软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(8):2148−2160 [doi: 10.13328/j.cnki.jos.005118]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

一种面向获取空间信息的潜在好友推荐算法*

俞 菲[1], 李治军[1], 车 楠[2], 姜守旭[1]

[1](哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

[2](哈尔滨理工大学 软件学院,黑龙江 哈尔滨 150001)

通讯作者: 姜守旭, E-mail: jsx@hit.edu.cn

摘 要: 随着社交网络的不断发展,朋友推荐已成为各大社交网络青睐的对象,在能够帮助用户拓宽社交圈的同时,可以通过新朋友获取大量信息.由此,朋友推荐应该着眼于拓宽社交圈和获取信息.然而,传统的朋友推荐算法几乎没有考虑从获取信息的角度为用户推荐潜在好友,大多是依赖于用户在线的个人资料和共同的物理空间中的签到信息.而由于人们的活动具有空间局部性,被推荐的好友分布在用户了解的地理空间,并不能满足用户通过推荐的朋友获取更多理信息的需求.采用用户在物理世界中的签到行为代替虚拟社交网络中的用户资料,挖掘真实世界中用户之间签到行为的相似性,为用户推荐具有相似的签到行为且地理位置分布更广泛的陌生人,能够增加用户接受被推荐的陌生人成为朋友的可能性,在保证一定的推荐精度的基础上,增加用户的信息获取量.采用核密度估计估算用户签到行为的概率分布,用时间熵度量签到行为在时间上的集中程度,选择可以为用户带来更多新的地理信息的陌生人作为推荐的对象,通过大规模 Foursquare 的用户签到数据集,验证了该算法能够在精度上保证与目前已有的 LBSN 上陌生人推荐算法的相似性,在信息扩大程度上高于上述已有算法.

关键词: LBSN(location-based mobile social network);朋友推荐;核密度估计;签到行为概率分布
中图法分类号: TP311

中文引用格式: 俞菲,李治军,车楠,姜守旭.一种面向获取空间信息的潜在好友推荐算法.软件学报,2017,28(8):2148−2160. http://www.jos.org.cn/1000-9825/5118.htm

英文引用格式: Yu F, Li ZJ, Che N, Jiang SX. Potential friend recommendation algorithm for obtaining spatial information. Ruan Jian Xue Bao/Journal of Software, 2017,28(8):2148−2160 (in Chinese). http://www.jos.org.cn/1000-9825/5118.htm

Potential Friend Recommendation Algorithm for Obtaining Spatial Information

YU Fei[1], LI Zhi-Jun[1], CHE Nan[2], JIANG Shou-Xu[1]

[1](School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

[2](School of Software, Harbin University of Science and Technology, Harbin 150001, China)

Abstract: Along with the development of online social networks, friend recommendation becomes the favor of the major social networks. It can help people to meet new friends for expanding the scale of social network, which in turn allows people to receive more information from their friends. Therefore, friend recommendation should be focused on expanding the scale of social network and obtaining information form recommended friends. However, existing friend recommendation methods barely consider the people information need, and they are mainly based on the simple and limited user profiles, and are agnostic to users' offline behaviors in the real world. Because human activity in the physical world has a spatial locality, the recommended friends through the existing recommendation methods are limited in geographic space which the target user know. As a result, the recommendation cannot provide more new information on geography to meet the target's need on information. This paper first proposes a new friend recommendation method based

* 基金项目: 国家自然科学基金(61370214, 61300210)
Foundation item: National Natural Science Foundation of China (61370214, 61300210)
收稿时间: 2015-06-23; 修改时间: 2016-03-18, 2016-06-12; 采用时间: 2016-06-30; jos 在线出版时间: 2016-10-11
CNKI 网络优先出版: 2016-10-12 16:27:01, http://www.cnki.net/kcms/detail/11.2560.TP.20161012.1627.028.html

on the offline check-in behaviors in the real world instead of the online user profiles, and mines check-in behavior similarity between users in the real world. The essential goal of friend recommendation is to provide users with more new information. In order to meet the requirements of user getting more geographical location information, the recommendation systems can recommend the strangers in broader check-in geography distribution for the target users. Meanwhile, when the recommended friends and the target users have the similar check-in behaviors, it is more probable for the users to accept recommended strangers. Kernel density estimation (KDE) is used to estimate each user's check-in behavior probability distribution and the time entropy to filtering some noise that have side effects on overall check-in behavior similarity, then the recommended strangers who can bring a wider range of new strangers geographic information for the target users are selected. Lastly, a large-scale user check-in data-set of Foursquare is used to validate recommendation precision and the degree of information expanding of this approach. The experimental results show that the proposed approach outperforms the existing friend recommendation methods on the aspect of the information expanding degree while maintaining the recommendation precision of the state-of-the-art stranger recommendation methods.

Key words: LBSN (location-based mobile social network); friend recommendation; kernel density estimation; check-in behavior probability distribution

近年来,基于位置的社交网(location-based social networks,简称 LBSN)得到了广泛应用,以 Foursquare 为代表的网站有 Gowalla、街旁、QQ 等.随着移动互联网深入到人们的生活、学习和工作中,人们对基于位置的网络服务的需求越来越多[1−5].LBSN 和传统的社交网络都为用户提供一个社交平台,朋友推荐[6−8]随之也成为 LBSN 上的主要应用.然而,LBSN 上推荐问题的研究与传统社交网络的本质区别在于:LBSN 上的推荐算法[9]利用了用户的物理行为信息;传统社交网络中的推荐问题[10]利用人们在虚拟的社交网络中的行为,这些线上行为不能真实地反映用户在物理世界中真实的性格特征和行为习惯.与此同时,社交网络中的用户具有社会需求(拓宽朋友圈)和获取信息的需求[11].然而,传统的好友推荐算法几乎没有从信息获取的角度考虑为目标用户推荐潜在好友.在 Marketing Letters 的社会调查中发现:大约有 4/5 的美国人在选择餐厅时,相对于传统的商业广告会更倾向于询问家人或者朋友,这说明人们获取信息的途径更倾向于从朋友那里获得.从美国社会学家马克·格拉诺维特(Mark Granovetter)于 1974 年提出的弱连接理论中可知:虽然弱连接不像强连接那样坚固,但在信息获取方面比强连接有明显优势[12].所以,为用户推荐陌生人体现出“弱链接”在社交网络中信息传播的高效能.目前,各大交友网站都推出了为用户推荐周围的陌生人的推荐系统[13],移动设备上具有定位功能的 APP 可以帮助用户与附近的陌生人认识,发现周围与自己兴趣相同的人.由此,本文提出一种在拓宽用户社交圈子的同时,可以满足用户信息获取需求的好友推荐算法.

