基于会面合并事件的社会关系强度度量模型

陈 增[1]，王科人[1*]，杨 铮[2]

(1. 盲信号处理重点实验室 成都 610041; 2. 清华大学软件学院 北京 海淀区 100084)

【摘要】针对时空数据条件下的网络用户社会关系挖掘，该文提出了一种社会关系强度度量模型—EPTDD(熵-个人-时间-时长-直径)模型，在会面合并事件基础上，从位置、时间、用户等多方面综合考虑会面事件对社会关系强度的贡献。首先，对用户之间会面事件进行检测，并将发生时间相近的会面事件进行合并处理，得到更加接近现实情况的会面合并事件；之后，以位置熵、位置个人背景、时间、时长和直径5种要素对会面合并事件的权重进行刻画；最后综合上述要素，分别实现社会关系强度度量的无监督和有监督方法。在3个真实数据集上的实验结果表明，该文提出的EPTDD模型能够有效度量用户之间的社会关系强度，且优于现有方法。

关 键 词 数据挖掘; 会面合并事件; 社会关系度量; 时空数据

中图分类号 TP391 N94 **文献标志码** A doi:10.3969/j.issn.1001-0548.2019.01.016

A Social Relationship Strength Measurement Model Based on Merged Meeting Events

CHEN Zeng[1], WANG Ke-ren[1*], and YANG Zheng[2]

(1. National Key Laboratory of Science and Technology on Blind Signal Processing Chengdu 610041;
2. School of Software, Tsinghua University Haidian Beijing 100084)

Abstract In order to mining the social relationship between users based on spatio-temporal data, a novel entropy-personal-time-duration-diameter (EPTDD) model is proposed for measuring relationship strength in this paper. The model considers the effect on relationship measurement of meeting events from several different sides including location, time and user on the basis of merged meeting events. Firstly, meeting events are merged according to their occurring times to obtain merged meeting events that are more correlated with real life. Each merged meeting event is then weighted from location entropy factor, location personal factor, time factor, duration factor and diameter factor. Finally, the five factors are synthesized to obtain unsupervised and supervised methods for measuring social relationship. Experimental results on three different real datasets demonstrate that our methods perform significantly more favorable than existing methods on the effectiveness.

Key words data mining; merged meeting events; social science computing; spatiotemporal

随着手机和空间定位技术的发展和广泛应用，大量位置信息的获取成为可能。目前对手机的位置信息获取与记录主要有两种方式：1) 手机可以利用基站信息来确定其所处的位置并进行记录，例如基站号就可以认为是手机所处位置的标识[1]；2) 越来越多的手机应用允许用户分享他们的位置和移动信息，例如在Facebook、微信等上，用户可以上传带有位置标签的文字和图片，而Foursquare等应用记录了大量用户的签到数据，其中包含位置信息[2]。

大规模时空数据吸引研究人员针对时空数据与用户社会关系的相关性开展研究[3]。这些研究工作对于广告投放[4]、朋友推荐[5]、经济发展[6]甚至犯罪检测[7]等大量应用具有重要的现实意义。文献[8]发现用户之间的社会关系与物理距离具有很强的相关关系，即距离较近的用户之间更有可能存在较强的社会关系。大量的研究主要通过社会关系理解人类的移动行为[9-11]，如文献[12]通过对用户的移动距离和社会关系的分析，发现短距离的周期移动与社会网络结构几乎没有关联，而长距离移动则受社会关系影响很大。

本文基于时空数据对用户之间的社会关系强度进行度量。针对此研究目标，文献[13]发现在相近时

收稿日期：2017 – 07 – 10；修回日期：2017 – 11 – 01
基金项目：国家自然科学基金(61361166009)
作者简介：陈增(1995 –)，男，主要从事时空数据挖掘和社会网络分析方面的研究.
通信作者：王科人，E-mail: cfan662003@163.com

间出现在同一个地方(即会面)是表征朋友关系的一个重要指标。文献[14]提出社会关系强度与会面事件的频次具有强正相关关系，即会面频次越高，则两个用户是朋友的概率更大。

