计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(1):269−274 [doi: 10.15888/j.cnki.csa.006749]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于情感文本数据筛选的感知节点选择机制①

张晓滨, 黄梦莹

(西安工程大学 计算机科学学院, 西安 710048)
通讯作者: 黄梦莹, E-mail: 1084566658@qq.com

摘　要: 通过分析移动群智感知的协作过程, 即感知节点的携带-存储-转发过程, 发现该过程忽略了对节点携带信息的内容筛选. 而对于有目的的数据获取而言, 这种先收集后筛选的方法导致在后续对数据的分析与筛选过程中会耗费更多的时间, 同时获取的有效数据占比不高. 考虑到这个因素, 本文结合遗传算法设计了一种在移动群智感知环境下基于情感文本数据筛选的节点选择机制. 该节点选择机制主要通过对节点携带数据类型的筛选来选择感知节点, 从而获取感知环境下移动用户的情感文本数据. 通过实验验证表明, 使用此方法在数据处理的效率上最大提高了 27.6%, 在有效的数据占比上最大提高了 21%, 因此该方法能够有效的提高对整体数据处理的效率.

关键词: 移动群智感知; 遗传算法; 情感文本; 数据筛选; 节点选择机制

引用格式: 张晓滨,黄梦莹.基于情感文本数据筛选的感知节点选择机制.计算机系统应用,2019,28(1):269–274. http://www.c-s-a.org.cn/1003-3254/6749.html

Sensing Node Selection Mechanism Based on Sentiment Text Data Screening

ZHANG Xiao-Bin, HUANG Meng-Ying

(School of Computer Science, Xi'an Polytechnic University, Xi'an 710048, China)

Abstract: It was found that the process of carrying, storing, and forwarding data of sensing nodes ignored the content filtering of information carried by nodes by analyzing the cooperative process of mobile crowd sensing. As for purposeful data acquisition, this method of collecting data first and screening data later spent more time in analyzing and screening data. At the same time, the proportion of efficient data was not high. Taking these into account, a node selection mechanism based on sentiment text data screening in mobile crowd sensing environment combining genetic algorithm is designed. In the node selection mechanism, the perceptual nodes are selected by screening the data types so as to obtain the emotional text data of the mobile users in the perceptual environment. The experimental results show that the efficiency of data processing is increased by 27.6%, and the proportion of effective data is increased by 21% by using this method. Therefore, the method proposed in this study can effectively improve the efficiency of the whole data processing.

Key words: mobile crowd sensing; genetic algorithm; sentiment text; data screening; node selection mechanism

引言

随着移动智能设备的不断普及与应用, 使得越来越多的传感器应用到移动智能设备里, 这样就使得人们获取数据的方式越来越多样化, 群智感知服务的发展也越来越普及. 群智感知服务就类似于一种众包服务, 是将原来分配给员工做的任务包给拥有移动智能设备的群体进行完成并上传数据实现感知服务. 当某一项研究需要收集某些数据时, 就会发布一些感知任务, 收到任务的移动感知节点根据需要完成该感知任务并上传数据, 收集到这些数据的人员对数据进行分

① 基金项目: 陕西省自然科学基金 (2015JQ5157)
Foundation item: Natural Science Foundation of Shaanxi Province (2015JQ5157)
收稿时间: 2018-06-21; 修改时间: 2018-07-20, 2018-08-21, 2018-08-29; 采用时间: 2018-08-30; csa 在线出版时间: 2018-12-26

析处理, 从而得到研究结果. 在现有的研究中, 有很多关于群智感知应用, 感知问题面临的挑战以及解决办法等方面的研究.

