

移动应用众包测试人员信誉度的模糊评估方法研究

成静¹, 薛峰², 张逸飞³, 张涛², 马春燕²

(1.西安工业大学 计算机科学与工程学院, 陕西 西安 710021;
2.西北工业大学 软件与微电子学院, 陕西 西安 710072;
3.西北工业大学 计算机学院, 陕西 西安 710072)

摘要:移动应用众包测试因众包模式的匿名性和无监督性,使得移动应用测试人员容易对测试任务产生懈怠或有意欺骗行为,致使众包测试质量的下降。针对该问题,提出一种基于模糊数学的移动应用众包测试人员信誉度评估方法以衡量测试人员的信誉水平。此方法的特点在于紧密结合众包测试的特性,通过引入测试人员评价机制,构建信誉度评估模型并研究信誉度在众包测试过程中的计算与更新算法,以实现对众包测试人员信誉度的合理评估,有效选拔出高信誉测试人员,保证移动应用众包测试质量。

关键词:移动应用测试;众包测试;信誉度;模糊数学

中图分类号:TP311 **文献标志码:**A **文章编号:**1000-2758(2018)04-0800-07

移动应用众包测试,是将移动应用众包测试任务,以自由自愿的方式外包给匿名网络测试人员^[1],具有灵活方便、可伸缩、成本低、测试场景真实等显著优势^[2]。但由于其匿名性和非监督性,使得众包测试人员可能存在测试行为欺诈,以获取最大收益。为此,通常众包平台采用最大期望、最大似然估计数算法^[3-4]来评估众包人员信誉度,存在运算代价高、实时性差等问题;另有学者提出针对信誉评估的数据分类算法^[5],却只适合简单标注类型的众包任务。

本文在充分考虑移动应用众包测试基本特性的前提下,提出了一种面向复杂众包测试任务的测试人员信誉度的评估方法,通过划分可信与不可信2个模糊集合,以移动应用众包测试任务发包方与测试人员相互间的评分为计算基础,利用模糊集合理论判断众包测试人员对2个集合的隶属程度,计算评估众包测试人员信誉度。

1 基于模糊集合的移动应用众包测试人员信誉度评估

本节从人员评价机制、信誉度评估模型、信誉度计算及更新3个阶段,完成对移动应用众包测试人员信誉度评估体系的构建。

1.1 人员评价机制

评分、评级等形式的评价机制已广泛应用于电子商务类网站之中,可以很好地反映交易双方的满意度,也体现了双方的信任关系。在众包测试中,引入这些机制可以有效地表现出对测试服务完成效果的评价,可以作为评估信誉度的基础数据。

图1表示一个常用的“5分制”的评价打分形式,可以较好地描述对某一服务或商品的满意程度。

非常满意	满意	一般	不满意	差
5分	4分	3分	2分	1分

图1 5分制评分

然而,与普通一对一的互评模式不同,在移动应用众包测试中,通常测试任务的发包方与测试人员是一种一对多的评价关系。测试任务的发包方需要对多名测试人员进行评价,评估工作量巨大,评分效果差。

为解决这一问题,本文设计了一种偏隐式的评分方式,即通过记录任务发包方与测试人员的交互行为,将其作为评分依据,以取代双方的主动评价。如表 1 所示,考虑以发包方对测试人员所发现缺陷

表 1 任务发包方对测试人员的评价

评分项	5 分	4 分	3 分	2 分	1 分
缺陷价值评价	高	中高	中	中低	低
人员可信	多次使用	收藏	无评价	加黑名单	举报欺诈
.....
任务完成度	高	中高	中	中低	低

表 2 测试人员对任务发包方的评价

评分项	5 分	4 分	3 分	2 分	1 分
缺陷评价是否公正	给出好评	确认结果	默认结果	提出异议	反对结果
人员可信	多次参与	关注	无评价	加黑名单	举报欺诈
.....
任务执行度	高	中高	中	中低	低

