郭小强. 迭代检测策略在众包质量评估中的应用研究[J]. 智能计算机与应用,2025,15(6):202-206. DOI:10.20169/j.issn.2095-2163.250631

迭代检测策略在众包质量评估中的应用研究

郭小强
(河南省工业互联网创新发展中心, 郑州 450001)

摘 要: 为了提高众包结果质量评估的准确性,提出一种众包迭代检测策略。该策略根据众包工作者完成的任务,采用少数服从多数原则评估任务结果,将评估任务中选项不唯一的结果集作为新的任务再次发布到众包平台,选择候选的工作者人群参与迭代检测操作,直到确定出每一个任务的最优结果。通过众包的迭代检测策略可以有效地识别出评估任务中存在的相近结果,提高众包质量评估的准确性。实验表明,与基于熵的经典质量评估算法相比,该策略能够取得较好的效果。

关键词: 众包; 质量控制; 熵; 迭代检测策略; 质量评估
中图分类号: TP399 **文献标志码**: A **文章编号**: 2095-2163(2025)06-0202-06

Application research of iterative detection strategy in the crowdsourcing quality evaluation

GUO Xiaoqiang
(Center of Industrial Internet Innovation and Development, Henan Province, Zhengzhou 450001, China)

Abstract: To improve the accuracy of crowdsourcing results evaluation, the paper presents a crowdsourcing iterative detection strategy. According to the task crowdsourcing workers complete, the paper could use the principle of the minority being subordinate to the majority to assess the results of the task. The result sets of the assessment task in which candidate option is not unique will be regarded as new task to release in the crowdsourcing platform. Candidated workers would be chose to participate in the iterative detection operations until the optimal result of each task has been determined. Experimental results show that compared with the classic quality assessment algorithm based on entropy, this mechanism could achieve better results.

Key words: crowdsourcing; quality control; entropy; iterative detection strategy; quality evaluation

0 引 言

众包是 Howe[1] 在 2006 年首次提出的概念,用来描述一种利用互联网分配工作、发现创意或解决技术问题新的商业模式。目前国内外已出现众包平台,如 CrowdFlower 和 Amazon 的 MTurk[2]。通过众包平台,企业和组织可以利用自由工作者的创意和能力来解决问题,这些自由工作者具备完成任务的技能,愿意利用业余时间工作获取一定报酬[3]。典型的众包模式如图 1 所示。随着众包技术的不断发展,众包在很多领域得到了广泛应用,例如,在信息检索领域的图片搜索[4]、数据挖掘[5]、微博信息的可信度计算[6]、在数据库研究领域的 CrowdDB 查询系统[7],而在传感器方面 Demirbas 等学者[8]提出一种基于众包的传感系统,都可以将任务以众包的方式分发给在线的网络用户。

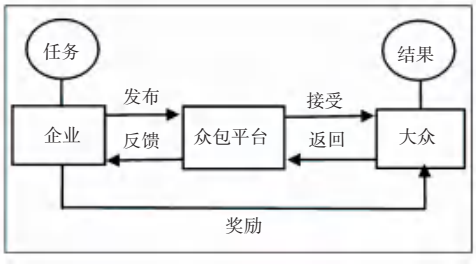

图 1 典型的众包模式

Fig. 1 A typical model of crowdsourcing

基金项目: 河南省国际科技合作项目(144300510007)。

作者简介: 郭小强(1987—),男,硕士,工程师,主要研究方向:云计算,智能信息处理。Email:543102893@qq.com。

收稿日期: 2024-01-03

由于众包任务的发布是面向互联网上所有用户，接受任务的工作者身份匿名，每一个工作者能力大小、工作态度各不相同，导致众包任务的结果具有较大不确定性[9]。本文针对以上问题提出一种众包迭代检测策略，采用少数服从多数的原则评估任务结果，有效地识别出评估任务中存在的相近结果，将其作为新的任务再次发布到众包平台，选择候选的工作者参与迭代检测，以提高众包质量评估的准确性。论文内容组织如下，第 1 节介绍众包质量控制的相关研究工作，第 2 节提出众包迭代检测策略架构，第 3 节进行实验和结果分析，第 4 节做出总结与展望。

1　相关研究

随着众包技术的广泛应用，众包的质量控制越来越受到关注，正如 Lease[10] 所指出那样，如果关心数据的质量，就必须考虑质量控制的问题。目前，关于众包质量控制方面的研究工作主要集中在 3 个方面[11]：