Xu 等[1]通过收集和研究现有移动群智感知在社会感知上的应用, 提出并分析了潜在的新应用在未来方向上的一些可能. 当然, 感知应用也需要考虑到用户成本、网络压力、云计算服务器架设、用户隐私等问题[2]. 在参与者的问题上, 刘琰, 郭斌等[3]针对多感知任务并发的情况下对任务参与者的选择问题, 提出了三种任务参与者选择的解决方法, 同时通过实验验证, 这三种方法针对不同的情景能够实现最优的参与者选择, 达到感知服务的目的. Wang 等[4]通过对感知平台数据传递的问题的分析, 提出了端对端的参与者选择模型, 来实现对参与者的选择. 这两者在不同程度上解决了参与者选择上的问题, 但是基于用户隐私问题的考虑, 大部分研究者的解决方法是对用户采取激励措施来降低用户对隐私问题的顾虑, 用户基于激励与隐私问题的平衡来决定是否参与任务. 因此, 激励机制是群智感知研究的一个重要问题[5], 吴垚等[5]对群智感知服务做了分析, 同时介绍了 4 类主要的激励机制及其核心研究问题. 为了补偿用户参与感知任务的代价, 需要设计一种合理的激励机制[6], 从而使得能有更多的人能参与到感知任务中. 南文倩等[7]基于感知任务用户参与度的问题, 提出了基于跨空间的动态激励模型, 相比较传统的方法来说, 该方法有效地提高了用户参与感知任务的积极性, 从而使得感知任务能顺利的进行. 为了让更多的人自愿的参与感知任务, Cai 等[8]提出了一种具有多项式时间计算复杂度的近最优算法和一种关键的支付方案. 此方法能让用户在隐私与激励上做出平衡, 从而决定参与感知任务. 另外, 为了激励更多的参与者完成感知任务, Zhang 等[9]提出了三种在线反向拍卖的激励机制, 并取得了很好的成效. 以及 Micholia 等[10]说明了用户最终参与和贡献的不确定性, 同时提出了激励分配问题和迭代算法, 也取得了不错的效果.

以上研究中在参与者的选择问题上做了许多改进, 使更多的参与者能够参与到感知任务中并完成感知任务, 但是对于任务完成的质量却没有保证. Luo 等[11]针对激励用户参与感知任务以及感知数据的真实性问题, 采用协同的方式解决了这两个问题. 在数据传输成本问题上, 徐哲等[12]利用感知节点的社会属性提出了多任务分发算法, 降低了数据传输的时间消耗成本, 实现了节点之间的高效感知与通信. Ma 等[13]基于感知角度探讨了如何利用人类移动的机会主义特征来高效、有效地收集数据. 这三者虽然在数据的可靠性和数据的传输收集效率上做了改进, 但并未对数据做出有效的筛选, 因此在后续的数据处理效率上并没有有效的提高.

另外, 在群智感知的研究中可以利用群智感知收集的数据进行分析与研究, 找到某些数据直接的关联, 从而实现对具体事件的分析. 张佳凡等[14]基于新浪微博进行研究, 根据词频变化特征以及热词上下文语境能有效的对热点事件进行区分, 达到事件的感知效果. 李静林等[15]分析了车联网群智感知服务的三种感知模型, 从而提出了一种有效的感知方法解决了车联网的高效性与时效性之间的冲突. 以上的大部分研究都是对群智感知的整体问题与挑战进行的研究与分析, 很少有对于局部的数据的筛选问题进行解决的具体解决方法. 同时, 在传统的数据筛选问题上都是对基础数据, 如数值数据、统计数据的筛选.

本文主要是针对群智感知环境下的情感文本数据的筛选来进行感知节点的选择. 首先进行感知网络模型的构建, 结合遗传算法, 根据节点携带的数据进行分类, 从而根据适应度函数得到节点最终值选择节点迭代. 同时, 对遗传算法的复杂度进行了分析. 一般感知服务收集到的数据包含有时间、地理位置以及用户上传的文本等信息, 而本文主要是通过筛选出带有情感词的文本数据的节点, 对节点所含情感文本数据进行处理, 从而实现后续的情感分析处理. 本文提出的方法在一定程度上减少了不必要的数据获取以及避免了较大数据量的筛选, 从而实现了数据处理效率的提高.

1 感知网络模型

本文是利用混合群智感知服务对基于情感文本数据筛选的感知服务节点选取机制进行的研究.

图 1 是混合群智感知服务的模型图, 任务发布机构进行感知任务的发布, 任务发布之后由感知中心进行任务的分解发布给各个感知节点, 感知节点需要在感知中心进行注册才能进行感知任务的参与和完成. 感知节点完成感知任务上传感知数据, 感知中心是对在感知中心进行注册的感知节点上传的数据进行处理, 将数据进行集成并将数据报告上传至任务发布机构. 由于该机制是混合式的群智感知服务, 所以每个感知节点也是服务节点, 也有对数据的处理能力, 当处理能力不够就需要在感知中心进行注册让感知中心进行处

理. 在进行感知任务时都需要对感知节点进行选取, 本文主要是采用遗传算法对感知节点进行选择, 实现对特定情感文本数据的高效处理.

