

基于移动网络位置信息的群体发现方法^{*}

刘 分, 汤红波, 葛国栋, 杨 森

(国家数字交换系统工程技术研究中心, 郑州 450002)

摘要: 当前群体发现研究主要利用通联关系挖掘用户群体, 未能充分利用网络中所隐含的用户社交关系, 致使挖掘的群体不能真实反映用户在社会生活中的群体关系。提出一种基于用户位置信息的群体发现方法, 利用序列模式挖掘算法挖掘用户位置规律序列, 建立位置序列相关性度量标准, 以位置规律相关性揭示用户社会活动的群体关系; 结合局部相似性度量方法计算用户通信距离指数, 反映用户之间的相识程度; 最后采用通信距离指数对位置相关性进行加权计算用户群体相关性, 再利用分裂聚类算法挖掘具有通信关系和社交关系的用户群体。实验结果表明, 该方法能够有效地挖掘用户社交活动中的通信相关性和位置相关性, 体现用户在现实社会活动中的群体关系。

关键词: 群体发现; 通联关系; 位置信息; 序列模式挖掘; 通信距离相似性

中图分类号: TP393.04 **文献标志码:** A **文章编号:** 1001-3695(2013)05-1471-04

doi:10.3969/j.issn.1001-3695.2013.05.048

Group discovery method in mobile communication network based on location information

LIU Fen, TANG Hong-bo, GE Guo-dong, YANG Sen

(National Digital Switching System Engineering & Technological R&D Center, Zhengzhou 450002, China)

Abstract: Currently, group discovery mainly mines group structure with certain features based on communications relationship, which can't take full advantage of the implicit social relations in networks and can't truly represent the group relationship in the real social life. This paper proposed a group discovery method based on location information, utilized sequence pattern mining algorithm to mine location patterns, established location sequence similarity measurement criterion, detected social activity group relations with location pattern similarity. Then it utilized local similarity method to calculate communication similarity, reflected the degree of acquaintance between users. Finally, calculating group relations based on location pattern similarity weighted with communication similarity, it utilized the clustering algorithm to mine groups with communication relations and social relations. Experimental result indicates that the proposed method can better combines communication relations and location relations to mine group relations, embodying the group relationship of users in the reality of social activities.

Key words: group discovery; communication relations; location information; sequential pattern mining; communication distance similarity

0 引言

近年来, 群体发现技术已成为通信领域的研究热点, 众多研究机构、会议、期刊纷纷开展相关的研究。一些研究成果广泛应用于社会、经济、政治等众多领域。例如根据用户网络行为特征, 挖掘用户群体关系, 打击恐怖主义、犯罪组织, 维护社会稳定; 了解客户需求、公司业务与市场前景之间相互作用的关系, 为公司发展提供参考; 根据政府议员关系进行模拟选举, 预测选举结果^[1,2]。

当前群体发现的研究方法大致可分为三类: 基于图分割的方法、层次聚类方法^[3]、基于谱思想的方法。此外, 还有许多基于经典算法的改进方法, 如文献[4]提出了一种自适应动态社区发现方法, 依据节点和边的增加与删除, 快速更新网络的

群体划分。然而, 这些群体发现的研究方法主要是利用用户的网络连接、属性或行为特征^[5]进行分析, 发现具有相似兴趣、爱好、背景或价值观等网络特征的用户群体, 不能够反映用户在社会生活中社会活动的群体关系。现实社会生活中的群体往往是联系密切、经常参与相同社会活动的用户, 如亲人、朋友、同事等。这些群体用户之间一般通过基本的通信方式进行交流沟通, 如电话、短信等, 而且这些群体用户往往在位置上表现出相似的规律或具有很高的重合率, 如可能经常处于同一小区等。

本文提出了基于移动通信网络位置信息的群体发现方法, 利用移动用户的通信关系及能够反映其日常社会活动规律的位置规律, 挖掘社会生活中联系密切、具有相同社会活动的群体。了解用户在现实生活中社会活动的群体关系, 对于掌握用户的社会行为与社会关系具有巨大的参考意义。

