheme via vehicular crowdsourcing[J]. Sensors (Basel), 2019, 19(6): 1274.

[4] AHMED M, KARAGIORGOU S, PFOSER D, et al. A comparison and evaluation of map construction algorithms using vehicle tracking data[J]. GeoInformatica, 2015, 19(3): 601-632.

[5] CIRQUEIRA D, VINÍCIUS L, PINHEIRO M, et al. Opinion Label: A Gamified Crowdsourcing System for Sentiment[C]// Anais Estendidos do XXIII Simpósio Brasileiro de Sistemas Multimídia e Web. SBC, 2017: 209-213.

[6] HAGERER G, SZABO D, KOCH A, et al. End-to-End Annotator Bias Approximation on Crowdsourced Single-Label Sentiment Analysis[C]// Proceedings of The Fourth International Conference on Natural Language and Speech Processing (ICNLSP 2021). 2021: 1-10.

[7] ENNAJI F Z, FAZZIKI A E, ABDALLAOUI H, et al. A Crowdsourcing Based Framework for Sentiment Analysis: A Product Reputation[J]. Journal of Communications Software and Systems, 2020, 16(4): 285-295.

[8] ZHOU J, JIN X, YU L, et al. TruthTrust: Truth Inference-Based Trust Management Mechanism on a Crowdsourcing Platform[J]. Sensors (Basel, Switzerland), 2021, 21(8): 2578.

[9] KHUDABUKHSH A, CARBONELL J, JANSEN P. Detecting Non-Adversarial Collusion in Crowdsourcing[C]// Proceedings of the AAAI Conference on Human Computation and Crowdsourcing. 2014, 2: 104-111.

[10] CHEN P P, SUN H L, FANG Y L, et al. Collusion-Proof Result Inference in Crowdsourcing[J]. Journal of Computer Science & Technology, 2018, 33(2): 351-365.

[11] NIAZI T M, AMINTOOSI H. Collusion-resistant worker selection in social crowdsensing systems[J]. Computer and Knowledge Engineering, 2018, 1(1): 9-20.

[12] AKKERHUIS T S, DE MAST J. Quantifying the random component of measurement error of nominal measurements without a gold standard[J]. Quality and Reliability Engineering International, 2016, 32(6): 1993-2003.

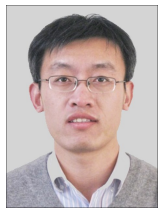
[13] JEONG S, LEE K. Spam Classification Based on Signed Network Analysis[J]. Applied Sciences, 2020, 10(24): 8952.

[14] MADHAVAN V M, PANDE S, UMEKAR P, et al. Comparative Analysis of Detection of Email Spam With the Aid of Machine Learning Approaches[J]. IOP Conference Series: Materials Science and Engineering, 2021, 1022(1): 012113.

[15] XU C, SHEN X, ZHU L, et al. A Collusion-Resistant and Priva-

cy-Preserving Data Aggregation Protocol in Crowdsensing System[J]. Mobile Information Systems, 2017, 2017: 1-11.
[16] LI M, WENG J, YANG A, et al. CrowdBC: A blockchain-based decentralized framework for crowdsourcing[J]. IEEE Transactions on Parallel and Distributed Systems, 2018, 30(6): 1251-1266.
[17] WANG Z, HU R, CHEN Q, et al. ColluEagle: collusive review spammer detection using Markov random fields[J]. Data Mining and Knowledge Discovery, 2020, 34(6): 1621-1641.
[18] KUANG L, ZHANG H, SHI R, et al. A spam worker detection approach based on heterogeneous network embedding in crowdsourcing platforms[J]. Computer Networks, 2020, 183: 107587.
[19] LUO J, SHAN H, ZHANG G, et al. Exploiting Syntactic and Semantic Information for Textual Similarity Estimation[J]. Mathematical Problems in Engineering, 2021, 2021: 4186750. 1-4186750. 12.
[20] LAURIOLA I, LAVELLI A, AIOLLI F. An introduction to deep learning in natural language processing: models, techniques, and tools[J]. Neurocomputing, 2022, 470: 443-456.
[21] OTT M, CARDIE C, HANCOCK J T. Negative deceptive opinion spam[C]// Proceedings of the 2013 Conference of the North Smerican Chapter of the Association for Computational Linguistics: Human Language Technologies. 2013: 497-501.
[22] MUKHERJEE A, VENKATARAMAN V, LIU B, et al. Fake review detection: Classification and analysis of real and pseudo reviews: UIC-CS-03-2013[R]. 2013.

**LIU Qingju**, born in 1997, postgraduate, is a member of China Computer Federation. Her main research interest is mobile crowdsourcing.



**PAN Qingxian**, born in 1979, Ph.D candidate, associate professor, is a member of China Computer Federation. His main research interests include artificial intelligence and machine learning.

（责任编辑：杨雪敏）