DOI：10.13203/j.whugis20130066 文章编号：1671-8860(2014)06-0734-05

利用手机定位数据的用户特征挖掘

陈 佳[1] 胡 波[2] 左小清[3] 乐 阳[4]

1 福建省基础地理信息中心，福建 福州，350003
2 重庆市勘测院，重庆，420020
3 昆明理工大学国土资源工程学院，云南 昆明，650093
4 深圳大学空间信息智能感知与服务深圳市重点实验室，广东 深圳，518060

摘 要：手机定位数据已经逐渐成为一类新兴的空间数据，可用于分析个体或大规模区域内群体的活动特征，服务于基于位置的服务和城市及交通规划等。提出了一种基于手机定位数据，结合区域内兴趣点(POI)、房产价格等，利用空间聚类及语义分析等手段，对用户特征进行分析和挖掘的方法。首先采用 DBSCAN 方法提取用户重点活动区域；其次，根据用户的活动规律假设对活动区域进行类别标注；最后引入自然语言处理方法对 POI 和楼盘描述信息进行词频分析。并结合区域内 POI 类别和房价信息推断用户可能的偏好特征及收入或消费能力等特征，对用户一个月的手机定位数据进行挖掘分析。结果表明，该方法对用户重点活动区域及个体喜好特征等能够进行较为有效的挖掘。

关键词：数据挖掘；手机数据；用户特征；位置服务
中图法分类号：P208 **文献标志码**：A

空间轨迹数据挖掘是 GIS 的一个热点研究领域。随着数据获取手段的丰富，针对空间数据的研究也不再集中在测绘、国土等专门领域。近几年，除了传统的 GPS 定位数据[1]，逐渐兴起了利用多种形式的空间位置数据，如手机定位数据[2-6]、公交卡[7]签到(check-in)[8]以及带有地理标签的图片[9]等数据，对人类的空间活动范围和规律进行研究[10]。本文旨在利用手机定位数据，结合区域内兴趣点(point of interest，POI)、房产价格等，利用空间聚类及语义分析等手段，对手机用户的个体特征如习惯、喜好、收入或消费水平等进行分析和挖掘。了解这些用户特征，可以为用户提供更加精准的个性化空间搜索服务及地理围栏服务等。

1 手机定位数据简介

基站定位是目前手机定位最常见的方式。每次手机用户接打电话、收发短信或者使用数据通讯服务时，系统就记录一个手机的位置点。当用户手机长时间处于待机状态时，系统有时会对用户进行周期性的位置更新，探测手机所处的基站范围，一般是 2 h。所以基于基站定位的手机数据位置点并不是手机用户当时的精确位置，而是以手机所在的基站位置为准。在郊区，基站的范围可达几 km；在人口稠密的城区，多为 500 m 左右[11]。

研究发现，尽管手机定位数据是离散和稀疏的，但利用手机数据仍然可以对人们的活动进行高精度的预测[2]。该结论为人的行为和特征研究提供了理论前提。除了用户轨迹的可预测性，用户的手机定位数据在某种程度上也能反映用户的个体特征，如收入或消费水平，习惯偏好甚至职业特征等。本文尝试对手机定位数据分析挖掘用户的个人特征，并对此类数据及方法的可行性进行分析。

2 用户特征提取方法

2.1 重点活动区域识别

首先使用 DBSCAN(density-based spatial

收稿日期：2013-04-22
项目来源：国家自然科学基金资助项目(41231171，41171348，41061043)；深圳市科技研发资金资助项目(JCYJ20120109111128765，JCYJ20130329144141856)；CCF-腾讯犀牛鸟科研基金资助项目(CCF-TencentARG20130115)。
第一作者：陈佳，硕士生。现主要从事数字城市及移动轨迹挖掘工作。E-mail：chenjia305@126.com
通讯作者：乐阳，博士，副教授。E-mail：yueyang@szu.edu.cn

clustering of application with noise)[12]方法对用户的位置点进行聚类,将用户在地理空间上分布较为密集的点集表示为重点活动区域。由于手机基站覆盖范围存在一定的不确定性,活动区域由点集的外包圆形表示,而不是外接多边形[13]。