进一步研究发现，在基于会面事件度量社会关系时，一对用户的不同会面事件并不是等权重的。文献[15]提出了一种基于位置熵的度量模型，该模型使用会面地点的信息熵对发生在热门地点的会面事件进行惩罚，这样可以降低偶然相遇对度量社会关系的影响。在文献[15]的基础上，文献[16]提出了一种综合了位置熵、个人背景和时间要素的方法，表现出比文献[15]更优的度量效果。除此之外，文献[17]提出了一种基于用户移动轨迹之间的距离的社交关系预测方法，这种方法可以用于预测不存在会面事件的用户之间的朋友关系，然而这种方法的应用场景受到一定的限制，无法应用到以离散基站标号记录位置的场景中。文献[18]则将会面特征与共现(两个用户出现在同一地点)特征进行结合，并使用地点的熵进行加权，取得了很好的效果。除此之外，还有一些研究将时间维度、空间维度和社交网络结合，通过有监督方法对朋友关系进行预测[19]。

考虑到两个用户发生会面时的时间、地点、时长以及会面过程中移动距离等要素的不同，表征着会面事件对于度量用户之间的社会关系强度的重要性也不同，本文在文献[15-16]的基础上综合多种要素，提出EPTDD模型用来描述和度量不同要素条件下会面事件的重要性。该模型首先对会面事件进行合并，然后基于会面合并事件的5个要素对会面事件的权重进行定量描述。实验表明，EPTDD模型可以更好地用于度量用户之间的社会关系强度。

1　EPTDD模型

基于会面事件对用户之间的社会关系进行度量时，会面事件的权重受以下5个方面影响：

1) 会面地点在所有用户中的全局热度。当一次会面事件发生在一个热点区域(如火车站等公共场所)时，则该会面事件可能是个偶然事件，其对度量会面双方社会关系强度的贡献可能较小。

2) 会面地点对于会面双方而言的“局部热度”。当会面事件发生在会面双方经常访问的地点(如办公室)时，此次会面的权重应当降低。

3) 会面时间的“局部热度”。如果用户通常在每天的同一时间与他人进行会面，则应该对在这个时间附近发生的会面事件进行惩罚，降低其对关系度量的影响。

4) 会面持续时长。显然，会面的持续时长越长，其权重越大。

5) 会面过程中用户的移动距离。具有较强社会关系的用户之间经常会发生一起外出等行为，即两个用户在一次较长时长的会面过程中会发生一定距离的移动。移动距离的大小在一定程度上反映了“会面”的真实性，应考虑将其作为度量社会关系强度的要素之一。

考虑到以上5个方面，本文提出EPTDD模型，在对会面事件进行合并后，从位置熵、位置个人、时间、时长和直径要素分别对以上5个方面进行描述和度量，其框架如图1所示。图中，EPTDD-U表示基于EPTDD模型的无监督方法，EPTDD-S表示基于EPTDD模型的有监督方法。

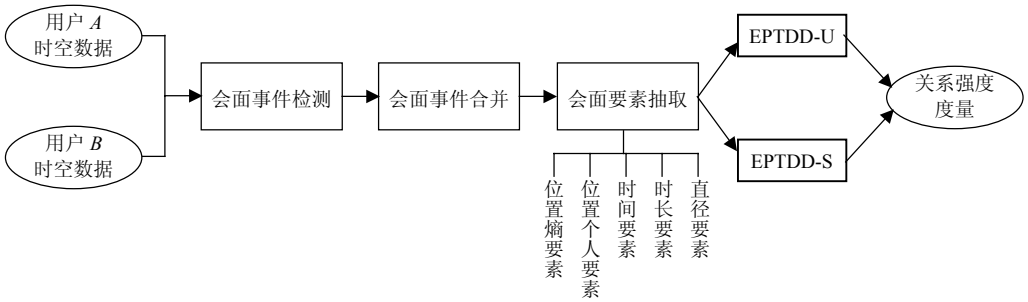

图1　基于EPTDD模型的社会关系强度度量框架

1.1　会面合并事件

定义一个含有 n 个用户的用户集 $U=\{U_i,i=1,2,\cdots,n\}$，其中任意一个用户 U_i 的时空数据可以表示为 $S_i=\{(t_r^i,\mathrm{loc}_r^i),r=1,2,\cdots;R_i\}$，其中 (t_r^i,loc_r^i) 为一个包含了时间戳和位置的二元组，R_i 是用户 U_i 的时空数据记录条数。由于数据收集工具和方法的不同，位置 loc_r^i 的形式可能有两种情况：1) 代表位置编号的一个离散数字；2) 使用经度和纬度表示的一个连续的地理坐标。

当两个用户 U_i 和 U_j 几乎同时到达接近的地点时，说明这两个用户之间发生了一次会面事件，可以表示为：

$$\begin{cases}\mathrm{loc}_p^i \sim \mathrm{loc}_q^j \\ |t_p^i - t_q^j| < \tau\end{cases} \tag{1}$$