基于位置的社交网络能够记录用户的签到轨迹,人们在物理世界中的行为信息(例如关于位置的签到信息)可以更加真实地描述人们的性格和偏好特征,所以能够提高推荐的精度[1].已有的 LBSN 上的潜在好友推荐算法[1,14,15]是基于用户空间活动位置相似性来推荐潜在好友.从文献[16–18]可知,人类运动轨迹显示出时空的规则性[16−18].文献[19]指出,位置偏好和访问评论、停留时间成正相关.由文献[20]提出的 Tobler 第一地理定律可知:相对于远距离的位置,人们更倾向于近距离的位置.文献[15]指出,人们倾向于访问家附近的空间位置.并通过 Foursquare 真实数据集的统计分析得出总结性结论:75%的用户访问 50 英里距离内的地方.进一步说明了人们的活动具有空间局部性.基于以上分析可知,基于活动位置相似性推荐的潜在好友主要分布在目标用户了解的地理空间.所以在本文提到的信息获取方面,他们不能提供更多新的地理信息给目标用户.基于用户的签到信息,两个陌生人签到行为越相近,则说明他们之间相似的真实世界中的性格偏好和行为习惯就越多,从而使得他们成为朋友的概率就越高.所以,文献[21]提出为目标用户即时推荐周围与其具有相似的签到行为的陌生人来扩大用户的社交圈的推荐算法.以往基于熟悉的社交圈的朋友推荐算法和基于签到行为相似性的即时朋友推荐算法[21]大部分是采用协同过滤[22]、随机游走[23]、遗传算法[23]、加权泰森图[24]等算法,这些推荐算法从推荐对象上划分为两种:(1) 基于用户在真实世界中熟悉的社交圈的朋友推荐算法;(2) 基于附近具有相似的社交网络拓扑信息的陌生人推荐算法.基于熟悉的社交圈子的朋友推荐主要利用两种信息:一种是社交网络的拓扑信息;另一种是非拓扑信息(共同的朋友;用户资料:名字,年龄,头像,性别,生日和家乡;空间信息等).其中,文献[25]通过确定网络结构的性质来确定节点之间的相似性;文献[26,27]采用对社交网络中的节点进行加权的方法来

确定节点之间的相似性.由于本文推荐的是与目标用户没有朋友关系且没有公共朋友的潜在好友,所以以上这些基于拓扑信息的推荐方法不适用于本文提出的问题.文献[28]采用非拓扑信息找到与目标用户相似的潜在好友.本文是在 LBSN 上做潜在好友的推荐,并且以往的基于非拓扑信息的潜在好友推荐算法[28]几乎没有利用真实世界中用户的离线社会活动行为.文献[24]在好友推荐算法中融入了用户的空间活动位置和停留时间,并结合了 GPS 信息和社交网络结构.文献[29,30]提出主要基于用户间移动轨迹相似性的好友推荐算法和位置推荐算法.然而,这些利用物理世界用户的活动信息的推荐算法没有从获取信息角度为目标用户推荐好友,并且在用户之间的相似度比较上主要着眼于拓扑信息和非拓扑信息,都没有利用物理世界的签到行为的相似性.所以本文提出一种不仅可以满足用户拓宽社交圈的需求,也可以满足用户获取空间地理信息需求的潜在好友推荐算法.并且利用用户真实世界中的签到行为挖掘用户间的相似性,为目标用户推荐可以给其带来更多空间信息的潜在好友.本文的贡献主要包括以下几点.

1) 本文提出利用用户在真实世界中的签到行为的相似性,为用户推荐可以带来更多信息的地理信息的潜在好友推荐模型,并且模型满足用户通过朋友获取更多新的地理信息的需求.
2) 由于真实签到数据的稀疏性,本文提出利用核密度估计来估算用户在一天 24 个时间槽中签到行为的概率分布,同时引入时间熵度量每段时间的签到行为的集中程度.在精度上,与以往在 LBSN 上基于核密度估计处理数据稀疏问题的朋友推荐算法相比具有比较好的效果.
3) 最后,本文用大规模真实用户 Foursquare 上的签到数据集验证本文提出的方法的有效性.实验结果表明,我们提出的方法为用户推荐可以获得更多新的地理信息的潜在好友.同时验证了目标用户与被推荐的用户成为朋友后,目标用户随后受到推荐用户的影响,也签到了推荐用户签到过的地方.

本文第 1 节提出基于签到行为相似性,为目标用户推荐获取更多地理信息的潜在朋友推荐算法.第 2 节是本文的实验部分,通过与已有的基于位置的社交网络陌生人推荐算法的对比,分析本文提出的算法的有效性.第 3 节是本文的总结.

1 基于签到行为的推荐算法

本节提出利用用户的签到行为分布度量陌生人之间的相似性.由于签到行为在一天 24 个时间槽中的分布的稀疏性,我们采用核密度估计去估算每个用户的签到行为概率分布,并基于签到行为相似性,挑选可以为目标用户带来更多新的地理信息的陌生人作为推荐的朋友.

本节首先列出本文用到的符号及其含义,并给出相关定义.然后介绍核密度估计的基本知识,最后介绍基于签到行为的推荐算法的基本原理与实现方法.

1.1 符号含义与相关定义

若无其他说明,本文出现的符号及其含义见表 1.

Table 1 Symbols and meanings

表 **1** 符号及含义

符号	含义
U	数据集中所有用户集合
u_T	目标用户 Target User
U_0	目标用户 u_T 的朋友集合 $U_0=\{u_1,u_2,\ldots,u_n\}$
U_C	候选推荐用户 Candidate User 集合(即与目标用户具有相似签到行为的陌生)
U_{RC}	最终被推荐用户 Recommendation Candidate User
S_T	Target User 的 Check-in Location ID 集合
S_C	Candidate User 的 Check-in Location ID 集合

1.2 信息熵

在信息论中,信息熵常用于度量随机变量的不确定性.不确定性越大,信息熵越大.

设 $E=\{e_1,e_2,\ldots,e_L\}$表示由一系列随机事件 e 构成的随机变量,则 E 的信息熵计算公式为

$$H(E)=-\sum_{l=1}^{L}p_l\log_2(p_l) \tag{1}$$

其中,p_l 表示随机事件 e_l 发生的概率.

1.3 核密度估计

核密度估计是一种非参数概率密度估计方法.