(1)结果质量评估方法研究。通过各种方法对工作者提交的结果进行评估，来识别恶意的工作者。

(2)工作者的组织模型。从建立一种好的工作者组织管理模式的角度来控制众包结果的质量[12]。

(3)众包任务的设计。从设计一个好的众包任务角度达到获得高质量结果的目标[13]。

本文主要针对众包结果质量评估方法进行研究，下面介绍常用的质量控制方法。

1.1　黄金标准数据评估策略

黄金标准数据是目前常用的质量评估手段之一，通过将工作者提交的任务结果和标准答案进行比较检测出工作者完成任务的优劣，识别出欺骗类型工作者并拒绝该类工作者提交的答案。对于一些简单的小任务，黄金标准数据是一种理想的选择，然而众包平台发布的任务绝大多数是没有固定答案的，很多问题带有主观性，需要对工作者进行客观公正的评估，对于这一类问题很难运用黄金标准数据评估方法。

1.2　阶段式动态众包质量控制策略

与黄金标准数据评估策略不同，参考文献[14]提出了阶段式动态众包质量控制策略，该方法实施阶段式动态质量控制。在众包任务的完成过程中设计若干个分段的检测点，依次来评估每个检测点上一个阶段完成的任务质量。如果发现提交的结果质量低下，则停止该工作者参与任务的资格，并删除所提交的结果，同时选择新的工作者继续本阶段的任务。采用这种策略可以更早地发现欺骗类型工作者，减少最终结果中不可靠结果比例，提高整体结果的质量。

虽然阶段式动态众包质量控制策略能够提高众包任务结果的质量，但是由于设置了检测点，每个阶段任务完成后需要多花费一些额外的时间来进行结果的检测和替换，这样会延长工作任务完成的时间，同时如何合理地设置检测点和替换策略也是必须着重考虑的问题，如果替换策略设计得不合理可能会陷入一直替换的恶性循环中，严重影响众包任务的完成。

1.3　基于熵的众包质量评估算法

最大期望估计(Expectation Maximization Algorithm)算法[15]是一种迭代算法，主要用来计算含有未知参数的极大似然估计[16-18]，是经典的众包质量评估算法之一。Ipeirotis 等学者在使用该算法时采用矩阵形式表示输出结果[19]。由于不能直观展示工作者完成任务情况，文献[20]提出一种基于熵的众包质量评估算法。该算法引入熵定义，每一位工作者完成任务后，通过计算所有参与任务工作者的结果，可以得出每一位工作者的得分，好的工作者得分接近 1，而差的工作者得分几乎为 0。通过工作者的得分，能够直观展示完成任务的情况。基于熵的评估算法中对未知参数的计算使用了 EM 算法，而 EM 算法采用少数服从多数的原则仅选择最多的选项作为最佳结果，忽略相近的结果，导致评估结果的准确性受到影响。

2　基于迭代的众包质量评估架构

本文针对目前众包质量控制研究工作中存在的问题与不足，提出一种基于众包质量评估的架构。通过分析工作者提交的结果，使用迭代控制的策略(Iterative Control Strategy)识别任务中存在的相近结果集，计算工作者提交结果的正确率，保留好的众包任务结果。

2.1　众包质量评估架构

图 2 描述了众包质量评估架构，主要包括众包任务发布、工作者人群分类和迭代检测策略 3 个部分。众包任务发布负责将众包任务池中的任务发布到众包平台；工作者人群分类使用分类算法将众包人群中优秀的工作者加入到候选人群；迭代检测策略对评估结果集中存在的相近结果进行迭代操作，确定出众包任务中的最佳结果，计算得到每个工作

者完成任务的正确率。

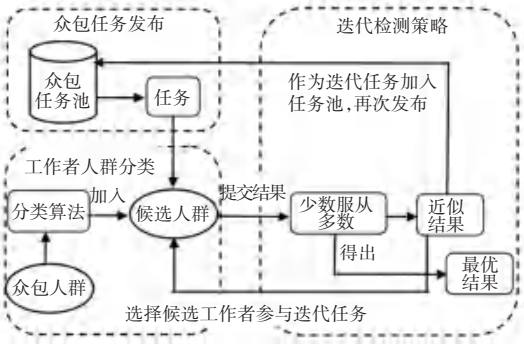

图 2 众包质量评估架构

Fig. 2 Crowdsourcing quality evaluation architecture

2.2 工作者分类算法

参与众包任务的工作者通常分为如下 3 类:

(1)勤奋类型的工作者。能够听从指挥,严格按照任务给出的要求,很好地完成任务,这类工作者通常能够给出比较满意的结果。

(2)草率的工作者。可能也具有良好的意图,由于没有认真阅读题目,给出低质量的结果。

(3)恶意的工作者。经常采用欺骗的手段随机地给出问题结果或者提交的问题结果全都一样。

迭代检测策略在每一次迭代过程中,筛选存在相似结果的众包任务,再次发布,并交由从众包人群中选出的候选工作者参与完成。文献[10]提出的工作者分类方法,检测出恶意和草率的工作者,保留勤奋型工作者。

对于检测草率类型的工作者,采用相关性特征距离的方法,通过每个工作者完成任务的结果和其他工作者完成的结果进行对比,计算出每一个工作者完成任务的随机分数。随机分数采用如下公式计算得到:

$$RandomSpam = \frac{\sum_{j\in J_w}\sum_{i\in J_{j,\bar{w}}} dis_{ij}}{\sum_{j\in J_w} |J_{j,\bar{w}}|} \tag{1}$$

其中,w 表示工作者;J_w 表示属于工作者 w 所做的相关性判断集合;$J_{j,\bar{w}}$ 表示除工作者 w 外其他工作者所做的相关性判断集合;dis_{ij} 表示对于同一问题 j,工作者 w 和其他工作者 i 两者之间所做判断的差异距离。如果 $dis_{ij}=0$,两者所做的判断相同;反之,如果 $dis_{ij}=1$,两者所做的判断不同。

同样,对于检查恶意类型的工作者,利用工作者完成的结果和其他工作者完成结果的不一致数,通过如下公式计算出每个工作者的分值:

$$UniformSpam_w = \frac{\sum_{s\in S}|s|\cdot(f_{s,J_w}-1)\{\sum_{j\in J_{s,w}}\sum_{i\in J_{j,\bar{w}}}(disagree_{ij})^2\}}{\sum_{j\in J_w}|J_{j,\bar{w}}|} \tag{2}$$

其中,s 表示所有任务的集合;$disagree_{ij}$ 表示对于同一问题 j,工作者 w 所做的相关性判断 $J_{s,w}$ 与其他工作者 $\bar{w}$ 提交的不一致数;f_{s,J_w} 表示工作者 w 标记的任务 s 在其所做判断集合 J_w 中出现的频数。

根据实验验证,式(1)选取评分大于 0.7,式(2)选取评分大于 1.6,能够有效地发现众包人群中存在的草率类型和恶意类型的工作者。

2.3 少数服从多数原则

在众包质量评估架构中,任务评判的标准采用的是少数服从多数的原则。该原则定义如下:假设有 1 项众包任务,m 个工作者参与,任务候选集 H 中有 s 个选项($h_1 \sim h_s$),基准值为 $k=\lfloor m/s \rfloor$,选项之间相近的阈值为 δ,每个选项的选择人数分别为 $c_1, c_2, \cdots, c_t$,其中 $c_1+c_2+\cdots+c_t=m$。如果对于任意选项 h_i 和 h_j,满足 $c_i > k, c_j > k (1 \leqslant i, j \leqslant t)$,且 $|c_i - c_j| < \delta$,则认为选项 h_i 和 h_j 是近似的结果选项;否则,若 $|c_i - c_j| > \delta$ 且 $c_i > c_j$,则认为选项 h_i 是最优结果选项。现在有 6 个工作者参与众包任务,该任务候选集中有 2 个选项,基准值为 3,工作者对任务的选择结果如图 3 所示。

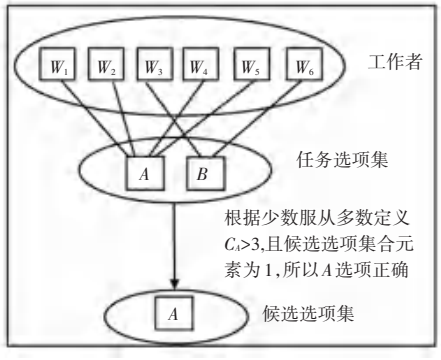

图 3 少数服从多数原则的模型图

Fig. 3 Model diagram of the minority being subordinate to the majority principle

2.4 迭代检测策略

本文提出一种迭代检测策略,其本质是对众包的任务结果进行预处理操作,将正确率低的结果过滤掉。该迭代策略的基本思想:首先,根据初始众包结果集合 $\boldsymbol{R}_{m\times n}$,采用 2.3 小节给出的少数服从多数的原则,消除最不可能正确的结果;然后,将每个任务中可能存在相近的结果作为一次新的任务重新发