式中，$\mathrm{loc}_p^i \sim \mathrm{loc}_q^j$ 表示两个位置十分接近，更确切的说，当位置使用离散的数字表示时，$\mathrm{loc}_p^i \sim \mathrm{loc}_q^j$ 等价于 $\mathrm{loc}_p^i = \mathrm{loc}_q^j$；而当位置使用经纬度坐标表示时，$\mathrm{loc}_p^i \sim \mathrm{loc}_q^j$ 等价于 $\mathrm{dist}(\mathrm{loc}_p^i, \mathrm{loc}_q^j) < \mathrm{Dist}_{th}$，其中，$\mathrm{dist}(\mathrm{loc}_p^i, \mathrm{loc}_q^j)$ 表示两个位置之间的距离(如欧氏距离)，Dist_{th} 为距离阈值。

当存在记录 $(t_p^i, \mathrm{loc}_p^i) \in S_i$ 与 $(t_q^j, \mathrm{loc}_q^j) \in S_j$ 满足式(1)时，则可以认为用户 U_i 和用户 U_j 发生了一次会面事件。会面事件可以用四元组表示：$oe_{i,j;k} = (t_{i,j;k}^1, t_{i,j;k}^2, \mathrm{loc}_{i,j;k}^1, \mathrm{loc}_{i,j;k}^2)$，其中 $(t_{i,j;k}^1, \mathrm{loc}_{i,j;k}^1)$ 代表两条记录中时间戳较小的记录。

给定用户 U_i 和 U_j，两个用户之间的所有会面事件可以表示为集合 $\mathrm{OE}_{ij} = \{\mathrm{oe}_{i,j;k}, k = 1,2,;R_{ij}^o\}$，其中 R_{ij}^o 表示会面事件的总频次。考虑到实际情况中，一次持续较长时间的会面事件可能被检测到多次，也就是说在集合 OE_{ij} 中，可能有多条记录对应现实中的一次会面事件的情况，因此本文考虑对时间上相邻的会面事件进行合并。

会面事件序列 $\{\mathrm{oe}_{i,j;k_1}, \mathrm{oe}_{i,j;k_1+1}, \cdots, \mathrm{oe}_{i,j;k_2}\}$ 能够被合并，当且仅当：

$$\left| t_{i,j;k_1}^2 - t_{i,j;k_2}^2 \right| < \tau_s \tag{2}$$

本文将合并之后的会面事件称为“会面合并事件”。每个会面合并事件可以用一个五元组表示：$e_{i,j;h} = (t_{i,j;h}, \mathrm{loc}_{i,j;h}^1, \mathrm{loc}_{i,j;h}^2, \mathrm{dur}_{i,j;h}, \mathrm{dia}_{i,j;h})$，其中 $t_{i,j;h}$ 表示会面合并事件的开始时间，$\mathrm{loc}_{i,j;h}^1$ 和 $\mathrm{loc}_{i,j;h}^2$ 分别表示会面合并事件发生时会面双方所在的位置，$\mathrm{dur}_{i,j;h}$ 表示时长，$\mathrm{dia}_{i,j;h}$ 为会面合并事件的直径，表示会面过程中会面双方的最大移动距离。在后文为了简化，在不引起歧义的情况下，将下标 $i,j;h$ 简写为 h。

会面合并事件五元组可以使用式(3)进行计算：

$$\begin{cases} t_h = t_{k_1}^1 \\ (\mathrm{loc}_h^1, \mathrm{loc}_h^2) = \mathrm{rand}((\mathrm{loc}_{k_1}^1, \mathrm{loc}_{k_1}^2), \cdots, (\mathrm{loc}_{k_2}^1, \mathrm{loc}_{k_2}^2)) \\ \mathrm{dur}_h = t_{k_2}^2 - t_{k_1}^1 \\ \mathrm{dia}_h = \begin{cases} \max\limits_{\substack{k_1 \le r,s \le k_2 \\ \alpha = 1,2}} (\mathrm{dist}(\mathrm{loc}_r^\alpha, \mathrm{loc}_s^\alpha)) & \text{for continuous locations} \\ \sum\limits_{k_1 \le m,n \le k_2} \delta(\mathrm{loc}_m^1 - \mathrm{loc}_n^1) & \text{for discrete locations} \end{cases} \end{cases} \tag{3}$$