在得到一系列评分后,需要对各项评分进行综合。这里采用几何平均数法对评分进行综合处理,公式为:

$$\bar{a} = \sqrt[n]{\prod a_i} \quad (1)$$

式中, \bar{a} 表示评价综合分, n 表示评分项数, a_i 表示第 i 个评分项的得分。例如,对一个 5 项评分的评价进行综合,则 $\bar{a} = \sqrt[5]{5 \times 4 \times 3 \times 5 \times 4} = 4.13$,即综合评分为 4.13。

采用综合评价机制的主要目的,是为了获取移动应用众包测试参与双方的相互反馈信息,进而将其作为基础数据开展移动应用众包测试信誉度评估方法的研究。

1.2 信誉度评估模型

通常,信誉度难以精确量化表示。在众包测试中,测试人员很难被界定为绝对可信或不可信,因此适合模糊数学理论评估信誉度。本文在获得移动应用众包测试任务发包方与测试人员的相关评价综合得分后,利用模糊数学理论来构建信誉度评估模型。首先,利用模糊集合思想,在以所有移动应用众

的价值评价为标准,间接反映测试人员在本次任务中体现的价值;也可以通过记录发包方是否偏向于使用某一位测试人员来体现对其认可的程度。同样,表 2 展示了测试人员对发包方的隐式评价,包括对发包方任务的关注、参与等行为,以及缺陷评价的公正性等。使用偏隐式评价的另一优势在于能较大程度地避免评价者主观性或恶意性评价,使得评价结果更为直接客观。

包测试参与人员为范围定义 2 个模糊子集,即“不可信”子集和“可信”子集,进而通过判断每位众包社区参与者分别隶属于 2 个集合的程度,见图 2。例如,当参与者的可信任程度大于其不可信程度时,即其更偏向信任子集,则认为这名参与者是一名可信任人员,否则为不可信。



图 2 众包人员可信及不可信程度关系

其次,利用从人员评价机制中获得的综合评价得分,仅筛选可信任人员的评分作为被评价人员的信誉计算基础。这样,在移动应用众包测试社区的整体运作过程中,每名众包测试参与人员信誉的计算,形成如图 3 所示的一种迭代循环过程。

在整个众包测试社区中,当想要获知哪些测试人员是可以信任时,通过向与测试人员有过任务接触的可信任任务发布人员进行确认作为其对测试人员

可信的判断依据。在众包测试社区整体的评价过程中,对一名参与者的评价是否可作为判断另一名参与人员信誉计算的依据,主要依赖于这名参与人员

本身是否被社区群体所认可,具有足够的可信度。这样,每一名参与人员在多次参与测试任务后,经过多轮的信誉迭代计算,其信誉的评估会越趋于准确。

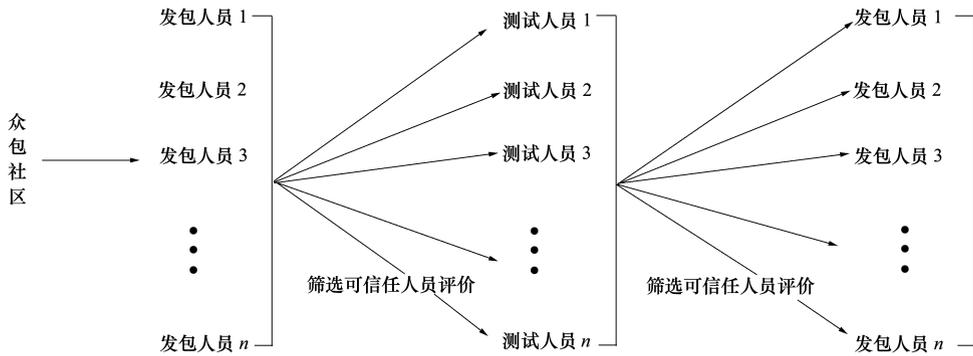


图 3 移动应用众包测试社区人员评价过程

1.3 信誉度的计算及更新

下面将详细说明移动应用众包测试人员信誉度评估模型的具体计算及更新过程。

1) 隶属度函数的确定

前面提到,通过划分出 2 个模糊子集“可信”和“不可信”以及使用可信人员提供的综合评分来估计参与者的可信情况。下面先分析一下综合评分的特点。

(1) 因采用 5 分制进行评分,并且根据评分的计算方式可以得出综合评分的结果必然是 1~5 范围内的实数;利用几何平均数计算出来的结果变化

趋势必然是一种呈直线状的线性趋势。

(2) 评分的高低在一定程度上反映出评价人对被评价人的信任关系,评分越低表明评价人对被评价人的不信任;反之,评分越高则体现为信任。

(3) 根据图 1 所示的 5 分制评分标准:评分为 3 分是最模糊的状态;评分为 4 分及以上的表明评价人对被评价人的肯定;而全部评分为 2 分及以下的表明评价人对被评价人的否定。