2.2　活动区域类别分析

已有研究利用精确定位信息(GPS 和蓝牙定位数据)根据各项活动的时间、位置及持续时间等特征,标注用户的家庭和工作及其他活动区域等[14]。本研究中,"工作"区域范畴包含了"学习"区域。本文针对手机数据粗时空粒度的局限,不需要用户的标注及任何的用户背景知识,仅基于用户的活动时间标注活动区域。本时间划分方法基于以下的假设前提:① 最多有两个区域是用户的家所在区域;② 用户至少有一个工作区域;③ 用户周一到周五每天都正常上班;④ 用户在工作时间内只出现在工作区域,不会出现在家、休闲娱乐等区域。基于以上假设,对用户的活动时间划分如下:家(24:00～06:00)、上班(08:30～11:30、14:30～17:30)、休闲娱乐/家(18:00～24:00)。由于人的活动特性比较复杂,在实际生活中,用户的作息规律可能并不符合本文的假设,所以它并不适用于所有用户。识别其他活动特征的用户,需要重新制定相应的规则。

本文提出的重点区域类别分析的流程按照如下顺序进行:"家"所在区域标签提取→"工作"区域标签提取→"休闲娱乐"区域标签提取。具体流程如下。

1) 提取"家"所在区域。将在 00:00～06:00 时间段内位置点比例最高的区域作为"家"所在区域;如果区域的位置点比例与最高比例的差值在 $1/n$ 之间(n 是聚类的个数),认为该区域的比例接近最大比例,该区域也被标记为"家";其他区域则标记为"休闲娱乐"区域。

2) 提取"工作"区域。分析在 08:30～11:30 和 14:30～17:30 时段内的聚类区域,如果在该时间段内的区域已被标记为"家",该区域重新标记为"家或工作";如果已被标记为"休闲娱乐",则更新标签为"工作或休闲娱乐";如果该区域未被标记,则将区域标记为"工作"区域。

3) 提取休闲娱乐区域。在 18:00～24:00 时间段内,如果区域已被标记为"家"、"家或工作"区域、"工作或休闲娱乐"区域中的任意一个标签,则该区域不被重新标记。如果该区域已被标记为"工作"区域,则将标签重新标记为"工作或休闲娱乐"区域;如果区域未被标记,将该区域标记为"休闲娱乐"区域。

4) 判断是否所有区域已被标记,如有区域未被标记,且该区域的位置点是用户在周末时间内形成的位置点,则将该区域标记为"休闲娱乐"。

即使已经分辨出了用户的工作地点、居住地点及经常光顾的休闲娱乐地点,这些空间位置信息仍然不足够用于标识用户的个体特征。因此,本研究引入自然语言处理(natural language processing, NLP)方法,对用户活动的重点活动区域进行语义分析。

2.3　活动区域及用户特征提取

1) 活动区域语义特征分析。本文尝试使用 POI 分类数据进行重点活动区域的语义特征分析,目的是找出用户活动区域的性质,如金融商业区、文教区、餐饮娱乐区等。首先使用计算机领域 NLP 的研究成果对 POI 名称进行词频分析,找出出现频率较高的几个词作为这个区域的标签。如"xx 银行"会被断为"xx"和"银行"两部分,那么在金融区,"银行"和"公司"及一些地名信息等出现的频率会比较高。利用开源的中文词频分析软件[15]进行分析,虽然词频分析有时仍会出现一些歧义,但是结合 NLP 进行土地利用分析会是一个有效的方法。另一方面,区域内 POI 的类别分布可以反映用户的某些兴趣偏好,本文选取频率排名前十的 POI 类别作为重点区域的特征和用户可能的偏好。

2)用户收入或消费水平分析。推断用户的收入或消费水平是推荐相关应用中重要的一个因素。本文通过用户活动的重点区域的房价信息分析用户的收入或消费水平。其假设是:收入或消费水平高的用户有能力购置地段较好的住房,其住房价格是评价用户消费能力的重要因素。本文将住房出售价格分为低、中低、中等、中高及高 5 个价格区间,统计用户住处所在区域内这 5 个价格区间所占的比例。同时,基于楼盘语义描述的区域标签提取,对重点区域内所有的楼盘语义描述信息进行词频分析,得出频率较高的 10 个词汇作为区域的语义标签。