定义. 设 $x_1,x_2,\ldots,x_n$ 为取值于 R 的独立同分布随机变量,$f(x)$是随机变量的概率密度函数,$x\in R$,定义函数:

$$\hat{f}_h(x)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x_i-x}{h}\right),x\in R \tag{2}$$

公式(2)称为密度函数 $f(x)$的核密度估计,其中,$K(\cdot)$称为核函数;h 为预先给定的正数,通常称为窗宽或光滑参数.同时,核函数 $K(\cdot)$需要满足:

$$\int_{-\infty}^{\infty}K(u)\mathrm{d}u=1,\int_{-\infty}^{\infty}uK(u)\mathrm{d}u=0 \tag{3}$$

1.4 签到行为概率分布估计

目前,大多数交朋友的应用软件都是基于静态的用户资料.一般的推荐系统只是提供一列目标用户附近的推荐用户列表;或者依据目标用户对关于个人资料的筛选,系统为用户推荐满足筛选要求的推荐列表.因此,用户只能依据在线网络中固有的用户资料做选择,而不能依照被推荐用户真实世界中离线行为做选择.然而,人们在物理世界中的离线行为暗含着用户自身真实的个人喜好和生活习惯的性格特征[19,31,32],因此,物理世界中的用户行为习惯可以提高推荐的质量.此外,基于位置的移动社交网络记录了每个用户在物理世界中的签到行为,包括签到的时间和位置(例如兴趣点(point of interest,简称 POI)).签到行为记录了每个用户在物理世界中真实的行动轨迹,并且体现出了用户真实性格特征和喜好.研究结果表明,人类的行为轨迹体现出高度的时空规则性[16−18].目前,社交网络中的朋友推荐算法[1,24]主要是基于公共的朋友和相同的活动经历,然而对于缺少这些相似性的陌生人推荐算法[21],则主要是基于爱好及行为习惯的相似性进行社交网络中潜在好友的推荐.所以,本文利用用户在物理世界中的签到行为概率分布来度量两个用户之间的相似程度.两个用户的签到行为概率分布越相近,则说明两个用户的签到行为越相似.签到行为概率分布考虑到用户在一天 24 个等分的时间槽中的签到行为分布情况,体现出用户真实世界中针对签到行为特征.

然而,如果本文只用每个用户综合已有的历史数据统计一天 24 个时间槽中各个时间槽签到的次数,得到的签到行为的频率来估计用户的签到行为概率分布,会存在如下问题.

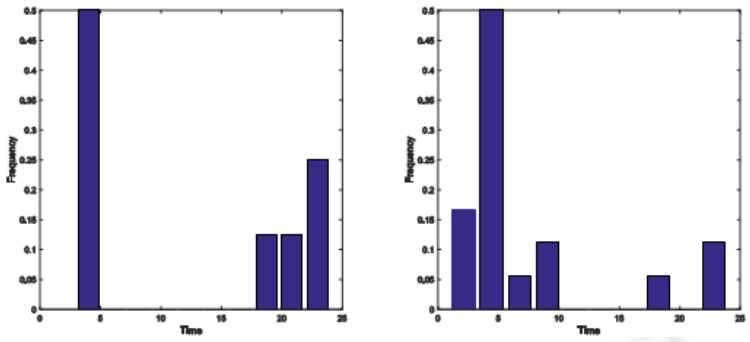
将一天分成 24 个时间槽,由于用户的签到行为在 24 个时间槽的签到次数统计频率是稀疏的.所以用签到频率估计用户的签到行为的概率分布是不精确的.例如,已知用户 u_a 和 u_b 的历史签到数据,对这些历史数据进行 24 个时间槽的签到次数的频率统计,如图 1(a)和图 1(b)所示.在 24 个时间槽中,用户 u_a 和 u_b 在大部分的时间槽中是没有签到行为的.所以,如果本文只用简单用户签到次数统计频率来估算用户的签到行为概率分布,是无法比较两个用户签到行为相似性的,因为签到次数频率只能给出在个别时间槽的离散的概率,而不能反映用户在 24 个时间槽中连续的签到行为分布.

本文依照采用核密度估计来估算用户在其他时间槽中的签到概率,并有效解决了数据稀疏问题.同时,也使比较用户之间签到行为相似性更加准确.

本文将要介绍如何采用核密度估计来估算用户的签到行为的概率分布.已知数据集 $X=(1,2,\ldots,24)$,已知用户 u_a 和 u_b 对应每个时间槽的频率分别为

$$Y_a=\{0,0,25.00\%,1,0,\ldots,12.5\%,12.5\%\},\ Y_b=\{5.56\%,0,\ldots,0,0,5.56\%,11.11\%\}.$$

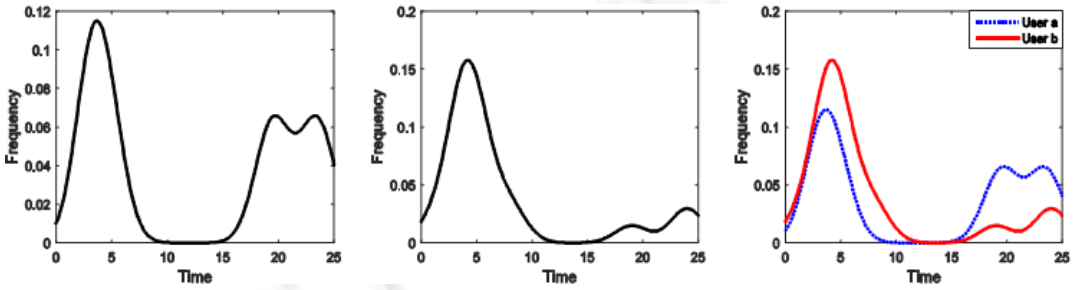
通过核密度估计,得到用户 u_a 和 u_b 在 24 个时间槽中的签到分布,如图 2 所示.



(a) u_a 的签到频率直方图 (b) u_b 的签到频率直方图

Fig.1 Frequency-Based check-in behavior probability estimation

图 1 基于频率的签到行为概率估计



(a) u_a 的签到概率密度 (b) u_b 的签到概率密度 (c) u_a 和 u_b 的签到概率密度比较

Fig.2 Probability-Based check-in behavior probability estimation

图 2 基于概率的签到行为概率估计

1.5 用户签到行为相似测度的计算

1.5.1 余弦距离

用户之间在物理世界中的离线签到行为差别,通过比较他们分别的签到行为概率分布刻画签到行为的相似性.本文采用余弦距离计算用户之间的签到行为的差别,余弦距离具体定义如下:

$$sim(x,y)=\cos(x,y)=\begin{cases}\dfrac{\sum_{s\in S_{xy}} r_{x,s}\cdot r_{y,s}}{\sqrt{\sum_{s\in S_{xy}} r_{x,s}^2 \sum_{s\in S_{xy}} r_{y,s}^2}}, & |S_{xy}|\geq 2\\ 0, & |S_{xy}|<2\end{cases} \tag{4}$$

其中,$sim(x,y)$表示用户 x 和用户 y 之间的相似度,$r_{x,s}$ 表示用户在第 s 个时间槽的签到概率,$S_{x,y}$ 表示用户 x 和用户 y 在 24 个时间槽的签到概率集合.