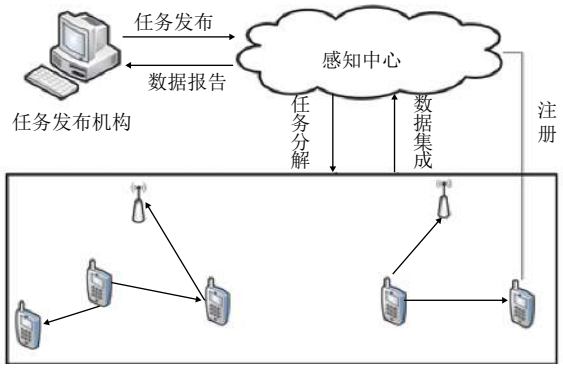


图 1 混合群智感知服务模型图

2 感知节点集选择

2.1 基于遗传算法的编码机制

在参与感知的服务节点集中选取符合条件的节点进行感知服务, 首先对节点进行编码, 本文采用遗传算法的编码机制, 采用常用的二进制编码. 现有服务节点集p里面包含 m 个服务节点, 即$p=\{p_1,p_2,\cdots,p_m\}$ 每个节点携带 m 个数据信息, 即$p_i=\{g_1,g_2,\cdots,g_m\}$, 当节点$p_i$被选取就将其归到服务节点集 中$p'$, 则$p'\subseteq p$, 同时将$p_i$对应位置的数据信息$g_i$标记为 1, 否则标记为 0, 即:

$$g_i=\begin{cases}1, \text{选取}p_i\text{节点}\\ 0, \text{不选取}p_i\text{节点}\end{cases} \tag{1}$$

2.2 适应度函数

在遗传算法中需要采取适应度函数对群体中的个体进行定量筛选, 然后根据适应度函数值的大小决定该个体是否淘汰. 本文是基于情感文本的数据筛选, 所以在节点信息携带的过程中本文主要考虑以下几个方面:

(1) 节点携带的信息数据类型. 定义数据类型 t, 取值为 1 和 0. 当 t=1 时表示数据类型为文本类型, 当 t=0 时表示数据类型是非文本类型.

(2) 节点携带的文本信息的内容是否包含情感词. 定义文本标记 s, 取值为 1 和 0. 当 s=1 时表示文本包含情感词, 当 s=0 时表示文本不包含情感词.

(3) 节点携带信息内容中包含情感词的多少. 用 c 表示, c_i表示包含情感词的数量.

结合以上三个方面的考虑, 对适应度函数的设计如下:

(1) 节点联合携带文本数据概率. 节点联合携带文本数据概率是指节点服务集中被选定的节点携带的信息数据中至少有一个是文本数据, 用 T 表示.

$$T=1-\prod_{p_i\in p'}(1-t_i) \tag{2}$$

其中, $t_i=\dfrac{m_i}{n_i}$表示被选定的节点i携带的信息数据中有文本数据的概率, m_i表示节点i携带的文本数据量, n_i表示节点i携带的信息数据量. T值越大说明节点携带文本数据的概率就越大.

(2) 联合包含情感词概率. 联合包含情感词概率是指节点服务集中被选定的节点携带的文本信息中至少有一个文本信息中包含情感词, 用S表示.

$$S=1-\prod_{p_i\in p}(1-s_i) \tag{3}$$

其中,

$$s_i=p(s|t)=\frac{p(st)}{p(t)} \tag{4}$$

即 s_i表示被选定的节点 i 携带的文本信息中包含情感词的概率, $p(st)$表示节点i携带的信息既是文本数据又包含情感词的概率, $p(t)$表示节点i携带的数据是文本数据的概率. S值越大说明节点携带的文本信息中包含情感词的概率就越大.

(3) 节点联合情感词数. 节点联合情感词数指节点服务集中被选定的节点携带的文本信息中包含情感词数量的平均值, 用C表示.

$$C=\frac{\sum\limits_{p_i\in p}s_i\times c_i}{n} \tag{5}$$

其中, c_i表示被选定的节点i携带的文本信息中包含情感词数量. C值越大, 表示被选定的节点组成的服务节点集中包含的情感词数越多.

由于本文主要是研究情感文本数据的筛选, 所以设置适应度函数$f(t,s,c)=C=\dfrac{\sum\limits_{p_i\in p'}s_i\times c_i}{n}$. 而基于上述, C的值与 T和 S的值相关, 所以在计算适应度函数值$f(t,s,c)$需先计算 T和 S的值.