收稿日期: 2012-10-10; 修回日期: 2012-11-25 基金项目: 国家“863”计划资助项目(2011AA010604)

作者简介: 刘分(1986-), 男, 河南商丘人, 硕士研究生, 主要研究方向为移动通信网络(wohucanglonglf@163.com); 汤红波(1968-), 男, 教授, 主要研究方向为移动通信网络; 葛国栋(1985-), 男, 博士研究生, 主要研究方向为可移动网络、移动通信网络; 杨森(1985-), 男, 助教, 硕士, 主要研究方向为移动通信网络。

1 位置规律挖掘

位置信息与用户真实的社会活动息息相关,用户的位置规律记录了用户在真实世界的社会活动轨迹,能够反映用户的生活位置规律及相关社会属性^[6]。现实社会中,经常一起参与相同社会活动的用户往往具有某种密切的群体关系,如一起上班的同事等。这些用户在社会生活中的密切群体关系不仅体现在通信连接的密切性方面,还体现在位置规律的相关性方面。

本文利用序列模式挖掘的方法^[7]从位置序列数据库中挖掘频繁子序列作为用户的位置规律。基于经典序列模式挖掘 PrefixSpan 算法^[8],提出位置规律序列挖掘算法(location-based sequence patterns mining, LSP),挖掘用户的频繁位置序列。

1.1 问题描述

将用户位置信息按时间顺序看做位置序列,建立用户位置序列数据库。采用序列模式挖掘的方法从用户位置序列数据库中挖掘频繁位置序列作为用户位置规律。序列模式挖掘方法中基本概念描述如下:

设位置序列 $S = \langle s_1 s_2 s_3 \dots s_n \rangle$, 其中 $s_j (1 \leq j \leq n)$ 为序列 S 的第 j 项位置集合, 称为项集。对于序列 $A = \langle a_1 a_2 a_3 \dots a_m \rangle$ 和序列 $B = \langle b_1 b_2 b_3 \dots b_n \rangle (0 \leq i \leq n - m)$, 当 $a_{i+1} \subseteq b_{i+1}$, $a_{i+2} \subseteq b_{i+2}$, $a_{i+3} \subseteq b_{i+3}$, $\dots$, $a_{i+m} \subseteq b_{i+m}$ 时, 称序列 A 是序列 B 的子序列。

定义 1 序列 A 的支持度是序列数据库中包含序列 A 的序列个数。序列 A 支持度为 3, 表示为 $A(3)$ 。

定义 2 对于序列 $A = \langle a_1 a_2 a_3 \dots a_m \rangle$, $B = \langle b_1 b_2 b_3 \dots b_n \rangle (m \leq n)$, 若序列 A 的前 $m-1$ 项和序列 B 的前 $m-1$ 项相同, 且 a_m 与 b_m 中的前面一部分或全部相同, 则称序列 A 是序列 B 的前缀。

定义 3 对于序列 A 和序列 A' 都是序列 B 的子序列, 当序列 A 是序列 A' 的前缀, 且序列 A' 是满足此条件的最大子序列, 则称序列 A' 是序列 A 在序列 B 上的投影。

定义 4 对于序列 A' 是序列 A 在序列 B 上的投影, 序列 A' 中去除前缀序列 A 的序列称为序列 A 在序列 B 上的后缀。

1.2 位置规律挖掘算法

位置信息在时间上的相关性体现用户社会活动的相关性, 位置序列需从位置和时间两个角度考虑位置的频繁性。与 PrefixSpan 算法相比, 位置项集、序列的支持度为对应时间上包含位置项集、序列的个数, 不同时间上的相同位置序列为不同的位置序列。