式中，$\mathrm{rand}(i,j,k,\cdots)$ 为一个随机选择函数，表示从输入 $i,j,k,\cdots$ 中随机选择一个元素。使用随机选择函数，是因为考虑到受数据采集手段的限制，用户可能在多个位置之间来回切换，此时应该倾向于使用出现频次较多的位置来代替用户的位置，使用随机选择的方法等效于给出现频次较多的位置赋予更大的被选择概率，而出现频次较少的位置被选择概率较小。另外从式(3)可以看到，dia_h 对于不同的位置表示有不同的计算方法，当位置使用离散数字记录时，会面合并事件的直径实际上表示会面过程中会面双方走过的不同位置编号个数。

1.2 度量会面合并事件权重

给定用户 U_i 和用户 U_j，使用 $E_{i,j} = \{e_h, h = 1,2,\cdots,R_{i,j}^e\}$ 表示两个用户之间的会面合并事件集合。基于会面合并事件，本节从5个要素考虑每次会面的权重。

1.2.1 位置熵要素和位置个人要素

位置熵要素主要考虑一个地点在所有用户中的受欢迎程度，也就是地点的全局热度。

用户 U_i 访问地点 loc 的所有记录集合为：

$$S_i(\mathrm{loc}) = \{(t_q^i, \mathrm{loc}_q^i) \in S_i : \mathrm{loc}_q^i \sim \mathrm{loc}\} \tag{4}$$

受数据采集手段的限制，不同用户的时空数据记录数可能存在较大差异。为避免记录条数不均匀导致的概率差异，在求取地点 loc 被用户 U_i 访问的概率时，本文将用户访问地点的频次 $|S_i(\mathrm{loc})|$ 使用访问地点的频率 $\Pr(i,\mathrm{loc}) = |S_i(\mathrm{loc})| / |S_i|$ 来代替：

$$P(i,\mathrm{loc}) = \frac{\Pr(i,\mathrm{loc})}{\sum_i \Pr(i,\mathrm{loc})} = \frac{|S_i(\mathrm{loc})|}{|S_i|} \Big/ \sum_i \frac{|S_i(\mathrm{loc})|}{|S_i|} \tag{5}$$

地点的热度可以使用信息熵来表征：

$$g(\mathrm{loc}) = -\sum_{i:P(i,\mathrm{loc})\neq 0} P(i,\mathrm{loc}) \log P(i,\mathrm{loc}) \tag{6}$$

这里使用香农熵的指数函数值计算每个地点的位置熵要素[15]：

$$\omega_{i,j}^{e'}(\mathrm{loc}) = \exp(-g(\mathrm{loc})) \tag{7}$$

为了更精确的度量位置熵要素的影响，针对会面事件，以两个地点的位置熵的几何平均对会面事件 $e_h = (t_h, \mathrm{loc}_h^1, \mathrm{loc}_h^2, \mathrm{dur}_h, \mathrm{dia}_h) \in E_{ij}$ 的位置熵权重进行度量：

$$\omega_{i,j}^e(e_h) = \sqrt{\omega_{i,j}^{e'}(\mathrm{loc}_h^1)\omega_{i,j}^{e'}(\mathrm{loc}_h^2)} \tag{8}$$

位置个人要素主要考虑的是对于会面双方而言会面地点的受欢迎程度，其计算式为[16]：

$$\omega_{i,j}^p(e_h) = -\log(\rho(i,\mathrm{loc}_h^1)\rho(j,\mathrm{loc}_h^2)) \tag{9}$$

式中，$\rho(i,\mathrm{loc})$表示用户U_i访问地点loc的频率。

1.2.2 时间要素

在文献[16]中，当一次会面事件与其他会面事件在时间上邻近时，则这次会面事件的权重会降低。然而，由于本文将时间上接近的会面事件进行了合并，得到的会面合并事件之间均具有相对较大的时间间隔，在这种情况下，文献[16]的这种方法将会失效。用户U_i如果经常在每天的同一时间与他人发生会面事件，则很有可能是该用户的工作或其他规律性活动的结果，这就意味着在该时间发生的会面事件可能并不能很好的表征社会关系强度，应该降低其权重。

与位置个人因素类似，定义会面事件$e_h \in E_{i,j}$的权重为：

$$\omega_{i,j}^{t}(e_h) = -\log(\rho(i,t_h)\rho(j,t_h)) \tag{10}$$

式中，$\rho(i,t)$为用户U_i在t时刻的会面事件密度函数：

$$\rho(i,t) = \sum_{e_h \in M_i} \exp(-c_t \,|\, t_h - t \,|)/|\, M_i \,| \tag{11}$$