用余弦距离来计算用户 u_a 和 u_b 的签到行为相似度,本文通过对 Foursquare 和 Gowalla 两个数据集中的 10 000 个用户的签到行为做统计分析发现:用户的签到在时间上也具有聚集性,并且集中程度大部分都在 0.8.从图 3 可知:两个数据集中签到行为在时间上的聚集程度达到 80%的用户,其所占数据集全体用户的比率最大.由于本文利用用户签到行为的概率分布来比较用户之间的签到行为的相似性,如果两个用户的签到行为概率分布有 80%以上相似,则本文视为两个用户具有相似的签到行为,所以本文设定相似性的阈值为 0.8.

$$sim(u_a,u_b)=\cos(u_a,u_b)=\frac{\sum_{s\in S_{u_au_b}} p_{u_a,s}\cdot p_{u_b,s}}{\sqrt{\sum_{s\in S_{u_au_b}} p_{u_a,s}^2 \sum_{s\in S_{u_au_b}} p_{u_b,s}^2}}=0.8428>\varepsilon=0.8,$$

其中,$|S_{u_au_b}| \geq 2.S_{u_au_b}$ 如公式(4)中的 $S_{x,y}$,表示用户 u_a 和用户 u_b 在 24 个时间槽的签到概率集合.本文定义两个用户之间签到分布之间的余弦距离超过 0.8,就视两个用户具有相似的签到行为,所以用户 u_a 和 u_b 具有相似的签到行为.

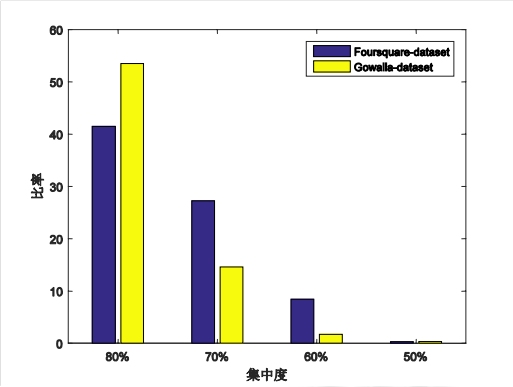


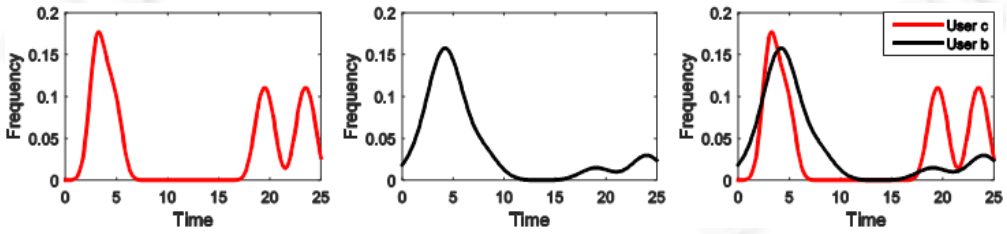
Fig.3 Setting similarity threshold

图 3 相似性阈值的设定

1.5.2 信息熵

直接用余弦距离比较用户之间的签到行为会存在一些问题.

余弦距离比较的是用户之间整体的概率分布的差别,比如用户 u_b 和 u_c 的签到行为概率分布可知,整体的比较余弦相似性是 0.7797<0.8.如图 4 所示,由于在前 10 个时间槽中的分布相似,只是两个用户概率分布的右侧尾部的分布趋势不同,从而降低了两个用户的签到行为的相似度.因为用户整体的签到次数是不同的,用户之间的相似的签到行为一般体现在 24 个时间槽中活动比较活跃的时间槽,所以在用户不活跃的时间槽做两个用户的签到行为相似性比较,会降低整体的签到行为相似性.所以,本文采用信息熵凸显出用户在活跃时间槽的行为特征,这样就降低了不活跃时间槽对整体相似性的负效应.



(a) u_c 的签到概率密度 (b) u_b 的签到概率密度 (c) u_a 和 u_b 的签到概率密度比较

Fig.4 Probability-Based check-in behavior probability estimation

图 4 基于概率的签到行为概率估计

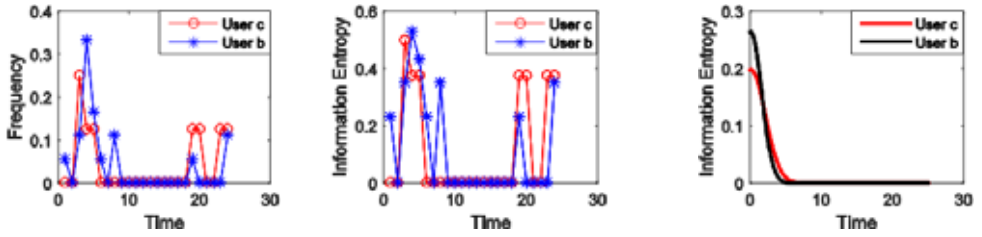
本文用基于信息熵,由于用户的签到行为在一天 24 个时间槽中是集中在用户比较活跃的时间槽,所以 24 个时间槽对于每个用户都有活跃和基本上不发生签到行为的时段.因此,本文针对于每个用户,为 24 个时间槽中每个时间槽做单位段的信息度量,如公式(5)所示.

$$Time_i=-p_i\log(p_i) \tag{5}$$

其中,p_i 表示用户在第 i 个时间槽签到行为发生的概率.

首先,用信息熵对用户在 24 个时间槽中的离散频率做计算,结果如图 5(b)所示;然后对离散的信息熵折线图用核密度估计做平滑的概率密度估计,结果如图 5(c)所示;最后,用户 u_b 和 u_c 平滑后的信息熵分布做余弦相似性比较,$sim(u_a,u_b)$=0.9811>0.8>0.7797.从得出的数值来看,对用户的签到概率做信息熵的计算后,可以明显提高用

户在活跃时间槽所占的签到信息的比重,从而提高了比较用户之间签到行为相似性的准确度.



(a) u_c 和 u_b 的签到频率 (b) u_c 和 u_b 的签到频率信息熵 (c) u_c 和 u_b 经过 KDE 的频率信息熵曲线

Fig.5 Information-Entropy-Based check-in behavior probability estimation

图 5 基于信息熵的签到行为概率估计

1.6 考虑信息获取的候选推荐用户查找

基于候选推荐用户与目标用户具有相似的签到行为,而本文的目标是为目标用推荐具有相似签到行为,并且可以为目标用户带来更多新的地理信息的陌生人.所以,本节我们将在候选推荐用户中选择与目标用户具有一定数量共同的签到过的地理位置,同时具有尽量多的目标用户之前没有签到过的地理位置.本文选择具有以上特征的候选推荐用户作为给目标用户推荐的陌生人.

- 问题定义

已知目标用户 Target User 为 u_T,候选推荐用户 Candidate User 集合为 U_C,Target User 的 Check-in Location ID 集合记为 S_T,Candidate User 的 Check-in Location ID 集合记为 S_C.

求:最终被推荐用户 Recommendation Candidate User,记为 u_{RC},使得:

$$u_{RC}=\arg\max_{p\in U_C} I\{|S_T\cap S_p|-\lambda>0\}\cdot(|S_T-S_p|) \tag{6}$$

λ是设定的已知阈值,$|S_T\cap S_p|$是目标用户与候选推荐用户具有的公共活动 Location ID 的数量.候选推荐用户与目标用户具有超过阈值的活动范围后,才具有与目标用户具有在地理空间活动较高的相似性.$|S_T\cap S_p|$是目标用户没有 Check-in 过的 Location ID 数量.公式(6)是当候选推荐用户与目标具有在地理空间上有较高活动相似性的基础上,寻找可以给目标用户带来更大没有访问过 Check-in 的 Location 范围的候选推荐用户.算法的伪代码描述如下.