LSP 算法主要流程: a) 扫描用户序列库, 合并重复序列, 找到频繁项集, 按照支持度大小排列; b) 产生每个候选序列的投影数据库, 合并重复投影序列; c) 递归构建频繁序列集、投影数据库, 最终产生定长频繁序列。例如某用户位置序列数据库如表 1 所示, 位置序列首个项集时刻相同。

表 1 用户位置序列库

序列编号	序列
1	$\langle (ad)(ac) cdbdc \rangle$
2	$\langle (bac) cc(fa) afa \rangle$
3	$\langle (ab) acdbd(cd) \rangle$
4	$\langle aacdddc \rangle$

应用 LSP 算法处理表 1 中用户序列数据, 设从首项开始提

取用户位置规律, 序列最小支持度为 2, 流程中仅以第一个最大频繁序列为例。

1) 扫描数据库, 产生频繁项集 $a(4)$ 、 $b(2)$ 。建立 a 投影数据库, $_ (ac) cdbdc$ 、 $_ cc(fa) afa$ 、 $_ acdbd(cd)$ 、 $_ aacdddc$ (以 a 为例)。同理, 产生 b 投影数据库。

2) 扫描投影数据库, 产生长度为 2 的频繁序列, $aa(3)$ 、 $ac(2)$ 。建立 aa 投影数据库, $_ cdbdc$ 、 $_ cdbd(cd)$ 、 $_ cdddc$ 。同理, 产生 ac 投影数据库。

3) 扫描投影数据库, 产生长度为 3 的频繁序列, $aac(3)$ 、 $acc(2)$ 。依次建立投影数据库, $_ dbdc$ 、 $_ dbd(cd)$ 、 $_ ddc$ 。同理, 产生 acc 投影数据库。

4) 依次循环扫描投影数据库, 产生支持度最大、序列长度最长的频繁序列为 $aacdbdc$ 。

从以上流程可以看出, 位置序列从位置和时间两个角度衡量相关性。因此, 与 PrefixSpan 算法相比, 位置规律序列挖掘算法(LSP 算法)能够大大减少产生投影数据库的数量, 具有更高的运算效率, 弥补 PrefixSpan 算法的突出缺陷^[9]。

2 基于群体相似性群体发现

利用能够揭示用户社会活动群体关系的位置规律相关性和反映用户通联关系的通信距离指数计算用户群体相关性, 结合分裂算法, 实现挖掘目标用户通信关系网络中联系密切、具有相同社会活动的群体。首先, 利用移动用户通信数据构建用户通信关系网络。基于移动用户的通信关系网络计算用户通信距离指数 $\beta(v_i v_j)$, 代表用户 v_i 和 v_j 之间的通信连接的相似性。其次, 将用户位置信息按照时间顺序看做位置序列, 利用序列模式挖掘的方法提取用户的频繁位置序列作为用户的位置规律。再根据建立的位置相关性标准计算用户位置规律序列的相关性。最后结合用户通信距离指数与位置规律序列的相关性计算用户的群体相关性, 采用不同精度实现对用户群体关系的划分。具体工作原理如图 1 所示。

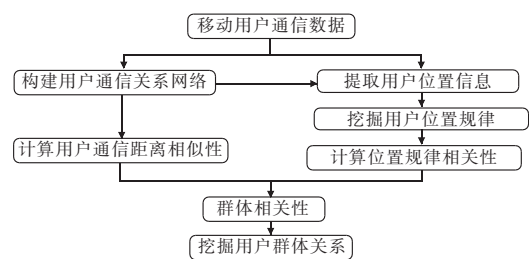


图 1 群体发现工作原理

2.1 群体相似性度量

群体相关性度量是用户通信距离相似性对用户位置相关性的加权函数。通信距离相似性是通信关系网络中用户之间通信距离的关联性, 反映用户之间的相识程度。位置相关性度量是用户在社会生活中社会活动位置的一致性, 反映用户参与社会活动的相关性。