根据上述 3 个特点,隶属函数可以直接套用实数域上的常用模糊分布,本文选择梯形分布作为隶属函数。梯形模糊分布如图 4 所示。

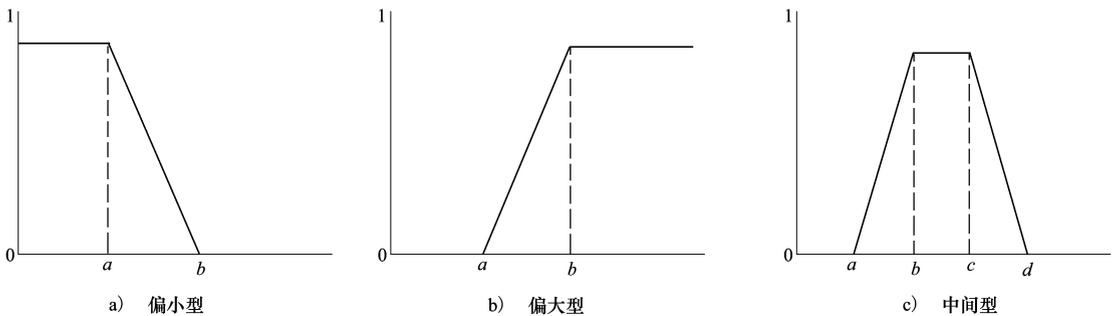


图 4 梯形模糊分布图

上述 3 种隶属函数的解析式如下:

$$A(x) = \begin{cases} 1 & x < a \\ \frac{b-x}{b-a} & a \leq x \leq b \\ 0 & b < x \end{cases}$$

$$A(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & b < x \end{cases}$$

$$A(x) = \begin{cases} \frac{x-a}{b-a} & a \leq x < b \\ 1 & b \leq x < c \\ \frac{d-x}{d-c} & c \leq x \leq d \\ 0 & x < a \text{ or } d < x \end{cases}$$

因为只将模糊集划分为 2 个模糊子集,所以仅使用到梯形分布的偏大型和偏小型 2 个部分。根据评分的数值变化趋势,令偏小型代表不可信子集,偏大型代表可信子集。而梯形分布参数的确定,则根据上述特点第 3 条,假设每个参与者都拥有 3 条被评价项,可令全部为 4 分为必然肯定值,2 分及以下情况为否定值,则参数 $a = \sqrt[3]{2 \times 2 \times 1} = 1.59, b = \sqrt[3]{4 \times 4 \times 4} = 4$ 。由此,可以得到 2 个子集的隶属度函数表示为

$$A(x) = \begin{cases} 1, & x \leq 1.59 \\ \frac{4-x}{4-1.59}, & 1.59 < x < 4 \\ 0, & x > 4 \end{cases} \quad (2)$$

$$B(x) = \begin{cases} 1, & x \geq 4 \\ \frac{x-1.59}{4-1.59}, & 1.59 < x < 4 \\ 0, & x < 1.59 \end{cases} \quad (3)$$

(2) 式中, $A(x)$ 表示不信任模糊子集的隶属度函数。(3) 式中, $B(x)$ 则表示信任模糊子集的隶属度函数。其分布性态形如图 5 所示。

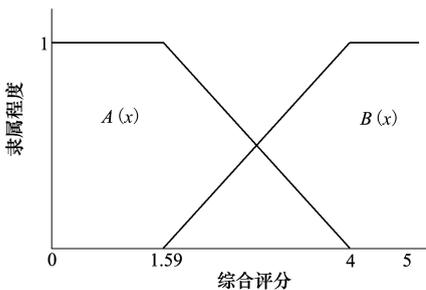


图 5 隶属度函数分布图形

2) 信誉度计算

移动应用众包测试参与人员在不断参与众包任务的过程中,对他的评价会进行积累,即每个参与者均会有一组评分,而这一组评分则形成了被评价人员的评分范围。如图 6 所示,其中 C_1 和 C_2 之间可能就是某位人员的被评分范围。