3 遗传算法分析

3.1 遗传算法优化过程

根据第二节提到的编码机制和适应度函数, 基于

遗传算子操作设计的遗传算法及其优化过程如下:

1) 随机产生 M 个初始个体;

2) 根据式 (2)–式 (5) 计算种群中各个个体的适应度函数值, 并判断是否达到最大迭代次数, 若满足终止条件, 则输出结果, 否则, 执行步骤 3);

3) 根据种群中各个个体的适应度函数值, 依此执行选择, 交叉, 变异遗传算子, 产生新的种群, 并执行步骤 2).

具体的算子选择如下:

1) 在进行选择操作时采用比例选择算子, 该算子是指个体被选中的概率与适应度函数值是成正比的, 即在轮盘选择的时候, 适应度函数值越大的, 在盘面上的刻度越长.

2) 交叉算子是通过交换个体之间某个基因从而产生新个体的过程. 在对父个体进行的重组操作采用的是单点交叉算子, 从而使得选取的节点冗余比较少.

3) 变异运算采用基本位变异算子, 该算子是最简单的变异算子, 即是在某一个基因位上取反操作. 在变异的操作过程中需要选择合适的迭代次数, 从而得到最优解.

3.2 遗传算法复杂度分析

复杂度分析包括时间复杂度分析和空间复杂度分析, 在这里忽略空间复杂度的分析, 只分析时间复杂度. 遗传算法的时间复杂度可以认为是算法的平均计算时间, 即是求适应度函数最优值的平均计算时间.

N 个个体的初始化种群, 记为ξ_0, 经过重组后产生的新种群记为ξ_c, ξ_c中的每个个体x_i每一位以概率 p 取反, 产生新个体y_i, 得到种群ξ_m, 最后从ξ_m中选取 N 个最好的个体组成下一代种群, 重复以上重组和变异操作, 直到选出满足条件的个体结束.

在此过程中对适应度函数的平均计算时间是不超过O(n^2)的, 即该遗传算法的时间复杂度为O(n^2).

4 实验验证

4.1 实验准备与操作

实验选取 100 个愿意上传移动应用数据的志愿者进行实验, 使用 Matlab 实现. 志愿者的年龄范围在 20~50 之间, 男女比例 1:1. 每个志愿者相当于一个个体, 代表着感知节点, 每个个体在 100×100 区域内随机分布. 计算每个个体中每个基因是文本数据的概率, 包含情感词的概率 和包含情感词的数量 的值. 根据这些值算出每个个体的适应度函数值, 并将函数值降序排列, 选出前 N 个个体作为候选感知节点.

设置遗传算法程序运行的参数, 初始化的种群个体数 M=20, 终止进化迭代次数 G=500, 交叉操作的交叉概率 Pc=0.3, 变异操作的变异概率 Pm=0.06, 运行程序得到实验程序选出的个体. 再将选出的个体中的包含的文本信息数据整理出来, 同时记录实验所需的时间和数据量, 最后将实验中单位时间内处理的数据量进行归一化, 即是实验中的数据处理效率指标. 因此实验分为两部分: 一是对节点的选取, 二是对选取的节点携带的数据进行处理. 所以, 实验的总时间就是这两部分时间的总和. 数据处理效率指标也就是整个实验的处理效率指标. 另外, 算出选择出的节点携带的有效数据占比, 即是文本类型又包含情感词的数据占比.

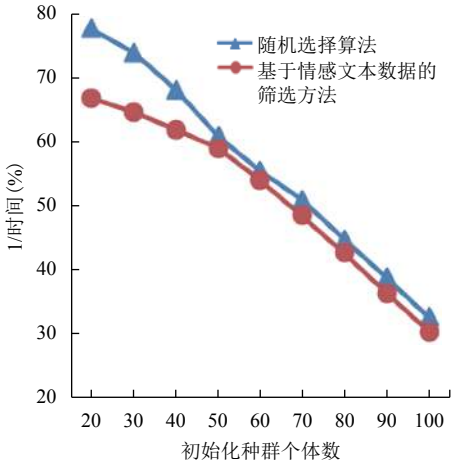
对于传统方法本文采用随机选择算法进行实验, 对于节点的选取是直到选择符合数量的感知节点为止. 后面的对数据处理的方法和本文中所用的数据处理方法一致.