算法 **1**. 考虑获取空间信息的候选推荐用户查找.

1. 输入:目标用户 ID u_T,候选推荐用户集合 U_C,u_T 和 U_C 签到过的 Location ID 集合 S_T 和 S_C.
2. 输出:最终推荐的用户集合 $U_{RC}=\{u_{RC_1},u_{RC_2},...,u_{RC_N}\}$,$U_{RC}\subset U_C$.
3. **For each** $n\in[1,2,\ldots,length(U_C)]$:
4. $M=length(S_T\cap S_{p_n})$; /$u_T$ 和 U_C 具有的公共活动 Location ID 的数量/
5. **If** $M>\lambda$ /公共活动数量 M 超过预设阈值λ,说明具有空间相似性/
6. $U_{CC}\leftarrow p_n$; /U_{CC} 是 u_T 具有空间性的候选推荐用户 p_n 的集合/
7. **End**
8. $N_1=length(U_{cc}),N_2=length(S_{p_n}-S_T)$; /$N_1$ 是与 u_T 具有空间性的候选推荐用户 U_C 的人数; N_2 是 S_{p_n} 中 u_T 没有签到过 Location ID 的数量/
9. **If** $N_2==\max\{(S_{u_{cc_1}}-S_T),(S_{u_{cc_2}}-S_T),....,(S_{u_{cc_{N1}}}-S_T)\}$ /若 N_2 是 U_{CC} 的签到位置集合中 u_T 没有签到过 Location ID 的数量中最大的/
10. $U_{RC}\leftarrow p_n$; /满足第 9 行的 p_n 就是最终推荐的用户/
11. **End**

12. **End**

基于签到行为相似性、空间活动范围相似性及地理信息扩大程度的度量,本文综合以上 3 个指标对满足以上要求的候选推荐用户排序,选择前 N 个陌生人作为最终推荐的陌生人.综合指标(comprehensive similarity,简称 CS)公式如下:

$$CS=w_1\cdot Similarity_{checkin}+w_2\cdot Similarity_{spacial}+w_1\cdot Similarity_{LocationInfor} \tag{7}$$

2 实验结果与分析

本节首先介绍实验所用数据集,然后说明评价标准及对比算法,最后给出本文提出的算法与其他方法的对比实验结果,并对实验结果进行了相应的分析.

2.1 实验数据集

本文选择社交网络上最流行的 Foursquare 签到数据集作为研究对象,数据集记录了 2010.4.14~2011.1.17 在 36 907 个 Location 共 1 048 575 次签到记录.每次签到记录由用户的 ID、签到位置的 ID 及经纬度、签到时间组成;朋友关系数据集记录了 36 907 个用户及这些用户的 231 148 对朋友关系.在 36 907 个 Location 中,平均每个 Location 有 28.411 3 次签到.

图 6 表示 Foursquare 数据集中用户签到的分布.虽然从数据集上不能直接得到用户基于一天中不同时间槽的签到行为偏好的信息,但是这些数据集蕴含用户基于位置的偏好信息及好友关系信息.因此,首先必须对 9 个月内数据集中所有用户签到行为信息进行统计学习,明确用户随着时间的签到轨迹及用户的签到行为信息,挖掘用户之间本身固有的签到行为的相似性,为用户推荐具有一定相似性的签到行为,并且可以为目标用户带来更多新的地理信息的陌生人,从而满足目标用户通过推荐的朋友获取新的地理信息的需求.

Fig.6 Check-In distribution of Foursquare dataset

图 6 Foursquare 数据集中的用户签到分布

2.2 目标用户的陌生人推荐

本节我们随机选择一个用户 ID 为 8880 作为目标用户,目标用户的陌生人数量为 8 947.通过签到行为相似性比较,与目标用户具有的相似性超过 0.8 的陌生人有 5 925 个,如图 7 所示为与目标用户 8880 签到行为相似性比较高的随机 5 个陌生人的频率信息熵曲线.

然后将 5 952 个满足相似签到行为要求的陌生人作为进一步做推荐的用户集合 U_C,在集合 U_C 中,选择与目标用户 u_T 具有超过阈值λ个数的共同签到位置数量的用户组成空间相似性候选推荐用户集合 U_{CC}.本文根据具体的实验得出经验值并设定为λ=2,在实验结果与分析中进行详细的说明.目标用户获得的集合 U_{CC} 包含 768 个

候选推荐用户满足空间相似性.基于之前与目标用户 u_T 具有相似的签到行为和空间相似性,最后在集合 U_{CC} 中寻找 Top N 个可以给目标用户 u_T 带来更多新的 Location 信息的用户作为最终推荐的陌生人集合 U_{RC}.本文选择 Top 10 个最终推荐的陌生人推荐给目标用户 8880,Top 10 个最终推荐的陌生人具体信息见表 2,他们与目标用户签到行为、空间相似程度、为目标用户带来的新的地理信息数量、综合指标这 5 方面的排序如图 8 所示.

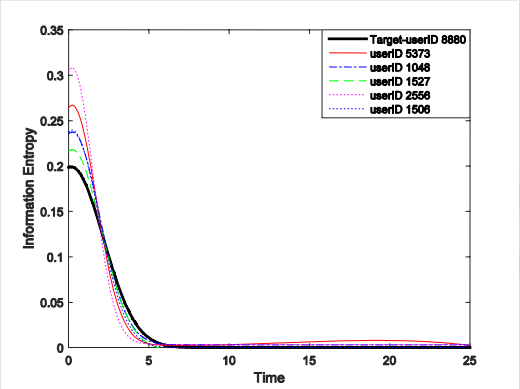


Fig.7 Information entropy based on check-in frequency of strangers

图 7 陌生人的签到频率信息熵曲线

Table 2 Recommendation users for target user 8880

表 **2** 为目标用户 8880 推荐的用户

Rank	User ID	Comprehensive similarity	Rank	User ID	Comprehensive similarity
1	404	0.913 714 799 226 866	6	4 471	0.832 590 243 368 972
2	5 194	0.861 375 939 405 830	7	5 975	0.832 319 707 987 596
3	588	0.859 998 075 959 129	8	0	0.832 266 836 621 552
4	3 812	0.841 555 411 455 260	9	111	0.830 231 210 394 441
5	5 728	0.841 555 411 455 260	10	4 739	0.829 620 119 704 323

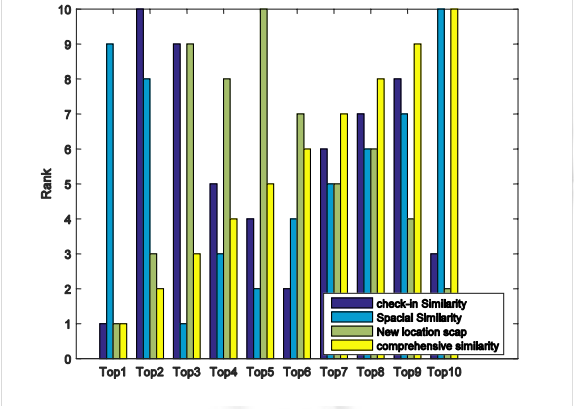


Fig.8 Rank of top 10 similarity

图 8 Top 10 相似性的排序

本文提出的推荐算法目标是为目标用户推荐可以提供更多新的地理信息的陌生人,所以为了刻画推荐的陌生人为目标用带来更多新的地理信息,本文用目标用户固有的签到地理位置和推荐的陌生人带来的新的地理位置的经纬度坐标图,通过经纬度坐标散点图可视化地理信息量的增加,如图 9 所示.