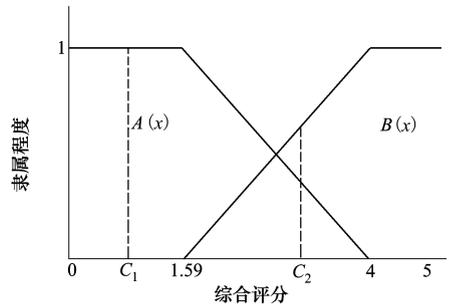


图 6 某被评价人员的评分范围

在评估某一位参与人员信誉度时,须更多关注该人员表现区间的变化,即其被评分范围的边界值,如图 6 中的 C_1 和 C_2 点。当该人员的评分在该区间内时,表明其信誉表现属于正常水平;而当该人员的表现小于 C_1 时,说明他有不良信誉行为;当大于 C_2 时,说明参与人员的信誉得到了更好肯定。

因此,对于每位参与人员的一组评分,当确定了评分范围后,依据隶属度函数分布图形,使用质心法来求出这一组评分的综合值,质心法的计算过程如下:

$$Z_A = \frac{\int_a^b A(x) x dx}{\int_a^b A(x) dx} \quad (4)$$

$$Z_B = \frac{\int_a^b B(x) x dx}{\int_a^b B(x) dx} \quad (5)$$

式中, Z_A 表示不可信范围内的质心, $A(x)$ 表示不可信模糊子集的隶属度函数; Z_B 表示可信范围内的质心, $B(x)$ 表示可信模糊子集的隶属度函数; a, b 表示评分范围的边界值。

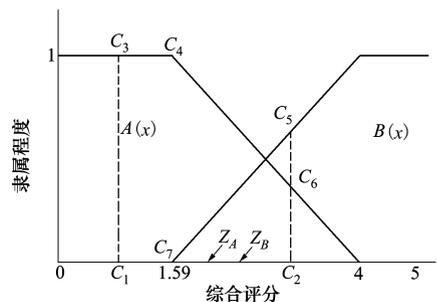


图 7 评分范围质心的计算

如图 7 所示,当要计算 C_1 到 C_2 范围内不可信和可信范围内的质心时。不可信范围内表示求 $C_3 -$

$C_4 - C_6$ 运动轨迹的质心,可信范围内表示求 $C_7 - C_5$ 运动轨迹的质心。

求得质心后,将其继续带入(2)式和(3)式即可求得被评价人员的一组评分数据,它们分别隶属于不可信模糊子集和可信模糊子集的程度。这样的计算方式从不可信及可信的双重角度来审视被评价人,使信誉估计值更具可信度。于是,可给出如下的综合信誉度值的计算公式:

$$C_R = \frac{R_e + (1 - \overline{R_e})}{2} \quad (6)$$

表 3 一组被评价人员的信誉度计算

被评价人员	最低评分 a	最高评分 b	不可信质心 Z_A	可信质心 Z_B	不可信程度 R_e	可信程度 $\overline{R_e}$	综合信誉度 C_R
A	$\sqrt[3]{5 \cdot 3 \cdot 3}$	$\sqrt[3]{5 \cdot 3 \cdot 4}$	3.70	4.11	0.132 7	1	0.933 7
B	$\sqrt[3]{2 \cdot 3 \cdot 2}$	$\sqrt[3]{4 \cdot 4 \cdot 4}$	2.86	3.32	0.504 4	0.699 1	0.597 3
C	$\sqrt[3]{3 \cdot 3 \cdot 4}$	$\sqrt[3]{4 \cdot 3 \cdot 4}$	3.45	3.53	0.243 4	0.792 0	0.774 3
D	$\sqrt[3]{2 \cdot 2 \cdot 1}$	$\sqrt[3]{3 \cdot 2 \cdot 2}$	1.92	2.10	0.920 4	0.159 2	0.119 4