3　实验与分析

3.1　实验数据

实验数据为用户一个月的手机位置数据。经过匿名处理后的手机位置数据示例如表 1 所示。其中,USE_ID 是用户编号;CELL_ID 是 Time 时刻用户所在的基站位置编号。使用的POI数据

表1 手机位置数据示例

Tab.1 Examples of Mobile Phone Location Data

Date	Time	Week	Pro_type	USE_ID	CELL_ID
2010-12-01	02:16:58	Wednesday	正常位置更新	812	22763
2010-12-01	01:12:15	Wednesday	短信	807	35343
2010-12-01	01:02:48	Wednesday	周期位置更新	895	51443
2010-12-01	00:29:56	Wednesday	周期位置更新	756	12412
2010-12-01	00:38:45	Wednesday	被叫 MTC	901	21773

包括名称、类别、经纬度信息;房产信息包括楼盘名称、地址、价格及经纬度和楼盘描述信息。POI数据及房价信息都可以从相应的网站上获取,如点评网及搜房网等。

3.2 算法实现及实验结果分析

以某一典型用户(用户865)为例(图1),用户手机定位数据的分析挖掘结果如下:活动天数为31 d,活动点数为1 572点,聚类点数为1 174点,聚类个数为4个,聚类比例为74.68%。

对图1中的4个用户重点区域$R1$、$R2$、$R3$、$R4$进行时间统计和区域类别标注,如图2所示。

1) 识别用户居住地。用户在00:00~06:00时间段有两个区域$R1$和$R4$,$R1$占的时间比例最大,因此将$R1$标记为"家"所在区域;$R4$所占比例较小,根据本文判断方法,将$R4$标记为"休闲娱乐"区域。

2) 识别用户工作区域。在08:30~11:30和14:30~17:30两个时间段内有4个区域,$R2$和$R3$未被标记,将$R2$和$R3$标记为"工作"区域;$R1$已被标记为"家",将该标签更新为"家或工作"区域;$R4$在1)中已被标记为"休闲娱乐"区域,将该标签更新为"工作或休闲娱乐"区域。

3) 识别用户休闲娱乐区域。在18:00~24:00时间段内,将标签中含有"家或休闲娱乐"的区域筛选掉,对标记为"工作"的区域,将标签更新为"工作或休闲娱乐"区域,未进行标记的区域标记为"休闲娱乐"区域。

在3)中,如果有区域已被标记为"工作"区域,根据本文判断法则,该区域的标签会被更新为"工作或休闲娱乐"区域;但是在实际情况中,有些性质的工作需要晚上加班,用户没有进行休闲娱乐活动,该区域仅仅是用户的工作区域。所以,如何设定判断前提是一个复杂的问题。

按以上分析方法,$R1$最有可能是此用户的居住地,$R2$是其主要工作地点。在工作时间,此用户还经常出现在$R3$和$R4$区域。所以,$R2$、$R3$和$R4$都被标注为此用户工作的地点,但在这三个区域内的时间,$R2>R3>R4$。另外,由图3所示,此用户并非典型的朝九晚五类型,所以$R2$、$R3$和$R4$也分别被标注了"休闲娱乐"的类别。因为涉及隐私,无法和真实的用户信息进行核对。有一种可能是,如用户在这三个区域内分别经营店铺。

对POI名称、楼盘描述信息进行词频分析,分别提取这4个区域的语义描述信息,并以标签云方式显示,图3(a)、3(b)、3(c)分别显示了$R1$、$R2$及$R3$三个区域对应的信息。虽然目前利用词频分析得到的结果并不十分理想,但是如果针对此类应用对通用的词频分析进行约束和改进,结合NLP进行土地利用分析会是一个值得关注的研究方法。

$R1$是用户的家所在区域,中高价位所占比例较大,$R1$区域的房价信息统计如图4(a)所示;$R2$是用户主要工作所在区域,$R2$区域的POI特征统计如图4(b)所示,手机销售、火锅店、餐厅等类别的POI所占比例较大,同时房价信息也显示$R2$内的中高价位与高价位楼盘居多。图4(c)显示了$R3$区域内的POI特征统计,且中高房价占有较大比例。总体来说,用户家和工作的区域,其房价都比较高,因此推测用户收入或消费能力可能比较高。