2.1.1 通信距离相似性

根据六度分割理论, 抽取目标用户 6 跳以内的用户 (仅对本地用户继续提取下一跳用户) 作为研究对象构建用户通信关系网络, 理论上基本涵盖所有与用户相关的邻近对象。文献 [10] 提出基于局部相似性度量的方法发现用户群体, 能够高效地挖掘用户之间的群体关系。然而, 计算用户局部相似性

时,没有考虑直接连接关系和间接连接关系对用户相似性的不同影响。随着间接连接关系的增加,直接连接关系存在被过度弱化的现象。

移动通信网络直接连接关系反映了用户之间具有较近的通信距离,通信指数较高,用户之间相识的可能性很大。移动通信网络中用户的直接通信连接对用户通信连接关系相似性计算应该被赋予较高的权重,局部相似性度量的方法不能完全适用于计算移动通信网络中用户之间的连接关系相似性。本文提出通信距离相似性的方法,定义用户的相邻用户集不包括用户自身,将用户通信关系分为直接通信关系和间接通信关系,并赋予不同的影响权重。

设用户通信关系网络为 $G = (V, E)$, 其中 $V = \{v_1, v_2, v_3, \dots\}$ 为顶点集, $E = \{E_i | E_i \in V \times V, i = 1, 2, \dots\}$ 为边集合, A 为邻接矩阵。 C_v 为用户 v 的相邻用户集(不含 v), m_v 为用户 v 相邻集的用户数。群体模块度 Q 定义为

$$Q = \sum_r (e_{rr} - a_r^2) \quad (1)$$

其中: e_{rr} 表示群体 r 内部边的条数, a_r 表示一端与群体 r 中任意一个节点相连的边的条数。

基于局部相似性度量的方法,定义用户 v_i 和 v_j 的通信距离相似性的通信指数 $\beta(v_i, v_j)$ 为

$$\beta(v_i, v_j) = \alpha A_{ij} + (1 - \alpha) \frac{\sum_{v_e \in C_{v_i} \cap C_{v_j}} \frac{1}{m_e}}{\sqrt{\sum_{v_e \in C_{v_i}} \frac{1}{m_e}} \sqrt{\sum_{v_e \in C_{v_j}} \frac{1}{m_e}}} \quad (2)$$

其中:当 v_i 和 v_j 相连时, $A_{ij} = 1$, 否则 $A_{ij} = 0$ 。 α 为通信距离参数,对直接连接和间接连接关系进行加权,赋予直接连接关系以固定通信距离指数 α ,赋予间接通信连接以通信距离指数 $1 - \alpha$ 的权重。

本文中通信距离相似性仅计算通信关系网络中三跳以内用户的相关性,以两跳连接衡量间接连接的相关性。在实际应用中对通信距离相似性精度要求较高的情况,可以根据需要扩展到计算更多跳内用户通信距离相关性的计算。

2.1.2 位置规律相关性

社会活动相同的群体一般具有相似的位置特征,具有相似位置特征的群体在一定程度上能够反映参与社会活动的相关性。建立位置规律序列相关性度量标准,利用挖掘的位置规律序列,计算用户位置规律序列相关性,衡量用户社会活动相关性^[11]。

定义5 相同时刻位置相同的项集完全相关;利用泊松分布表示不同时刻位置项集之间的相关性,不同时刻用户位置相同项集时间间隔越小,位置项集相关性越大,用户社会活动的相关性越大。相同时刻位置不相同的项集相关性,为项集与其他时刻位置项集相关性的总和。

设用户 i 和 j 的位置规律序列分别为 $S_i = \langle s_{i1} s_{i2} s_{i3} \dots s_{in} \rangle$ 、 $S_j = \langle s_{j1} s_{j2} s_{j3} \dots s_{jm} \rangle$, 初始位置时刻相同。记项集 s_{im} 和 s_{jm} 的相关性为 $d(s_{im}, s_{jm})$ 。

当 $s_{im} = s_{jm} (1 \leq m \leq n)$, $d(s_{im}, s_{jm}) = 1$;