3) 信誉度更新

在信誉值的更新方面,需要考虑如下 2 个因素:

①历史评分对信誉度计算的影响,考虑影响力的衰减问题,次数越近的评分越能反映出测试人员的当前信誉水平;②不同评价人应具备不同影响力,自身可信程度越高的评价人,对他人做出的评价,自然会有更高的影响。针对上述问题,在信誉度更新的问题上,引入评分影响力因子来控制每一个评分对众包测试参与人员信誉度的影响。

评分影响力因子的基本作用是确定评分参与信誉度评估计算的有效次数。一般而言,信誉度是一种较为固化的属性,随时间增长的变化缓慢,因此,是以使用次数而非时间作为历史评分的衰减单位。众包测试参与人员每次参加一项任务,在获得新评分的同时,历史评分的影响力会进行一次衰减,当某个评价值的影响力因子衰减为 0 时,则不再影响众包人员的评分。

此外,评分影响力因子还包含对不良表现的惩罚系数,评分低则会加强该评分存在次数。这样,将令众包测试参人员尽量避免获取较差的评分。

评分影响力因子由(7)式表示,主要由 3 个参数构成。其中, α 表示评价人员不可信程度和可信程度的贴近度,利用质心数据相减求得, α 越小则说明可信和不可信程度越贴近,即该评价人的表现较模糊,影响力低;而 α 越大则说明该评价人的表现越

式中, C_R 表示综合信誉度, R_e 表示不可信模糊子集隶属程度, $\overline{R_e}$ 表示可信模糊子集隶属程度。

作为例子,表 3 给出了一组被评价人员的信誉度计算数据,最低评分和最高评分即对应(4)式和(5)式中的参数 a 与 b ,套用公式分别求得 4 名被评价人员在不可信范围和可信范围的质心 Z_A 和 Z_B ;将求得的质心 Z_A 和 Z_B 再带入到隶属度函数中用(2)式和(3)式求得 4 名被评价人员对于不可信模糊子集的隶属度 R_e 及可信模糊子集的隶属度 $\overline{R_e}$;最后使用(6)式求得信誉度综合值 C_R 。

不模糊且越好,影响力增大; β 是惩罚系数,当评分小于 5 分制的中值 2.5 时开始生效,以 2 为底数可以令激活惩罚系数带来的数影响最小,达到 0.5 倍; γ 是影响力基数,表示一般情况下评分的影响力,例如设置为 5 次。

$$S_i = \lceil \alpha \left(1 + \frac{1}{2^\beta} \right) \gamma \rceil \quad (7)$$

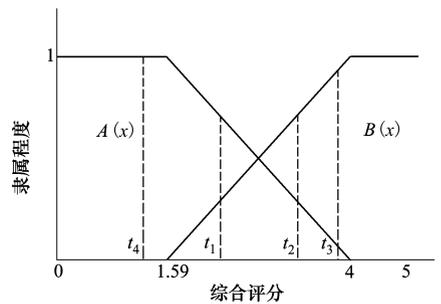


图 8 某人员参与 t_1 至 t_4 次任务的评分边界变化

图 8 举例说明了评分影响力因子的作用。例如,在 t_1 至 t_4 次,某人员参与任务后边界值的变化,在 t_1 至 t_3 次时,该人员的最高评分不断向上,使得其评分区间逐渐向可信任区间靠拢;但在 t_4 次时表现较差,评分突破了原来的最低边界 t_1 ,使得评分区间被拉大,增大了评分区间所包含的不可信区间范围,导致质心位置的变化,并影响信誉度值;同时, t_4 次评分将会持续比 t_1 至 t_3 次更大的次数,所以该名人

员需要在后续被更好认可,否则将会进一步导致信誉评分的降低。

由此可见,在加入评分影响力因子的信誉度更新机制后,高信誉需要不断积累而保持,而惩罚机制的存在,使不良评分对信誉度会造成更持久的不良影响。

2 移动应用众包测试人员信誉度的实验验证

因众包测试社区的可用数据目前较为稀少,所以本次采取模拟数据的方式进行实验验证。首先,通过分析,利用设置不同的缺陷发现概率范围和发现缺陷可能性概率范围,来模拟生成6种具备代表