表2总结了以上信息,其中,H表示家所在区域;W表示工作区域;E表示休闲娱乐场所;H-W表示该区域是家所在区域或工作区域;W-E表示该区域是工作区域或休闲娱乐区域。

由于篇幅限制,本文仅对一个位置点较多的用户进行了说明,而对于位置点较少的用户,需要相应地调整聚类参数,以生成有意义的聚类区域。虽然每个聚类区域并不一定能提取出有意义的特征标签,但是在目前的数据及隐私条件下,这并不失为一个可行和值得继续探讨的方法。

表2 用户标签

Tab.2 User Labels

	重点区域			
	$R1$	$R2$	$R3$	$R4$
区域类别	H-W	W-E	W-E	W-E
区域标签	银河 烟草 盘龙	滇池 张官营道口	商业街 霖雨	西核村
消费能力	中高	中高-高	中高	无
POI标签	火锅 网吧 棋牌室	火锅 网吧 川菜	火锅店、云贵菜、川菜	无

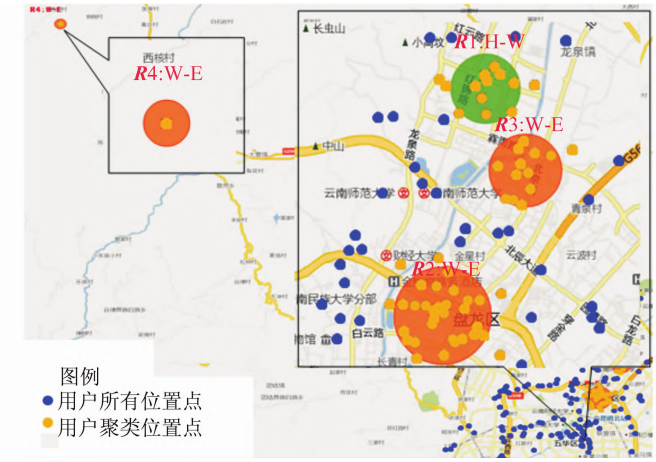


图 1　用户 865 位置点及聚类区域显示

Fig. 1　Location Points and Clusters of User 865

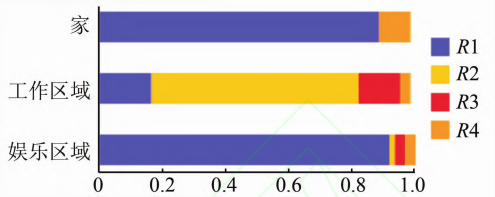


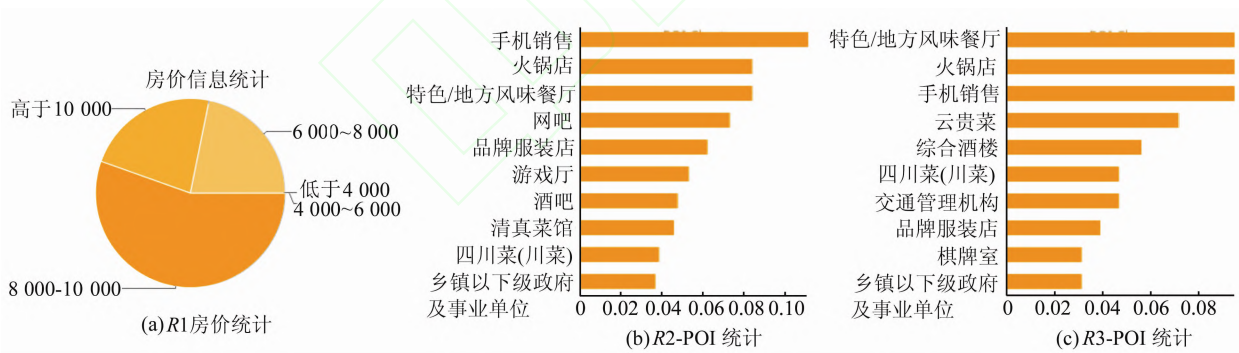
图 2　用户 865 活动区域时间统计图

Fig. 2　Activity Time Distribution of User 865

(a) *R*1　(b) *R*2　(c) *R*3

图 3　重点活动区域的云标签

Fig. 3　Tag Clouds of Hot Regions



(a) *R*1房价统计　(b) *R*2-POI 统计　(c) *R*3-POI 统计

图 4　重点活动区域的 POI 特征

Fig. 4　POI Features of Hot Regions

3.3　结果分析

为验证以上方法,本文对 32 名不同年龄段、分属 10 个城市的用户进行了问卷调查。首先获得调查者的重点活动区域;然后用以上方法根据用户提供的活动区域,利用 POI 提取用户特征,再根据房价分析用户收入或消费水平;最后让用户进行反馈评价。根据 POI 进行用户特征挖掘准确率的结果为:对用户特征评价为“准确”的百分比(49%)高于“部分准确”(33%)和“不准确”(18%);根据房价估计用户收入或消费水平准确率的结果为:收入或消费水平评价为“符合”的百分比(45.83%)高于“基本符合”(37.5%)与“不符合”(16.67%)。这说明本文提出的通过获取手机定位数据,结合 POI 及房价数据来挖掘用户的特征及收入消费水平的方法具有一定的可行性。

4 结 语

手机及可穿戴等个人定位数据越来越普遍，本文尝试利用时间和空间分辨率都较低的手机定位数据识别用户的重点活动区域，并引入NLP算法，结合网站上公开的POI数据和房价信息挖掘用户个体特征，如兴趣爱好及收入消费水平。提出了研究方法和技术路线，并对实例用户进行了分析。由于涉及个人隐私问题，本实验结果无法进行对应的个人用户验证，但是运用本文方法对不同年龄、不同地区的个体进行了问卷调查，说明利用位置数据、POI和房产数据的语义特征进行用户特征挖掘具有一定的可行性。研究中发现，词频分析仍具有一定的局限性，未来可以考虑将话题模型(topic model)引入土地利用分析，对空间对象的属性信息进行语义分析，从而更准确地得出重点区域的语义特征。

参 考 文 献

[1] Ashbrook D, Starner T. Using GPS to Learn Significant Locations and Predict Movement Across Multiple Users[J]. *Personal and Ubiquitous Computing*, 2003, 7(5): 275-286