当 $s_{im} \neq s_{jm} (1 \leq m \leq n)$, $d(s_{im}, s_{jm})$ 为

$$d(s_{im}, s_{jm}) = \sum_{s_{il}=s_{jl}, l=1}^n \frac{\lambda_k}{2k!} e^{-\lambda} + \sum_{s_{il} \neq s_{jl}, l=1}^n \frac{\lambda_k}{2k!} e^{-\lambda} \quad (3)$$

其中: $k = \begin{cases} |m-l| & |m-l| \leq n/2 \\ n-|m+l| & |m-l| \geq n/2 \end{cases}$ 。位置序列是时间上的

闭合序列,项集之间的距离是闭环上的最近距离, λ 为其泊松分布参数。

最后,对序列所有项集进行相关性叠加,求得位置规律序列 S_i 和 S_j 的相关性 DS_{ij} :

$$DS_{ij} = \sum_{m=1}^n d(s_{im}, s_{jm}) \quad (4)$$

群体相关性度量式表示如下:

$$Q_{LC} = \beta^{1/k}(v_i, v_j) \cdot DS_{ij} \quad (5)$$

通信距离相似性与用户之间的连接关系、连接的数目相关,单个连接对通信相似性的影响随着连接跳数的增加而逐渐减小。通信距离相似性计算中仅利用具有相邻用户的2跳连接关系衡量间接连接通信距离指数,致使通信距离指数偏小,采用 k 次方根对通信距离指数进行处理,增大用户之间的通信距离相似性,减弱用户通信距离指数对用户群体相关性度量的影响。随着用户通信距离相似性的减小,其对用户群体相关性的影响逐渐增大,当通信距离相似性较小时,用户的群体相关性随之急剧减小。

2.2 基于分裂聚类方法的社区发现

经典的分裂聚类算法需要重复计算所有节点之间的边介数,具有计算复杂度高的缺陷。经典的分裂聚类算法采用边介数衡量边的权重,通过边的最短路径数目越多,边介数越大,连接边对相邻两个节点的权重越小。本文利用节点的群体相似性取代边介数,提出一种基于群体相似性的分裂聚类算法(QGN算法)。与边介数相比,群体相似性的计算更加简单,可以大大降低算法的计算复杂度。基于群体相似性分裂聚类算法基本思想如下:

a) 计算网络中所有节点群体相似性。

b) 移除群体相似性最小的两节点之间的连接边,更新两节点相邻节点的群体相似性。

c) 计算划分群体的模块度,重复 b),直至找到模块度最大的划分。

3 仿真与结果分析

本文实验数据来源为某市合作运营商网络数据。实验抽取某用户通信关系网络中6跳内用户(仅对本地用户继续提取下一跳用户)的通信关系与位置数据作为研究对象,位置数据按照固定时间间隔抽取,抽取时间为一个月。

通信距离相似性方法中的通信距离指数 $\beta(v_i, v_j)$ 与用户 v_i 和 v_j 之间是否具有直接连接关系、通信距离参数 α 、共同相邻节点(共邻节点)数量及其连接数有关。以通信距离参数 $\alpha = 0.4$ 为例,选取某特定用户通信关系网络中具有直接通信关系、不同非共邻节点比例的用户,计算用户与特定用户之间的距离相似性和局部相似性。具有直接连接关系用户的相似性指数随非共邻节点比例的变化关系如图2所示。

可以看出,局部相似性方法的相邻用户相关性指数随用户非共邻节点比例的增加不断减小,直接连接的相关性被过度弱化,具有直接通信关系的用户相关性因通信用户数量的增加而被减弱。距离相似性的方法通过区分直接连接或间接连接,加权均衡两者对相关性指数的影响比重,使间接连接对用户相似性的影响被限定在一定范围之内,有效地减弱了非共邻节点连接边对用户相似性指数的影响,解决了直接连接关系被过度弱