4.2 实验结果与分析

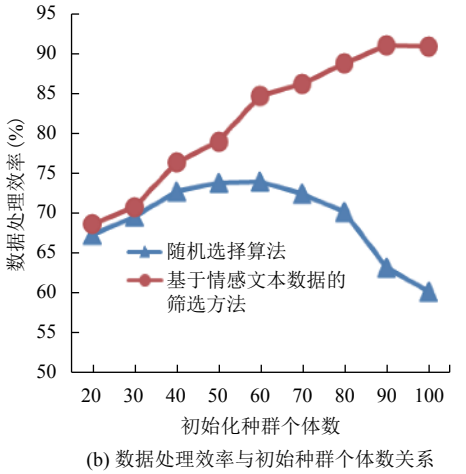
图 2 是在迭代次数、交叉概率和变异概率相同, 初始化种群个体数不同的情况下, 传统的数据处理与使用本文提出的算法进行的数据处理的对比. 图 2(a) 节点选取时间的倒数与初始种群个体数关系, 由图可以看到随着初始化种群个体数的增加, 两种方法在节点选取所需时间上的差距越来越小, 虽然在节点选取上本文方法在时间上会有一点延时, 但是由图 2(b) 可以看出, 两者的整体处理效率上本文方法相对较好, 并且随着初始化种群个体数的增加, 两者的整体处理效率相差越来越明显, 在初始化种群个体数达到 90 时, 差距达到最大, 相差 27.6%. 因此, 该方法能够有效的节省数据的处理时间.

图 3 是在初始化种群个体数为 20、交叉概率为 0.03 和变异概率为 0.06, 迭代次数不同的情况下, 传统的数据处理与使用本文提出的算法进行的数据处理的对比. 图 3(a) 是选取的感知节点个数与迭代次数之间的关系图, 图 3(b) 是数据处理效率与迭代次数对应选出的感知节点数之间的关系图. 由图 3(a) 可以看出, 随着迭代次数的增加, 节点选取的数量逐渐趋于稳定, 即在迭代次数达到 700 时, 最优的感知节点个数一定. 当最优感知节点个数达到一定时, 需要处理的数据量就是最优感知节点携带的数据量, 所以由图 3(b) 可以看

出, 在迭代次数达到 700 时, 数据处理的效率时趋于稳定的. 同时图 3(b) 可以看出, 使用本文提出的方法, 数据处理的效率优于传统的数据处理的效率. 图 4 是在迭代次数、交叉概率和变异概率相同, 初始化种群个体数不同的情况下, 传统的数据处理的有效数据占比与使用本文提出的算法进行的数据处理的有戏数据占比的对比. 由图 4 可以看出, 随着初始化种群个体数的增加, 两种方法的有效数据占比相差越来越明显, 在初始化种群个体数达到 90 时, 差距达到最大, 相差 21%. 实验表明, 该方法能够有效的提高获取的数据的有效性.



(a) 节点选取时间的倒数与初始种群个体数关系

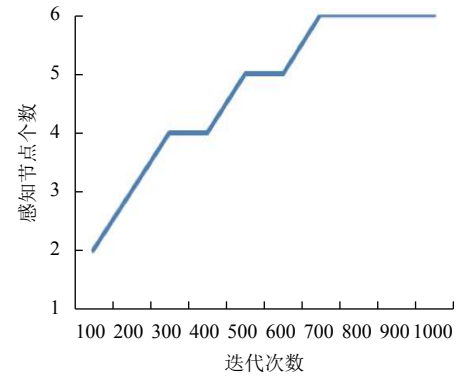


(b) 数据处理效率与初始种群个体数关系

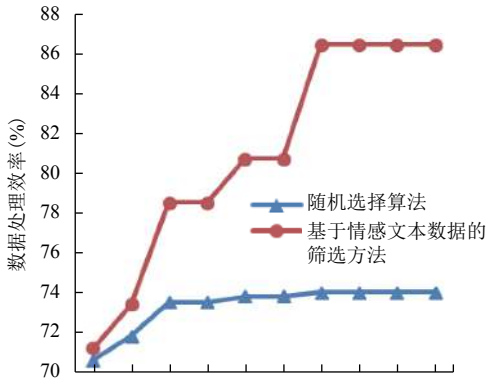
图 2 感知节点评价指标与初始种群个体数关系

5 结束语

本文通过对感知环境下节点数据的获取方法的分析, 提出了一种新的针对基于特定的情感文本数据筛选的节点选择机制. 基于遗传算法的优化选择过程, 得到所有节点集中符合优化条件的节点集, 从而实现感知服务对数据的获取. 实验结果表明, 本文提出的方法减少了对不必要数据的获取, 从而减少了在后续的数据处理的工作量, 能够提高对整体数据进行分析的效率. 在后续的研究中, 还需要根据实际情况考虑到遗传算法中遗传算子的选择问题, 实现对算法的进一步优化.