式中，M_i表示用户U_i的所有会面事件，即：$M_i = \{e_h : e_h \in E_{i,j}, \forall U_j \in U\}$。

1.2.3 时长要素

在真实世界中，持续时长很短的会面事件通常是偶然发生的，而持续时长较长的会面事件则通常发生在具有较强社会关系的用户之间，因而需要考虑会面时长要素对于度量社会关系强度的影响。

描述时长要素的权重表达式应满足3个条件：

1) 当一次会面合并事件的时长为0时，则这次会面事件的权重应该为0；

2) 会面合并事件的权重随时长单调递增；

3) 时长权重应该能够灵敏捕捉时长在较小值时的波动，也就是时长权重的二阶导数小于0。

基于以上3个条件，给定会面合并事件$e_h \in E_{i,j}$，考虑到权重不应该为负值，使用偏差值为1的对数函数计算会面事件的时长权重：

$$\omega_{i,j}^{d}(e_h) = \log(c_d \mathrm{dur}_h + 1) \tag{12}$$

1.2.4 直径要素

关系亲密的两个用户之间可能存在“伴随移动”行为，即两个用户在一段时间内具有相近的移动轨迹。例如，朋友之间边走边聊，或是家人朋友共同旅行等。相应地，关系强度较弱的用户之间即使偶然出现一次会面事件，也很难出现“伴随移动”行为。在单次会面合并事件上，这里采用直径要素来刻画此类“伴随移动”行为对用户关系度量的影响。

与时长要素类似，给定一次会面合并事件$e_h \in E_{i,j}$，使用其直径dia_h的对数函数计算直径权重：

$$\omega_{i,j}^{m}(e_h) = \log(c_m \mathrm{dia}_h + 1) \tag{13}$$

2 基于EPTDD模型的社会关系强度度量

为了对社会关系强度进行度量，本节分别针对无监督和有监督情况，提出基于EPTDD模型的社会关系强度度量方法。其中，无监督方法适用于无任何先验知识的情况，而在有监督方法中，则利用已知社会关系强弱或是否具有真实社会关系(例如朋友关系)的用户对进行训练。

2.1 基于EPTDD模型的无监督方法

记两个用户U_i和U_j的所有会面事件集合为$\mathrm{OE}_{i,j} = \{oe_h, h = 1,2,\cdots,R_{i,j}^{o}\}$，其中$R_{i,j}^{o} \geqslant R_{i,j}^{e}$。大量相关研究基于会面事件集合进行社会关系强度的度量和预测[11, 15-16]。比如，一种最简单的度量方法就是直接使用会面事件的频次作为度量值[14]，即：

$$G_{o,\mathrm{MF}}(\mathrm{OE}_{i,j}) = |\,\mathrm{OE}_{i,j}\,| \tag{14}$$

后续研究发现，不同会面事件对度量社会关系应该具有不同的权重[15-16]，EPTDD模型也反应了这一点。通过对会面合并事件的5种要素进行综合，实现用户之间社会关系的强度度量。文献[16]使用位置个人要素的最大值对位置个人权重进行建模，但最大值过分强调了单次特殊会面的重要性，放大了数据采集阶段和会面事件检测阶段误差所导致的影响。因而本文使用平均值与标准差的和来代替最大值，得到度量值为：

$$G_{e,\mathrm{EPTDD}}(E_{i,j}) = \prod_{x \in \{p,t,d,m\}} K_{e_h}(\omega_{i,j}^{x}(e_h)) \sum\nolimits_{e_h} \omega_{i,j}^{e}(e_h) \tag{15}$$

式中，$K_{e_h}(\cdot)$表示两个用户所有会面合并事件单一权重的平均值与标准差的和。

2.2 基于EPTDD模型的有监督方法

5种要素隐含了有利于度量用户之间社会关系强度的若干信息。为了尽可能地保留这些信息，在有监督方法中，针对5种要素分别提取多维统计特征，并结合分类器从已标注数据中学习训练分类器参数，期望得到更好的社会关系强度度量结果。

给定用户U_i和U_j的所有会面事件的集合$E_{i,j} = \{e_h, h = 1,2,\cdots,R_{i,j}^{e}\}$，提取5种要素所有权重的最大值、平均值和标准差以及会面频次(meeting frequency, MF)、不同会面之间的最大时间间隔($\max(\mathrm{interval})$)和平均时间间隔($\overline{\mathrm{interval}}$)作为有监