化的问题。

当 $\eta=0$ 时,用户之间所有连接都是为共邻节点连接或直接连接关系,所有连接均可达,用户相似性指数为 1,实验中只计算两跳用户连接,忽略多跳连接,故存在一定误差。当 $\eta=1$ 时,用户之间所有连接都是为非共邻节点连接,所有连接均不可达,直接连接关系所占比例趋于零,局部相似性趋于零,通信距离相似性指数为 $\alpha(0.4)$ 。

图 3 是位置序列中不同时刻位置项集之间的相关性随着项集之间距离变化的分布函数,不同时刻位置项集相关性分布函数随着项集距离增大趋于 1。实验提取序列长度为 24 的位置序列,根据不同时刻位置项集的相关性,计算不同项集距离的位置项集相似性的分布。当 $k=0$ 时,表示相邻时刻位置项集之间的相关性。以分布函数参数 $\lambda=0.8$ 为例,使得随着项集距离的增大,位置项集的相关性逐渐减小,项集相关性分布函数的增加趋于平缓。当 $n=12$ 时,位置项集的相关性趋于 0,项集相关性分布函数趋于 1。

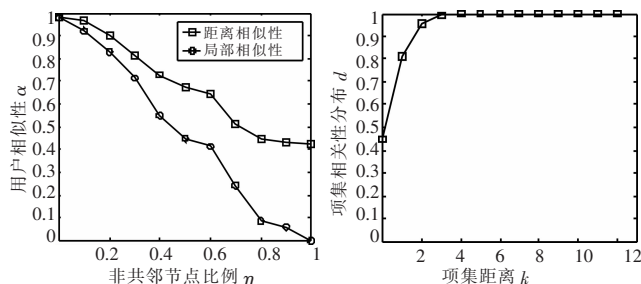


图2 通信相似性度量

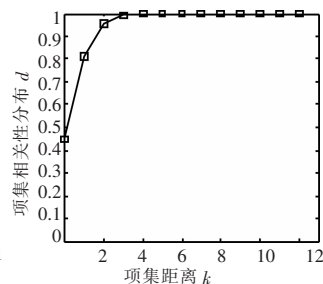


图3 位置相关性度量

实验抽取特定用户 v_0 及其通信关系网络中用户 $v_1 \sim v_9$ 的通信数据与位置数据,计算特定用户 v_0 与用户 $v_1 \sim v_9$ 的通信距离相关性、位置序列相关性和群体相关性度量。群体相关性度量中以 $k=2$ 为例,一组实验数据计算结果如表 2 所示。

表2 用户群体相关性度量

用户	v_0v_0	v_0v_1	v_0v_2	v_0v_3	v_0v_4	v_0v_5	v_0v_6	v_0v_7	v_0v_8	v_0v_9
β	1	0.5106	0.4581	0.4177	0.6372	0.0125	0.4156	0.3286	0.2736	0.4128
DS_{ij}	1	0.6915	0.0190	0.3047	0.3424	0.4011	0	0.1017	0.5109	0.0022
Q_{ic}	1	0.4941	0.0129	0.1969	0.2733	0.0448	0	0.0583	0.2673	0.0014

从表 2 可以看出,用户 $v_1 \sim v_9$ 为用户 v_0 通信关系网络中抽取用户,与用户 v_0 都具有一定的通信关系,多数通信距离相似性指数较大。然而,用户 $v_1 \sim v_9$ 中只有少数用户与用户 v_0 具有较高的位置相似性。现实社会生活中,用户通信关系网络中只有一部分用户与其具有较高的位置相关性。

由于用户位置分布时间具有一定特点,与用户在现实社会生活中的社会活动有关。因此,用户位置相关性的不同分布时刻特征,在一定程度上能够体现用户社会的不同社会活动,反映用户的不同社会活动群体关系,如普通用户同事关系,一般工作时间为 8 h,其位置相关性在 0.33 左右。当群体相关性划分尺度为 0.1 时,与用户 v_0 具有较高群体相关性的用户群体为 v_1, v_3, v_4 和 v_8 。