布在众包平台上，从候选人群中选择工作者参与任务，再次采用少数服从多数的原则进行结果集的筛选和近似计算；通过多次迭代操作逐步达到对上一次完成任务结果的精确度，最终确定每一个任务的最佳结果；在此基础上将得到的评估结果集 $G_n = \{g_1, g_2, \cdots, g_n\}$ 和每个工作者提交的初始结果集 $\boldsymbol{R}_{m\times n} = \begin{vmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ r_{m1} & r_{m2} & \cdots & r_{mn} \end{vmatrix}$ 进行比较，计算出每个工作者(w_i) 的正确率 $rate[i]$，得到的公式为：

$$rate[i] = \frac{\sum_{j=1}^{n} R_{ij}}{\sum_{j=1}^{n} G_j} \tag{3}$$

下面给出迭代检测的算法：

输入　众包任务集合 $T = \{t_1, t_2, \cdots, t_n\}$，任务 $t_i(1 \leqslant i \leqslant n)$ 选项集合 $H_i = \{h_1, h_2, \cdots, h_s\}$，评估结果集 $G = null$

输出　评估结果集 $G = \{g_1, g_2, \cdots, g_n\}$，工作者($w_i$) 的正确率 $rate[i]$

begin

初始化：根据 2.2 节中的工作者分类算法，从众包人群中选择优秀工作者加入候选人群，选择 m 个工作者(w_i) 参与众包任务

While ($|G| < n$)

{

*S*1：发布众包任务集 & 合 T，任务 t_i 的选项集合到众包平台，m 个工作者(w_i) 参与评估；

*S*2：计算得到众包结果集合 R，根据 2.3 节中的少数服从多数原则，对众包结果集合 R 计算，将任务 t_i 中最优选项结果加入到 G 中；

*S*3：更新众包任务集合 T、任务 $t_i(1 \leqslant i \leqslant n)$ 选项集合 H_i 和评估结果集 G；

}

根据迭代结束后的结果集 G，计算每一个工作者的正确率(见式(3))，返回每个工作者(w_i) 正确率 $rate[i]$

end

3　实验及结果分析

为了验证众包迭代检测策略的有效性，搭建一个众包实验平台。实验方案如下：结合具体的教学环境，在众包平台上发布一组任务，由 20 位学生来参与任务的完成。在实验的过程中对于同一组任务，采用 2 种不同的质量评估方法。首先，将迭代检测策略集成到众包实验平台，执行众包任务，在实验中针对存在相近的结果，考虑设置合理的阈值，具体的参数值根据实际参与众包的人数来设置，计算出众包的评估结果；然后，使用基于熵的经典质量评估算法对众包结果进行评估；最后，以基于熵的质量评估结果作为基准，将 2 种算法评估的结果进行对比说明。

3.1　实验环境

(1)硬件：曙光 1420r-G 服务器，32 GB 内存，主频 3.07 GHz。

(2)软件：MyEclipse9.0。

3.2　实验结果分析

在众包平台上发布一组众包任务集合，包括 50 个任务，任务所涉及的内容都是计算机专业相关的知识。每个任务有 5 个可选项，分别用‘A’、‘B’、‘C’、‘D’和‘E’表示，其中这些任务均具有主观性。从众包候选人群选择 20 位学生参与众包任务的完成。表 1 显示了 20 位学生参与第一次众包任务完成后部分结果的统计。

表 1　工作者完成众包任务部分结果的统计

Table 1　Statistics on the results of a worker´s completion of a crowdsourced task

任务	选项				
	A	B	C	D	E
1	3	7	2	2	6
2	3	3	2	7	5
3	7	3	2	4	4
4	2	2	5	5	6
5	5	3	9	1	2

(1)阈值设置。在实验中，如何设置合理的阈值是最为关键的，这是由参与众包任务的人数所决定的。以表 1 数据为例，采用少数服从多数的原则，将基准设置为 5，即只要一个问题的选项有 5 个以上的工作者选择时，认为该选项可能是正确的结果；阈值设置为 2，即如果可能的结果选项之差的绝对值在 2 的范围内，则认为这些选项是相近的结果。针对任务 1 可以看到 B 和 E 都有可能是正确结果，考虑到阈值为 2，B 和 E 相差的绝对值在设置的阈值范围内，因此任务 1 中存在相近结果 B 和 E，需要将问题 1 和可能的结果 B 和 E 作为新的众包任务重新在众包平台上发布，选择候选的工作者参与迭代完成。同理，对于任务 2 可以看到 D 和 E 都有可