[2] Song C, Qu Z, Blumm N, et al. Limits of Predictability in Human Mobility[J]. *Science*, 2010, 327(5 968): 1 018-1 021

[3] Song C, Koren T, Wang P, et al. Modelling the Scaling Properties of Human Mobility[J]. *Nature Physics*, 2010, 6(10): 818-823

[4] Ahas R, Silm S, Järv O, et al. Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones[J]. *Journal of Urban Technology*, 2010, 17(1): 3-27

[5] Zhu Lili, Wang Jinhua. Cell Phone Location Service System[J]. *Journal of East China University of Science and Technology (Natural Science Edition)*, 2007, 33(z1): 21-23(朱丽莉，王金华. 手机定位服务系统的研究[J]. 华东理工大学学报(自然科学版)，2007, 33(z1): 21-23)

[6] Luo Yong, Wang Yanmin, Zhang Jianqin. Research of Residents' Travel Information Mining and Analysis Methods Based on Mobile Phone Location Data [J]. *Journal of Beijing Institute of Civil Engineering and Architecture*, 2012, 28(1): 40-44(罗勇，王晏民，张健钦. 基于手机位置数据的居民出行信息挖掘和分析方法研究[J]. 北京建筑工程学院学报，2012, 28(1): 40-44)

[7] Long Ying, Zhang Yu, Cui Chengyin. Identifying Commuting Pattern of Beijing Using Bus Smart Card Data[J]. *Acta Geographic Sinica*, 2012, 67(10): 1 339-1 352(龙瀛，张宇，崔承印. 利用公交刷卡数据分析北京职住关系和通勤出行[J]. 地理学报，2012, 67(10): 1 339-1 352)

[8] Lian Defu, Xie Xing. Learning Location Naming from User Check-In Histories[C]. The 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM GIS 2011, Chicago, USA, 2011

[9] Arase Y, Xie Xing, Hara T, et al. Mining People's Trips from Large Scale Geo-tagged Photos[C]. The 18th ACM International Conference on Multimedia 2010, Firenze, Italy, 2010

[10] Wakamiya S, Lee R, Sumiya K. Crowd-based Urban Characterization: Extracting Crowd Behavioral Patterns in Urban Areas from Twitter[C]. The 3rd ACM SIGSPATIAL International Workshop on Location-based Social Networks, Chicago, Illinois, USA, 2011