类型的测试人员1000名。根据测试人员自身水平,将其划分为优秀型、专精型、稳定型、学习型、欠缺型、欺骗型等6个类型。其次,针对每一种测试任务的情况,设置缺陷数范围和缺陷被发现概率,来模拟生成测试任务情况。

从模拟出的1000名测试人员数据中,多次选择出分别对应优秀型、专精型、稳定型、学习型、欠缺型和欺诈型具有代表性的6名人员数据,并分析他们在任务执行过程中的信誉度变化情况。在实验中,每名参与人员的信誉度初始值均设定为0.5,即处于可信与不可信的中间模糊状态。图9分别展示了6名代表人员,分别进行5种不同难度任务10次的信誉度迭代计算,以及综合5种难度情况的20次任务迭代计算结果。

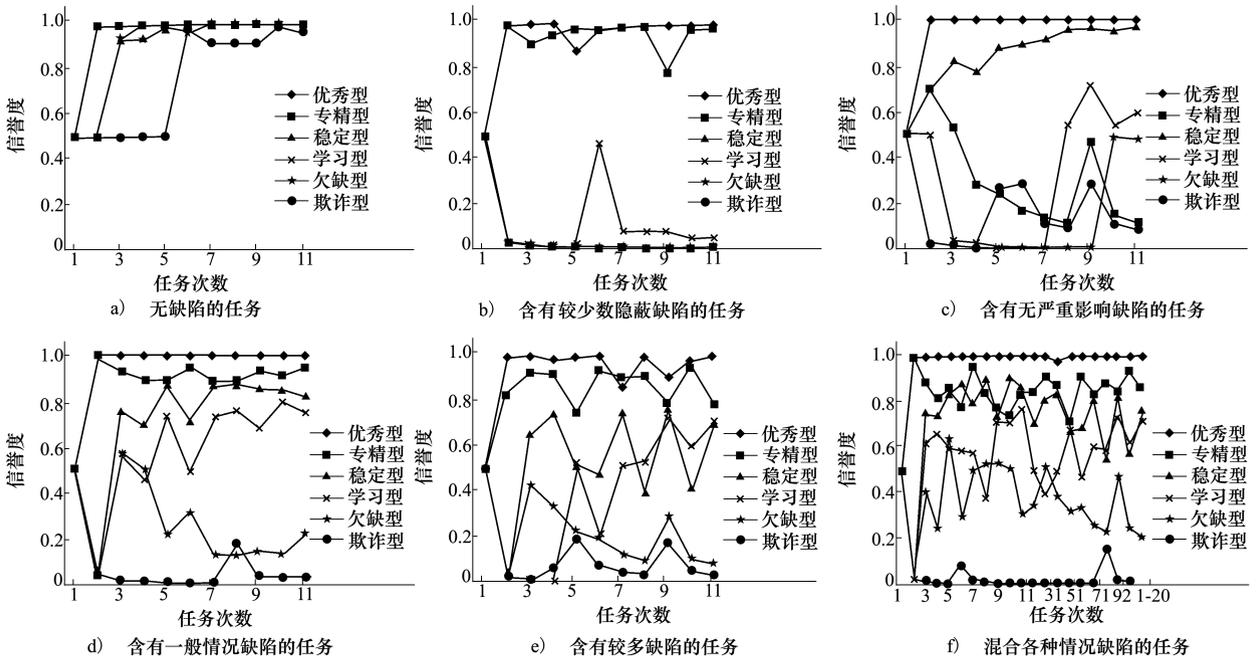


图9 不同类型测试人员在各类型任务下的信誉表现

对于无缺陷任务,如图9a)所示,这是一种特殊情况,信誉度只增加不减少,因此信誉度变化并不明显;图9b)是含有较少数隐蔽缺陷的任务,这类任务具备一定难度,因此优秀型和专精型人员的信誉度在高位持续;图9c)是含有无严重影响缺陷的任务,稳定型人员比较擅长此类任务,因此持续增长,而专精型人员则不适合此类任务,导致信誉度逐渐下降;图9d)是一般的普通任务情况,各类型人员表现为正常水平,信誉度趋势划分显著符合预期;图9e)含