能是正确的结果,2 个选项之差的绝对值在阈值的范围内,因此 D 和 E 是相近的结果;任务 3 中,可以确定 A 就是正确结果;任务 4 中,C、D 和 E 都有可能是正确结果,3 个选项的绝对值之差都在阈值的范围内,C、D 和 E 都是相近的结果;任务 5 中 A 和 C 都有可能是正确结果,由于 2 个选项的绝对值之差超出了阈值的范围,因此认为选项最多的 C 是正确答案。对于众包结果中存在不确定结果的问题,仅仅需要将问题和可能的结果集重新在众包平台上发布,从候选工作者人群中选择工作者参与众包任务的迭代完成。下面以 20 位同学参与任务为例,分别将阈值设置为 2、3 和 4,采用迭代检测的策略对结果进行评估。实验结果如图 4 所示。

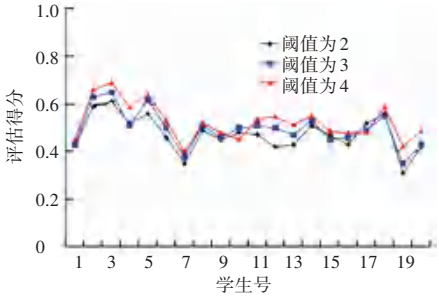

图 4 阈值设置结果对比

Fig. 4 The threshold setting contrast

从图 4 中可以得出:阈值设置的合理性直接影响到结果评估的准确率,如果阈值设置过大可能有多个相近的结果,虽然能够提高众包任务的准确率,但是会增加迭代检测的次数;反之,如果阈值设置过小可能使相近的结果给筛漏掉,导致众包任务的准确率低下。例如,表 1 中的任务 5,A 和 C 都是可能正确的结果。当阈值设置为 2 和 3 时,不存在相近的结果,经过一次的迭代检测就能确定 C 是正确的结果;当阈值设置为 4 时,A 和 C 是相近的结果,需要进行下一次的迭代操作才能确定,但实际上在第二次的迭代中正确的结果仍然是 C,由于阈值设置得过大,使相差甚远的 2 个选项作为相近结果,导致迭代的次数增多,实际上,这些迭代操作是徒劳的,增加额外的开销成本。经过实验的验证,本实验中阈值设置为 3 比较合理。

(2)算法比较。分别使用迭代检测的策略和基于熵的质量评估算法对这 20 位同学完成的结果进行评估,阈值为 3,实验结果如图 5 所示。

通过实验结果对比分析可以得出:与基于熵的质量评估算法相比,迭代检测策略充分考虑到采用少数服从多数的原则进行质量评估时,同一个任务中可能存在相近的结果,需要对相近的结果进行多次的迭代处理确定最佳的结果,从而提高评估结果的准确性。而基于熵的质量评估算法在计算每一个工作者得分时,对于涉及到的未知参数计算使用 EM 算法,该算法初始化参数时采用少数服从多数的原则,仅仅选择最多的选项作为最佳结果进行质量评估,最终的质量评估结果与初始化的参数值有很大的关系。例如,对于表 1 中的任务 1,最佳答案是 E,但由于 B 选项中有恶意的工作者参与使该选项的人数增多,EM 算法认为 B 是最佳结果,而 E 相对来说是一个相近结果被忽略。此时,若将正确答案认为是 B 将影响工作者的得分,导致评估结果的准确性降低。迭代检测策略充分考虑上述情况,对相似的结果进行处理提高众包结果质量评估的准确性。

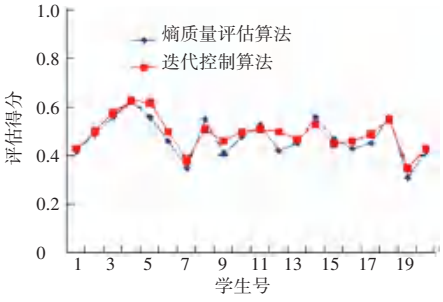

图 5 算法结果对比

Fig. 5 Results comparison of algorithms

4 结束语

针对目前众包质量控制的不足,本文提出一种迭代检测策略。根据众包工作者完成的任务,采用少数服从多数原则,将评估任务中存在的相近结果集作为新的任务再次发布到众包平台,选择候选的工作者人群参与,反复多次迭代检测,有效地区分出评估任务中存在的相近结果,提高众包任务结果质量评估的准确性。迭代检测策略中的关键点是设置合理的阈值。通过将迭代检测策略和基于熵的质量评估算法对同一组众包任务的结果进行评估对比,实验表明迭代检测策略的评估结果优于基于熵的质量评估结果,符合预期的结果。