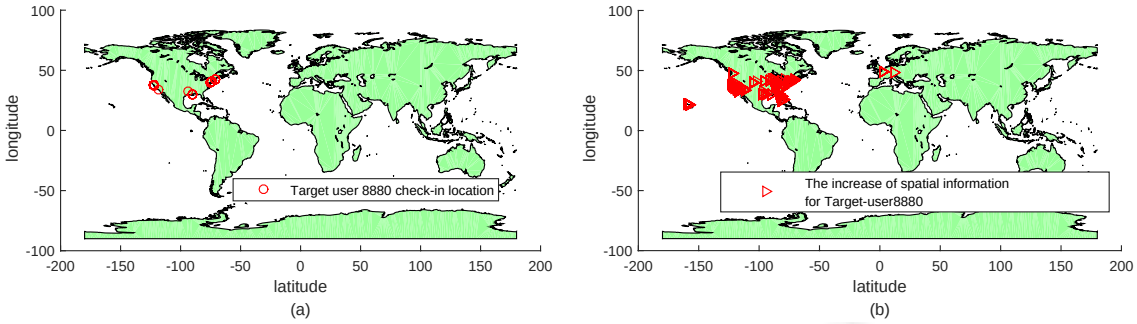


Fig.9 Latitude and longitude coordinates a scatter diagram

图 9 经纬度坐标散点图

2.3 评价指标

2.3.1 推荐精确度的度量

通过分析签到数据,我们发现朋友之间的签到行为要比不是朋友的更具有相似性,因为通常朋友们在生活中经常一起参加活动和共同爱好.因此,一种较好的推荐算法应该是对朋友之间的相似性导向要比陌生人之间的导向更高.本文从目标用户的朋友中移除一些朋友,假设移除的朋友为目标用户的陌生人.然后,查看通过本文提出的推荐算法,为目标用户最终推荐的陌生人中是否有这些假设的陌生人.如果为目标用户最终推荐的 N 个陌生人中包含的假设的陌生人较多,就说明本文提出的推荐算法的有效性.基于以上的考虑,本文采用文献[21]中的指标来评估本文提出的推荐的准确性.

$$Q_1 = \sum_{i=1}^{N} N_{i,top_n} / nN \tag{8}$$

其中,N 是实验次数, N_{i,top_n} 是目标用户 u_i 的前 n 个推荐的陌生人中真正朋友的数量,Q_1 反映了推荐的精确度.

2.3.2 获取信息量的度量

朋友推荐的本质目标是为目标用户提供新的信息,本文基于满足目标用户,通过朋友推荐获取更多新的信息的要求为目标,为目标用户提供满足具有一定的签到行为相似性的陌生人,同时可以为目标用户带来更多新的地理信息的陌生人作为系统最终为目标用户推荐的陌生人.基于以上的考虑,我们提出一个评价推荐算法为目标用户带来空间信息量的指标:

$$R = \sum_{i=1}^{N} M_{i,top_n} / N \tag{9}$$

其中,N 是实验次数; M_{i,top_n} 是目标用户 u_i 的推荐列表中的前 n 个推荐的陌生人提供新的 Location 信息的数量;R 反映了推荐带来新的 Location 信息度,R 越大,说明本文提出的方法给目标用户带来的新的 Location 信息越多.

2.3.3 实验结果及分析

本文随机选择 200 个用户作为目标用户,N=200.本文做 200 次推荐,每次为 200 个目标用户中的 1 个用户做推荐,并且记录每次实验的 N_{i,top_n} 和 M_{i,top_n}.然后,本文将提出的新的推荐算法记作 Rec_{new},并与存在的同样是基于签到数据做朋友推荐的算法做比较,目前存在的算法有基于时空相关的即时陌生人推荐算法[21],记作 Rec_{st}.然而算法 Rec_{st} 侧重的是为目标用户推荐即时所在空间的小范围内的陌生人,同时具有与目标用户相似的签到行为的陌生人.Rec_{st} 中的签到行为是针对目标用户即时空间范围内的地理位置做时空概率估计,基于这种在小范围内的时空上的签到行为相似性,为目标用户推荐相似性较高的前 N 个陌生人.然而,这种推荐算法只是考虑到目标用户推荐具有在空间上小范围内签到行为相似的陌生人,并没有考虑推荐的本质目标是为目标用户提供新的信息,也没有满足目标用户通过朋友推荐获取更多新的地理信息的要求.

本文首先基于与目标用户具有签到行为的相似性,然后选择在空间上有超过一定阈值λ个数的共同签到位置数量的用户组成空间相似候选推荐用户集合.通过对阈值λ的取不同的值,采用本文提出的潜在好友推荐算

法为 200 个目标用户推荐好友,在精度和获取的空间信息量上做了实验对比.实验结果如图 10 所示,本文选取 λ=2 作为设定的阈值,在精度和获取的空间信息量上好于其他取值.