督方法的18维特征。除此之外，会面发生时间段的不同也一定程度上表征了会面事件对于关系强度度量权重的不同。比如，朋友之间在周末发生会面事件的频次更高。将所有会面事件按照发生的时间分为工作日的白天、工作日的晚上和周末3个时间段内的集合，并分别提取以上18维特征。

将一周的时间记为 $\mathrm{seg}_0=[0,7)$，其中0表示周一零时，则3个时间段集合 $\mathrm{seg}_i(i=1,2,3)$ 分别为：

$$\begin{cases}\mathrm{seg}_1=\left\{t:t\in[0,5)\text{且}t\%1\in\left[\dfrac{7}{24},\dfrac{17}{24}\right)\right\}\\ \mathrm{seg}_2=\left\{t:t\notin\mathrm{seg}_1\bigcup\mathrm{seg}_3\right\}\\ \mathrm{seg}_3=\left\{t:t\in\left[4\dfrac{17}{24},7\right)\bigcup\left[0,\dfrac{7}{24}\right)\right\}\end{cases}\tag{16}$$

则时间段 $\mathrm{seg}_i(i=0,1,2,3)$ 内的会面事件集合可以记为：$E_{i,j}^x=\{e_h\in E_{i,j}:t_h\in\mathrm{seg}_x,x=0,1,2,3\}$。从中提取的72维特征的简单描述如表1所示。

表1 基于EPTDD模型的有监督方法特征描述

	特征	维数
总体描述特征	对 $E_{i,j}^x(x=0,1,2,3)$ 分别求取的MF、max(interval)、$\overline{\text{interval}}$	3×4
5种要素特征	对 $E_{i,j}^x(x=0,1,2,3)$ 分别按每种要素加权的最大值、平均值和标准差	3×4×5

在应用时，若训练数据标签为{“弱关系”“强关系”}信息，可利用随机森林等分类器对其进行训练，并将识别阶段输出的预测概率作为社会关系强度度量值；若训练数据集中用户之间关系强度为连续值，则可以通过回归方法进行训练。

3 实验结果

为了证明EPTDD模型的有效性，在3个真实数据集上对基于EPTDD模型的社会关系强度度量方法进行了测试。

3.1 数据集

本文所使用的公开数据集分别为MIT现实世界数据挖掘集(MIT数据集)、Gowalla数据集和Brightkite数据集。这3个数据集使用两种完全不同的方式收集。其中MIT数据集通过手机确认其所处的基站编号得到用户的位置信息，因而数据集中的位置使用离散编号记录。Gowalla和Brightkite数据集通过用户分享的签到信息收集用户的时空数据，其位置数据为GPS定位得到的地理坐标值。

MIT数据集的时空数据是从2004年9月~2005年5月收集到的106位用户的记录。本文选择其中时空数据记录条数超过200条的共87个用户的所有记录进行实验。Gowalla和Brightkite数据集是从基于位置的社交网络服务中收集的用户签到数据。其中，Gowalla的时间跨度为2009年2月~2010年10月，共包含了107 092个用户，Brightkite的时间跨度为2008年4月~2010年10月，共包含了58 228个用户，本文从这两个数据集中各提取时空数据记录最多的5 000名用户进行实验。3个数据集均含有朋友关系的社交网络，可以用作实验的基准标定数据。3个数据集提取后的其他统计数据如表2所示。

表2 提取后数据集的统计数据

数据	数据集		
	MIT	Gowalla	Brightkite
提取用户数	87	5 000	5 000
朋友数	47	27 678	68 228
记录总条数	291 176	2 563 771	3 508 326
平均记录条数	3 346	512	701

3.2 实验方法及参数配置

为了验证所提出EPTDD模型的有效性，本文选取MF作为基准度量方法，选择文献[16]中的无监督方法(personal global time, PGT)方法和文献[18]中结合会面事件和共现事件特征的有监督方法(vlocation, VLoc)作为对比方法。由于会面合并事件在本文中首次提出，已有方法均基于合并之前的会面事件进行度量，因而为了保证结果的可信性，对比MF方法和PGT方法基于会面事件进行度量，而EPTDD-U和EPTDD-S均基于本文提出的会面合并事件进行度量。考虑到MIT数据集中时空数据记录的采样间隔在一段时间内比较固定且大约为1.5 min一次，因而在使用MIT数据集进行实验时，设置 $\tau=3$ min，$\tau_s=0.5$ h。而在Gowalla和Brightkite数据集中的时空数据的时间间隔要大得多(15 min以上)且不固定，因而在使用Gowalla和Brightkite数据集时设置 $\tau=1$ h，$\tau_s=2$ h。除此之外，考虑到一个人在一分钟行走的距离大约为50 m，因而在使用Gowalla和Brightkite数据时设置 $\mathrm{Dist}_{th}=50$ m。