(a) 感知节点个数与迭代次数关系



(b) 数据处理效率与迭代次数关系

图 3 感知节点评价指标与迭代次数关系

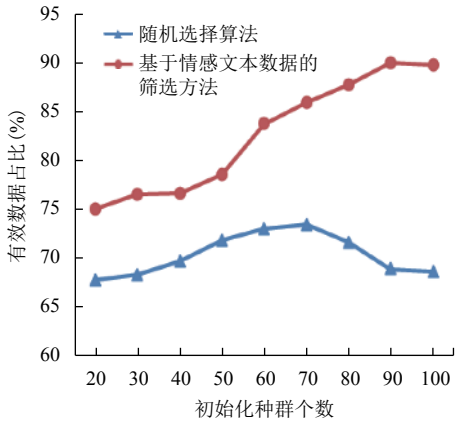


图 4 有效数据占比与初始种群个体数关系

参考文献

1 Xu Z, Mei L, Choo KKR, *et al*. Mobile crowd sensing of human-like intelligence using social sensors: A survey. Neurocomputing, 2018, 279: 3–10. [doi: 10.1016/j.neucom.2017.01.127]

2 陈荟慧, 郭斌, 於志文. 移动群智感知应用. 中兴通讯技术, 2014, 20(1): 35–37. [doi: 10.3969/j.issn.1009-6868.2014.01.008]

3 刘琰, 郭斌, 吴文乐, 等. 移动群智感知多任务参与者优选方法研究. 计算机学报, 2017, 40(8): 1872–1887.

4 Wang Y, Li HS, Li T. Participant selection for data collection through device-to-device communications in mobile sensing. Personal and Ubiquitous Computing, 2017, 21(1): 31–41. [doi: 10.1007/s00779-016-0974-0]

5 吴垚, 曾菊儒, 彭辉, 等. 群智感知激励机制研究综述. 软件学报, 2016, 27(8): 2025–2047. [doi: 10.13328/j.cnki.jos.005049]

6 Zhao D, Li XY, Ma HD. Budget-feasible online incentive mechanisms for crowdsourcing tasks truthfully. IEEE/ACM Transactions on Networking, 2016, 24(2): 647–661. [doi: 10.1109/TNET.2014.2379281]

7 南文倩, 郭斌, 陈荟慧, 等. 基于跨空间多元交互的群智感知动态激励模型. 计算机学报, 2015, 38(12): 2412–2425.

8 Cai H, Zhu YM, Feng ZN. A truthful incentive mechanism for mobile crowd sensing with location-sensitive weighted tasks. Computer Networks, 2018, 132: 1–14. [doi: 10.1016/j.comnet.2017.12.012]

9 Zhang XL, Yang Z, Zhou ZM, *et al*. Free market of crowdsourcing: Incentive mechanism design for mobile sensing. IEEE Transactions on Parallel and Distributed Systems, 2014, 25(12): 3190–3200. [doi: 10.1109/TPDS.2013.2297112]

10 Micholia P, Karaliopoulos M, Koutsopoulos I. Mobile crowdsensing incentives under participation uncertainty. Proceedings of the 3rd ACM Workshop on Mobile Sensing, Computing and Communication. Paderborn, Germany. 2016. 29–34.

11 Luo T, Kanhere SS, Tan HP. SEW-ing a simple endorsement web to incentivize trustworthy participatory sensing. 2014 11th Annual IEEE International Conference on Sensing, Communication, and Networking. Singapore. 2014. 636–644.

12 徐哲, 李卓, 陈昕. 面向移动群智感知的多任务分发算法. 计算机应用, 2017, 37(1): 18–23. [doi: 10.3969/j.issn.1005-8451.2017.01.004]

13 Ma HD, Zhao D, Yuan PY. Opportunities in mobile crowd sensing. IEEE Communications Magazine, 2014, 52(8): 29–35. [doi: 10.1109/MCOM.2014.6871666]

14 张佳凡, 郭斌, 路新江, 等. 基于移动群智数据的城市热点事件感知方法. 计算机科学, 2015, 42(S1): 5–9.

15 李静林, 袁泉, 杨放春. 车联网群智感知与服务. 中兴通讯技术, 2015, 21(6): 6–9. [doi: 10.3969/j.issn.1009-6868.2015.06.002]