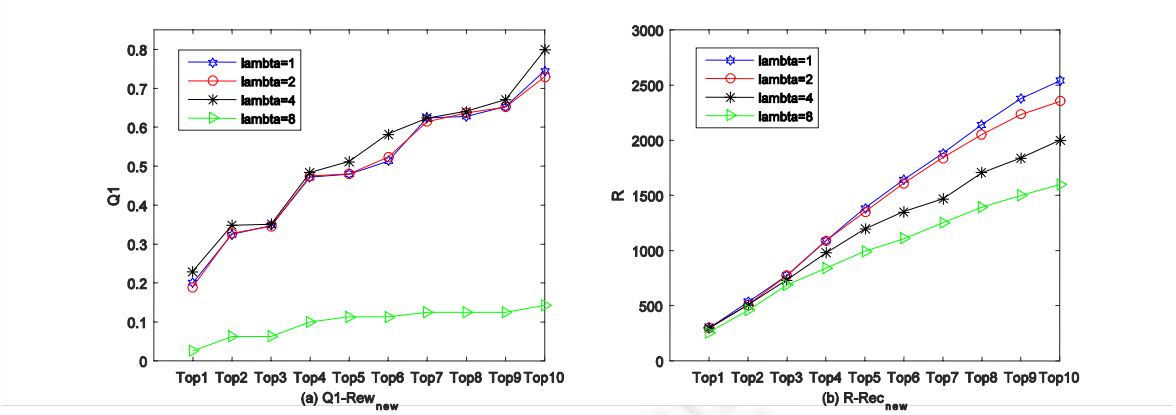


Fig.10 Evaluation index with the different values of λ

图 10 不同取值的λ的评估指标

与此同时,本节采用 Rec_{new},Rec_{st} 算法及文献[21]中提到的基于余弦相似性的推荐算法 Rec_{cos}、基于频率相似性的推荐算法 Rec_{freq} 为 200 个目标用户推荐陌生人,并比较推荐结果的精度和信息量扩大程度.从图 11 可知,Rec_{new} 在推荐精度上与 Rec_{st} 算法相似,并高于另外两种算法.同时,本文通过加入对签到数据的时间熵对 Rec_{st} 进行改进,并且对改进的算法 Rec_{est} 与其他 4 种算法在精度和信息获取量上做比较.改进的算法 Rec_{est} 在信息获取量上略高于 Rec_{st},但是在精度上明显高于其他几种算法,如图 11 和图 12 所示.

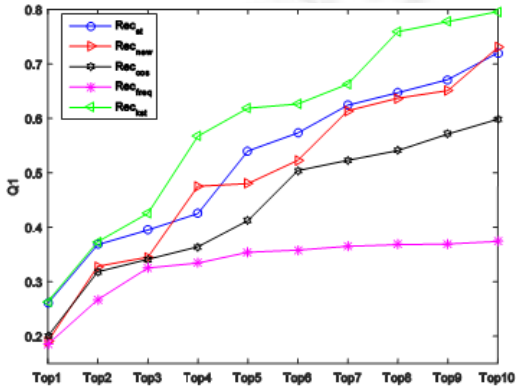


Fig.11 Q_1 of the four methods

图 11 4 种方法的 Q_1

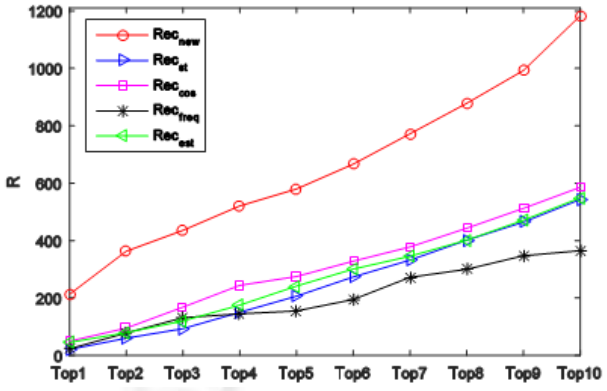


Fig.12 R of Rec_{new} and Rec_{st}

图 12 Rec_{new} 和 Rec_{st} 的 R

从图 12 可知,本文采用 Rec_{new} 和 Rec_{st} 算法为 200 个目标用户推荐陌生人,平均每个目标用户的前 10 个推荐的陌生人带来新的地理信息的数量,Rec_{new} 算法得到的 R 值明显高于其他 3 种算法.说明本文提出的推荐算法不仅可以保证在精度上不下降,而且给目标用户提供更多新的地理信息,证明了本文提出的推荐算法的有效性.由于本文主要研究的是考虑用户获取信息的潜在好友推荐算法,对于推荐系统中不可避免的冷启动问题,本文采取的方法是没有任何访问记录的新用户,开始可以采用根据用户所在的地理位置,基于附近用户中被推荐的次数,选择 top-k 用户推荐给新用户.随着新用户的访问记录的增加,但是访问记录信息不足,则可以通过采用推荐与新用户的朋友访问过的位置相似或者具有共同好友的潜在好友[1,24].直到当新用户的访问记录增加到可以采用本文提出的推荐算法时,才采用本文提出的算法.

3 结论及进一步工作

传统的朋友推荐算法都是基于目标用户熟悉的社交圈子进行推荐的,推荐的人大都是用户熟知的人或者是目标用户的朋友.推荐根本的目标是为用户提供新的信息,所以本文考虑满足用户通过朋友获取更多新的地理信息的需求,提出利用用户在真实世界中的签到行为的相似性,为用户推荐可以带来更多新的地理信息的陌生人.实验结果表明:该推荐算法在精度上保证了目前较好的基于签到数据的陌生人推荐算法的相似性;同时,在信息扩大程度上具有显著的优势.在下一步的研究工作中,我们将挖掘移动用户在 Foursquare 上的伴随签到行为的语言信息,研究移动用户位置预测方法以及基于签到行为的异地朋友推荐算法.

References:

[1] Xu B, Chin A, Wang H. Using physical context in a mobile social networking application for improving friend recommendations. In: Proc. of the 2011 Int'l Conf. on Internet of Things, and 4th Int'l Conf. on Cyber, Physical and Social Computing. IEEE, 2011. 602−609. [doi: 10.1109/iThings/CPSCom.2011.76]

[2] Liu S, Cao H, Li L, Zhou MC. Predicting stay time of mobile users with contextual information. IEEE Trans. on Automation Science and Engineering, 2013,10(4):1026−1036. [doi: 10.1109/TASE.2013.2259480]

[3] Reilly J, Dashti S, Ervasti M, Bray JD, Glaser SD, Bayen AM. Mobile phones as seismologic sensors: Automating data extraction for the iShake system. IEEE Trans. on Automation Science and Engineering, 2013,10(2):242−251. [doi: 10.1109/TASE.2013.2245121]

[4] Santos AC, Cardoso JMP, Ferreira DR, Diniz PC, Chaínho P. Providing user context for mobile and social networking applications. Pervasive and Mobile Computing, 2010,6(3):324−341. [doi: 10.1016/j.pmcj.2010.01.001]

[5] Amir A, Efrat A, Myllymaki J, Palaniappan L, Wampler K. Buddy tracking—Efficient proximity detection among mobile friends. Pervasive and Mobile Computing, 2007,3(5):489−511. [doi: 10.1016/j.pmcj.2006.12.002]

[6] Wang H, Chin A, Wang H. Interplay between social selection and social influence on physical proximity in friendship formation. In: Proc. of the SRS 2011 Workshop. 2011. 1−8.