[11] Trevisani E, Vitaletti A. Cell-ID Location Technique, Limits and Benefits: An Experimental Study [C]. The 6th IEEE Workshop on Mobile Computing Systems and Applications, Windermere, Cumbria, UK, 2004

[12] Ester M, Kriegel H P, Sander J, et al. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise[C]. The 2nd International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, 1996

[13] Eddy W F. A New Convex Hull Algorithm for Planar Sets[J]. *ACM Transactions on Mathematical Software (TOMS)*, 1977, 3(4): 398-403

[14] Adams B, Phung D, Venkatesh S. Sensing and Using Social Context[J]. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 2008, 5(2): 11

[15] Institute of Computing Technology Chinese Academy of Science. ICTCLAS Chinese Lexical Analysis System[EB/OL]. http://www.ictclas.org/, 2012(中国科学院计算技术研究所. ICTCLAS汉语分词系统[EB/OL]. http://www.ictclas.org/, 2012)

(下转第744页)

Driving Behavior Analysis Based on Trajectory Data Collected with Vehicle-mounted GPS Receivers

REN Huijun[1] *XU Tao*[1] *LI Xiang*[1]

1 Key Laboratory of Geographical Information Science, Ministry of Education, East China Normal University, Shanghai 200241, China

Abstract: Traffic congestion is becoming an increasingly serious problem in highly urbanized areas. One reason for congestion are frequent happened traffic accidents caused by risky driving behaviors. The accurate evaluation of a driver behavior as well as safety analysis has therefore become a research hotspot. In this paper, a new method of using the vehicle-mounted GPS model to collect the vehicle track data and analyzing the safety of driver's behavior is presented. In the approach, information about driving over the speed limit, quick acceleration, sharp slowdown and sharp turns is extracted from the trajectory data. Then, the security of a drivers' behavior assessed to provide scientific evidence for the transportation department managers when evaluating and managing a drivers' driving skill.

Key words: driving behavior analysis; GPS; safe driving; trajectory data

First author: REN Huijun, master, specializes in the GIS for transportation and spatial analysis. E-mail: fjbsm_jcl@163.com

Corresponding author: LI Xiang, PhD, professor. E-mail: xli@geo.ecnu.edu.cn

Foundation support: The Doctoral Fund of Ministry of Education of China, No. 20130076110014; the National 863 Program of China, No. 2013AA122302; the National Natural Science Foundation of China, No. 41271441; the National Natural Science Foundation of Shanghai in China, No. 11ZR1410100.

(上接第738页)

Personal Profile Mining Based on Mobile Phone Location Data

CHEN Jia[1] *HU Bo*[2] *ZUO Xiaoqing*[3] *YUE Yang*[4]

1 Fujian Provincial Geomatics Center, Fuzhou 350003, China

2 Chongqing Survey Institute, Chongqing 420020, China

3 Faculty of Land Resource Engineering, Kunming University of Science and Technology, Kunming 650093, China

4 Shenzhen Key Laboratory of Spatial Smart Sensing and Services, Shenzhen University, Shenzhen 518060, China

Abstract: Understanding personal profiles like preferences, income levels, and geographical areas is the basis of providing a person with personalized and accurate services. In order to acquire personal profiles we propose a reasonable technical route that first extracts the geographic regions from personal mobile phone location data based on a density-based clustering algorithm. Then, the geographic regions are tagged with semantic meaning and we analyze house descriptions by NLP(Natural Language Processing). A division method for people's daily time is given, based on the assumed the activity patterns of people. At last, an individual's taste for something or his income levels is analyzed using the statistics for POIs and house prices in the extracted places. An experiment with real data shows that this method is an effective solution to mining personal profiles.

Key words: data mining; mobile phone location data; user profile; location based services

First author: CHEN Jia, postgraduate, specializes in digital city and mobile trajectories mining and analysis. E-mail: chenjia305@126.com

Corresponding author: YUE Yang, PhD, associate professor. E-mail: yueyang@szu.edu.cn

Foundation support: The National Natural Science Foundation of China, Nos. 41231171, 41171348, 41061043; Shenzhen Scientific Research and Development Funding Program, Nos. JCYJ20121019111128765, JCYJ20130329144141856; CCF-TencentARG20130115.