有较多的缺陷任务,与缺陷正常水平任务相比,各类人员的表现均有更为明显的波动,主要因为在大量测试缺陷存在的情况下,测试人员因工作量大会漏掉某些缺陷的发现,而对于欺诈型人员因为缺陷的增多,也因他们使用猜测方式命中缺陷的概率有一定提升;图9f)是混合执行各类型任务的表现,优秀型、专精型、稳定型人员信誉度均呈现较为稳定的持续状态,学习型人员则有明显的上升趋势;欠缺型人员尽管个别有突出表现,但因表现的不持续性,信誉

度会持续下降;欺诈型与欠缺型人员较为类似,不持续性更为明显。

根据上述实验,当测试人员的信誉突破边界值时,信誉指标会有较大幅度的变化,但例如欠缺型和欺诈型人员,因为惩罚因素的存在,边界值下限存在的时间更长,因此信誉度会被迅速降低。随着迭代过程的不断进行,信誉度将不断稳定并趋于精确。上述实验初步证明,本文所提出的基于模糊集合的移动应用众包测试人员信誉度评估方法,在面向不同任务情况下,能够对不同类型测试人员的信誉特征进行有效评估。

3 结 论

本文提出了一种基于模糊集合的移动应用众包测试人员信誉度评估方法,该方法利用信誉度的模糊特性,通过可信人员间的迭代评价,以准确估计出众包测试人员的信誉水平。下一步,将继续围绕信誉评估进行扩展性研究,包括相近信誉度人员的行为一致性分析、基于信誉度的测试任务结果可信程度的智能化判断等,不断完善众包测试可信评估,以提高众包测试质量。

参考文献:

- [1] Leicht N, Blohm I, Leimeister J M. Leveraging the Power of the Crowd for Software Testing[J]. IEEE Trans on Software, 2017, 34(2):62-69
- [2] Zhang T, Gao J, Cheng J. Crowdsourced Testing Services for Mobile Apps[C]//Proceedings of the 2017 IEEE International Symposium on Service-Oriented System Engineering, 2017: 132-137
- [3] Peer E, Vosgerau J, Acquisti A. Reputation as a Sufficient Condition for Data Quality on Amazon Mechanical Turk[J]. Behavior Research Methods, 2014, 46(4): 1-9
- [4] Ye B, Wang Y, Liu L. Crowd Defense: A Trust Vector-Based Threat Defense Model in Crowdsourcing Environments[C]//Proceedings of the 2017 IEEE International Conference on Web Services, 2017:245-252
- [5] Lee S, Park S, Park S. A Quality Enhancement of Crowdsourcing Based on Quality Evaluation and User-Level Task Assignment Framework[J]. Applied Mathematics & Information Sciences, 2014, 9(2): 60-65

A Reputation Assessment Approach Based on Fuzzy Mathematics Mobile Application Crowdsourced Testers

Cheng Jing¹, Xue Feng², Zhang Yifei³, Zhang Tao², Ma Chunyan²

(1.School of Computer Science and Engineering, Xi'an Technological University, Xi'an 710021, China;
2.School of Software and Microelectronics, Northwestern Polytechnical University, Xi'an 710072, China;
3.School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: The anonymity and unsupervision of mobile application crowdsourced testing, makes the tester easy to slacken or cheat on test tasks, resulting in a drop in the quality of crowdsourced testing. To solve the problem, this paper proposes a method for assessing the reputation of a mobile application crowdsourced testers based on fuzzy mathematics. This method closely integrates crowdsourced testing features. By introducing a tester evaluation mechanism, building a reputation assessment model and studying the calculation and update of reputation in the crowdsourced testing, it achieves a reasonable assessment of the reputation of a crowdsourced tester and can effectively screen out high reputation testers, thus ensuring the quality of mobile application crowdsourced testing.

Keywords: mobile application testing; crowdsourced testing; reputation; fuzzy mathematics