[7] Bellavista P, Montanari R, Das SK. Mobile social networking middleware: A survey. Pervasive and Mobile Computing, 2013,9(4): 437−453. [doi: 10.1016/j.pmcj.2013.03.001]

[8] Conti M, Das SK, Bisdikian C, Kumar M, Ni LM, Passarella A, Roussos G, Tröster G, Tsudik G, Zambonelli F. Looking ahead in pervasive computing: Challenges and opportunities in the era of cyber—Physical convergence. Pervasive and Mobile Computing, 2012,8(1):2−21. [doi: 10.1016/j.pmcj.2011.10.001]

[9] Kefalas P, Symeonidis P, Manolopoulos Y. New perspectives for recommendations in location-based social networks: Time, privacy and explainability. In: Proc. of the 5th Int'l Conf. on Management of Emergent Digital EcoSystems. ACM Press, 2013. 1−8. [doi: 10.1145/2536146.2536202]

[10] Xu HL, Wu X, Li XD, Yan BP. Comparison study of Internet recommendation system. Ruan Jian Xue Bao/Journal of Software, 2009,20(2):350−362 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3388.htm [doi: 10.3724/SP.J.1001.2009.03388]

[11] Mahmud J, Zhou MX, Megiddo N, Nichols J, Drews C. Recommending targeted strangers from whom to solicit information on social media. In: Proc. of the 2013 Int'l Conf. on Intelligent User Interfaces. ACM Press, 2013. 37−48. [doi: 10.1145/2449396.2449403]

[12] Granovetter MS. The strength of weak ties. American Journal of Sociology, 1973,78(6):1360−1380. [doi: 10.1086/225469]

[13] Guy I, Ur S, Ronen I, Perer A, Jacovi M. Do you want to know? Recommending strangers in the enterprise. In: Proc. of the ACM 2011 Conf. on Computer Supported Cooperative Work. ACM Press, 2011. 285−294. [doi: 10.1145/1958824.1958867]

[14] Cho E, Myers SA, Leskovec J. Friendship and mobility: User movement in location-based social networks. In: Proc. of the 17th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2011. 1082−1090. [doi: 10.1145/2020408.2020579]

[15] Yin H, Sun Y, Cui B, Hu Z, Chen L. Lcars: A location-content-aware recommender system. In: Proc. of the 19th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2013. 221−229. [doi: 10.1145/2487575.2487608]

[16] Gonzalez MC, Hidalgo CA, Barabasi AL. Understanding individual human mobility patterns. Nature, 2008,453(7196):779−782. [doi: 10.1038/nature06958]

[17] Zhang Y, Wang L, Zhang YQ, Li X. Towards a temporal network analysis of interactive WiFi users. EPL (Europhysics Letters), 2012,98(6):68002. [doi:10.1209/0295-5075/98/68002]

[18] Zhang YQ, Li X. Temporal dynamics and impact of event interactions in cyber-social populations. Chaos: An Interdisciplinary Journal of Nonlinear Science, 2013,23(1):907-R. [doi: 10.1063/1.4793540]

[19] Froehlich J, Chen MY, Smith IE, Potter F. Voting with your feet: An investigative study of the relationship between place visit behavior and preference. In: Proc. of the Int'l Conf. on Ubiquitous Computing. Berlin, Heidelberg: Springer-Verlag, 2006. 333–350. [doi: 10.1007/11853565_20]

[20] Tobler WR. A computer movie simulating urban growth in the Detroit region. Economic Geography, 1970,46:234–240. [doi: 10.2307/143141]

[21] Qiao X, Yu W, Zhang J, Tan W, Su J, Xu W, Chen J. Recommending nearby strangers instantly based on similar check-in behaviors. IEEE Trans. on Automation Science and Engineering, 2015,12(3):1114–1124. [doi: 10.1109/TASE.2014.2369429]

[22] Bian L, Holtzman H. Online friend recommendation through personality matching and collaborative filtering. In: Proc. of the UBICOMM. 2011. 230–235.

[23] Yu X, Pan A, Tang LA, Li Z, Han J. Geo-Friends recommendation in gps-based cyber-physical social network. In: Proc. of the 2011 Int'l Conf. on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, 2011. 361–368. [doi: 10.1109/ASONAM.2011.118]

[24] Chu CH, Wu WC, Wang CC, Chen TS, Chen JJ. Friend recommendation for location-based mobile social networks. In: Proc. of 2013 the 7th Int'l Conf. on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS). IEEE, 2013. 365–370. [doi: 10.1109/IMIS. 2013.68]

[25] Silva NB, Tsang IR, Cavalcanti GDC, Tsang IJ. A graph-based friend recommendation system using genetic algorithm. In: Proc. of the 2010 IEEE Congress on Evolutionary Computation (CEC). IEEE, 2010. 1–7. [doi: 10.1109/CEC.2010.5586144]

[26] Narayanam R, Narahari Y. A shapley value-based approach to discover influential nodes in social networks. IEEE Trans. on Automation Science and Engineering, 2011,8(1):130–147. [doi: 10.1109/TASE.2010.2052042]

[27] Adamic LA, Adar E. Friends and neighbors on the Web. Social Networks, 2003,25(3):211–230. [doi: 10.1016/S0378-8733(03)00009-1]

[28] Debnath S, Ganguly N, Mitra P. Feature weighting in content based recommendation system using social network analysis. In: Proc. of the 17th Int'l Conf. on World Wide Web. ACM Press, 2008. 1041–1042. [doi: 10.1145/1367497.1367646]

[29] Zheng Y, Zhang L, Ma Z, Xie X, Ma WY. Recommending friends and locations based on individual location history. ACM Trans. on the Web (TWEB), 2011,5(1):5. [doi: 10.1145/1921591.1921596]

[30] Xiao X, Zheng Y, Luo Q, Xie X. Finding similar users using category-based location history. In: Proc. of the 18th SIGSPATIAL Int'l Conf. on Advances in Geographic Information Systems. ACM Press, 2010. 442–445. [doi: 10.1145/1869790.1869857]

[31] Braga RB, Tahir A, Bertolotto M, Martin H. Clustering user trajectories to find patterns for social interaction applications. In: Proc. of the In'l Symp. on Web and Wireless Geographical Information Systems. Berlin, Heidelberg: Springer-Verlag, 2012. 82–97. [doi: 10.1007/978-3-642-29247-7_8]

[32] Zheng Y, Zhang L, Xie X, Ma WY. Mining interesting locations and travel sequences from GPS trajectories. In: Proc. of the 18th Int'l Conf. on World Wide Web. ACM Press, 2009. 791–800. [doi: 10.1145/1526709.1526816]

附中文参考文献:

[10] 许海玲,吴潇,李晓东,阎保平.互联网推荐系统比较研究.软件学报,2009,20(2):350−362. http://www.jos.org.cn/1000-9825/3388.htm [doi: 10.3724/SP.J.1001.2009.03388]

俞菲(1989－),女,黑龙江哈尔滨人,博士生,CCF 学生会员,主要研究领域为地理位置数据挖掘,推荐系统.

车楠(1980－),男,博士,副教授,CCF 专业会员,主要研究领域为传感器网络,节点部署,网络优化.

李治军(1977－),男,博士,副教授,CCF 专业会员,主要研究领域为物联网,无线传感器,地理位置数据挖掘.

姜守旭(1968－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为普适计算,无线传感器网络,智能交通系统,物联网.