在进行测试时，EPTDD-S方法每次随机选取数据集中70%的用户对的数据作为训练集，其余30%作为测试集，并在每个数据集上运行20次取平均值作为最终结果。

3.3 评价指标

本文使用准确率-召回率(precision-recall)曲线、$F1$ 指标、AUC(ROC曲线下的面积)、

AP(precision-recall曲线下的面积)、正确率(Acc)和Gm[20](G-mean)等指标对EPTDD模型进行评估。使用TP、FP、TN和FN分别表示真正、假正、真负和假负样本，则准确率和召回率可以表示为：

$$\begin{cases}\text{Precision}=\dfrac{|\text{TP}|}{|\text{TP}|+|\text{FP}|}\\ \text{Recall}=\dfrac{|\text{TP}|}{|\text{TP}|+|\text{FN}|}\end{cases}\tag{17}$$

正确率、$F1$指标和G_m可以分别定义为：

$$\begin{cases}\text{Acc}=\dfrac{|\text{TP}|+|\text{TN}|}{|\text{TP}|+|\text{FP}|+|\text{TN}|+|\text{FN}|}\\ F1=\dfrac{2\cdot\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}\\ G_m=\sqrt{\dfrac{|\text{TP}|}{|\text{TP}|+|\text{FN}|}\times\dfrac{|\text{TN}|}{|\text{TN}|+|\text{FP}|}}\end{cases}\tag{18}$$

3.4 实验结果

5种方法的P-R曲线如图2所示，其他指标的结果如表3所示。

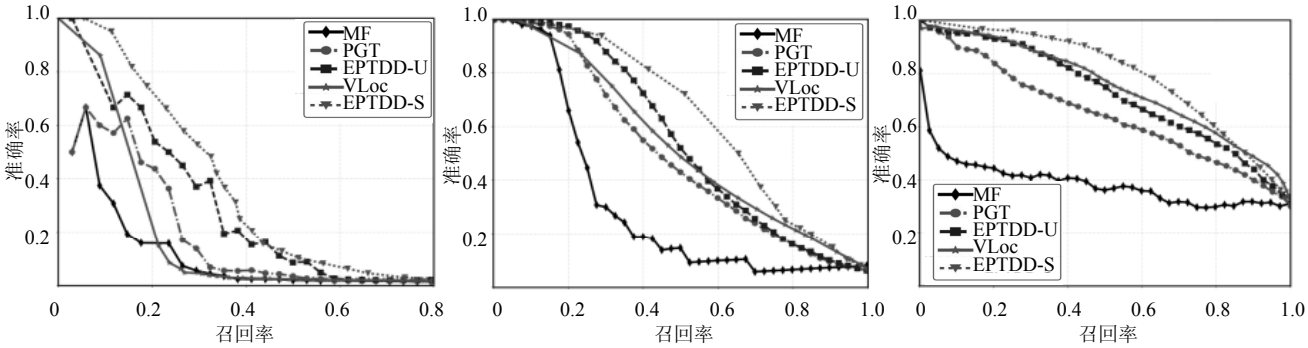

图2　用户社会关系强度度量P-R曲线比较

表3　用户社会关系强度度量性能比较

方法	MIT					Gowalla					Brightkite				
	$F1$	AUC	AP	Acc	G_m	$F1$	AUC	AP	Acc	G_m	$F1$	AUC	AP	Acc	G_m
MF	18.82	61.10	8.39	98.60	59.53	32.21	70.07	31.06	94.75	64.69	41.25	64.06	38.40	75.14	62.48
PGT	28.57	64.10	14.46	98.64	63.21	46.81	84.09	52.24	95.08	77.63	60.60	78.69	64.59	76.68	75.56
EPTDD-U	35.48	71.14	22.89	98.69	69.79	52.58	84.97	55.10	95.52	78.07	65.11	83.15	72.46	80.00	75.56
VLoc	25.91	64.51	17.87	98.65	60.67	49.67	88.64	53.69	95.17	80.64	67.50	86.09	75.03	80.62	77.41
EPTDD-S	39.22	78.04	32.06	98.73	73.17	59.91	89.48	63.86	95.91	82.04	70.46	86.75	79.43	83.57	78.89