QGN 算法采用群体相似性取代边介数来衡量相邻两点连接边的权重,既体现用户社会活动位置规律的相关性,又体现用户之间的通信连接紧密性。针对特定用户通信关系网络数据,利用群体平均群体相关性衡量群体划分质量,比较不同算法划分效果,如图 4 所示。

可以看出,局部相似性方法和 GN 算法的群体用户平均群体相关性基本相似,与群体划分大小无关,说明用户群体相关性与通信连接相似性无明显相关性。QGN 算法划分出的群体

随着划分过程群体用户平均群体相关性不断增大,说明 QGN 算法划分的群体能够体现用户的位置相关程度。

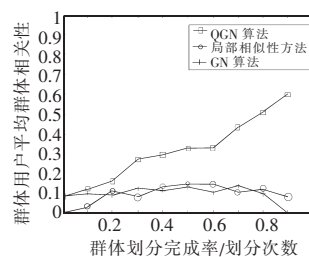


图4 群体平均位置相关性

4 结束语

本文提出了一种基于移动通信网络位置信息的群体发现方法。该方法综合考虑移动通信网络用户的通信距离相似性和位置规律相关性,挖掘社会生活中联系密切、具有相同社会活动规律的用户群体。理论分析及仿真实验表明,该方法能够较好地发现用户真实社会生活中的社会活动群体关系。然而,对于单一位置相关性或通信相关性较强的关系存在被弱化的问题,下一步将针对这个问题进行研究改进。此外,根据用户位置规律序列的相关性及其关联时刻可进一步推断用户关系。

参考文献:

- [1] PORTER M A, MUCHA P J, NEWMAN M E J, et al. A network analysis of committees in the United States House of Representatives[J]. *Proceedings of the National Academy of the United States of America (PNAS)*, 2005, 102(20): 7057-7062.
- [2] SCHNEIDER J J, HIRTREITER C. The impact of election results on themember numbers of the large parties in Bavaria and Germany[J]. *International Journal of Modern Physics C*, 2005, 16(8): 1165-1215.
- [3] CHANG Cheng-shang, HSU C Y, CHENG J, et al. A general probabilistic framework for detecting community structure in networks[C]// *Proc of IEEE INFOCOM*. 2011: 730-738.
- [4] NGUYEN N P, DINH T N, YING Xuan, et al. Adaptive algorithms for detecting community structure in dynamic social networks[C]// *Proc of IEEE INFOCOM*. 2011: 2282-2290.
- [5] ADAMIC L A, ADAR E. Friends and neighbors on the Web[J]. *Social Networks*, 2003, 25(3): 211-230.
- [6] BACCELLI F, BOLOT J. Modeling the economic value of the location data of mobile users[C]// *Proc of IEEE INFOCOM*. 2011: 1467-1475.
- [7] PEI Jian, HAN Jia-wei, MORTAZAVI-ASL B, et al. FreeSpan: frequent pattern-projected sequential pattern mining[C]// *Proc of the 6th ACM SIGKDD International Knowledge Discovery and Data Mining*. New York: ACM Press, 2000: 355-399.
- [8] PEI Jian, HAN Jia-wei, MORTAZAVI-ASL B, et al. Mining sequential patterns by pattern-growth: the PrefixSpan approach[J]. *IEEE Trans on Knowledge and Data Engineering*, 2004, 16(11): 1424-1440.
- [9] 张利军, 李战怀, 王森. 基于位置信息的序列频繁模式挖掘算法[J]. *计算机应用研究*, 2009, 26(2): 529-531.
- [10] 刘旭, 易东云. 基于局部相似性的复杂网络社区发现方法[J]. *自动化学报*, 2011, 37(12): 1520-1529.
- [11] BATYRSHIN I, HERRERA-AVELAR R, SHEREMETOV L, et al. Association networks in time series data mining[C]// *Proc of Annual Meeting of the North American Fuzzy Information Processing Society*. 2005: 754-759.