本文提出的迭代检测策略只是考虑了众包结果质量控制方面。由于大部分的众包任务是需要支付给工作者报酬,论文下一步工作将考虑众包经济方面的因素,即如何在保证众包任务结果质量的前提下,使众包发布者付出最少的报酬。

参考文献

[1] HOWE J. The rise of crowdsourcing[J]. Wired, 2006, 14(6): 176-183.

[2] CALLISON-BURCH C, DREDZE M. Creating speech and language data with Amazon's mechanical Turk[C]//Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. ACL,2010: 1-12.
[3] 魏拴成. 众包的理念以及我国企业众包商业模式设计[J]. 技术经济与管理研究,2010(1):36-39.
[4] YAN Tingxin, KUMAR V, GANESAN D. CrowdSearch: Exploiting crowds for accurate real-time image search on mobile phones[C]//Proceedings of the International Conference on Mobile Systems, Applications, and Services. New York:ACM,2010: 77-90.
[5] LEASE M, CARVALHO V R, YILMAZ E. Crowdsourcing for search and data mining[J]. Journal of SIGIR Forum(SIGIR), 2011, 45(1):18-24.
[6] CASTILLO C, MENDOZA M, POBLETE B. Information credibility on twitter[C]//Proceedings of the WWW. New York: ACM, 2011:675-684.
[7] FRANKLIN M J, KOSSMANN D, KRASKA T, et al. CrowdDB: Answering queries with crowdsourcing[C]//Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2011: 61-72.
[8] DEMIRBAS M, BAYIR M A, AKCORA C G, et al. Crowdsourced sensing and collaboration using twitter[C]//Proceedings of 2010 IEEE International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM). Piscataway,NJ: IEEE,2010:1-9.
[9] BAIO A. Cheap, easy audio transcription with mechanical Turk [EB/OL]. (2008-09-22). http://waxy.org/2008/09/audio_transcription_with_mechanical_turk/.
[10] LEASE M. On quality control and machine learning in crowdsourcing[C]//Proceedings of Human Computation Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence. California,USA:AAAI, 2011:97-102.
[11] 张志强,逄居升,谢晓芹,等. 众包质量控制策略及评估算法研究[J]. 计算机学报,2013,36(8):1636-1649.
[12] KOCHHAR S, MAZZOCCHI S, PARITOSH P. The anatomy of a large-scale human computation engine[C]//Proceedings of the ACM SigKDD Workshop on Human Computation. New York: ACM, 2010: 10-17.
[13] KITTUR A, CHI E H, SUH B. Crowdsourcing user studies with Mechanical Turk[C]//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York:ACM, 2008: 453-456.
[14] IPEIROTIS P G, PROVOST F, WANG Jing. Quality management on Amazon mechanical turk [C]//Proceedings of the ACM SIGKDD Workshop on Human Computation. New York:ACM, 2010: 64-67.
[15] DAWID A P, SKENE A M. Maximum likelihood estimation of observer error-rates using the EM algorithm[J]. Applied Statistics,1979,28(1):20-28.
[16] 孙大飞,陈志国,刘文举. 基于EM算法的极大似然参数估计探讨[J]. 河南大学学报(自然科学版),2002,32(4):35-41.
[17] SUN Dafei, DEMPSTER A P, LAIRD N M, et al. Maximum likelihood from Incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society Series B,1997,39(1):1-38.
[18] MENG X L, RUBIN D B. Recent Extension to the EM algorithm [M]. Oxford:Oxford University Press,1992.
[19] RAYKAR V C, YU Shipeng. An entropic score to rank annotators for crowdsourced labeling tasks[C]//Proceedings of 2011 Third National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG). Piscataway, NJ: IEEE, 2011: 29-32.
[20] VUURENS J B P, VRIES D A P. Obtaining high-quality relevance judgments using crowdsourcing[J]. IEEE Internet Computing, 2012, 16(5): 20-27.