从图2和表3可以得到以下结果：

1) 从指标结果来看，本文的EPTDD-U方法和EPTDD-S方法在所有指标上均优于MF方法和PGT方法。以F1指标为例，EPTDD-U方法在MIT数据集、Gowalla数据集和Brightkite数据集上的性能相比于无监督的PGT方法分别提升了24.19%、12.33%和7.44%；EPTDD-S方法相比于VLoc方法则分别提升51.37%、20.61%和4.39%。另外值得注意的是，由于3个数据集中正负样本具有很高的不均衡性，负类样本占比很高，所有方法的正确率(Acc)均较高，此时不能使用正确率作为度量方法有效性的重要指标。

2) 从数据集来看，5种方法在Gowalla和Brightkite数据集上的结果优于在MIT数据集上的结果，这可能有两方面原因：① Gowalla和Brightkite收集的是用户主动签到的位置数据，而MIT数据集收集过程对于用户来说是被动的，由于朋友之间会面时更倾向于在社交网络上签到，因而Gowalla和Brightkite数据集的数据对于挖掘社会关系更有利；② 相比于MIT中使用基站编号记录位置信息，Gowalla和Brightkite中使用GPS记录用户经纬度信息更为精确。另外还可以看到，本文方法相对于PGT方法在MIT数据集上性能提升最大，这可能有两个原因：Ⅰ. 相比于Gowalla和Brightkite数据集，MIT数据集的时空数据更密集，因而基于时间间隔的会面事件的合并更可信，得到的会面时长更接近真实会面时长；Ⅱ. 同样由于时空数据更密集，在MIT数据集中可以检测到更多的伴随移动现象，因而直径因素能够起到较好的效果。

3) 从度量方法上来看，有监督方法利用了已标注数据的信息，并综合了要素的多种统计值，相比

于无监督方法度量效果明显更优。

4 结 束 语

针对基于时空数据的用户社会关系强度度量问题，本文提出了基于会面合并事件的EPTDD模型。该模型在对会面事件进行检测、合并的基础上，从5个要素对会面事件进行加权。通过综合这些要素，本文提出了基于EPTDD模型的无监督和有监督方法用于度量用户社会关系强度的度量。实验结果表明，在社会关系强度度量上，本文提出的EPTDD模型优于已有方法，且在精确而密集的时空数据条件下表现更优。一方面，精确位置数据条件降低了会面事件检测阶段引起的误差；另一方面，密集时空数据条件为本文EPTDD模型的基础——会面事件的合并提供了有效的支撑，提高了会面合并事件的会面时长和伴随移动距离的可信度。

考虑到网络用户之间除了在时空数据层面上会发生会面事件外，还可能会存在一定通联关系，因而本文下一步的研究将围绕时空数据与通联数据结合条件下的社会关系强度度量展开。

参 考 文 献

[1] ASGARI F, GAUTHIER V, BECKER M. A survey on human mobility and its applications[EB/OL]. [2017-03-01]. https://www.researchgate.net/publication/244989928_A_survey_on_Human_Mobility_and_its_applications.

[2] BAO J, ZHENG Y, WILKIE D, et al. Recommendations in location-based social networks: a survey[J]. Geoinformatica, 2015, 19(3): 525-565.

[3] WANG D, PEDRESCHI D, SONG C, et al. Human mobility, social ties, and link prediction[C]//ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, USA: ACM, 2011: 1100-1108.

[4] DHAR S, VARSHNEY U. Challenges and business models for mobile location-based services and advertising[J]. Communications of the ACM, 2011, 54(5): 121-128.

[5] ZHENG V W, ZHENG Y, XIE X, et al. Collaborative location and activity recommendations with gps history data[C]//International Conference on World Wide Web. Raleigh, North Carolina, USA: [s.n.], 2010: 1029-1038.

[6] HOLZBAUER B O, SZYMANSKI B K, NGUYEN T, et al. Social ties as predictors of economic development[M]. [S.l.]: Springer International Publishing, 2016: 178-185.

[7] GE Y, XIONG H, LIU C, et al. A taxi driving fraud detection system[C]//International Conference on Data Mining. Vancouver: IEEE Computer Society, 2011: 181-190.

[8] DESCIOLI P, KURZBAN R, KOCH N, et al. Best friends alliances, friend ranking, and the myspace social network[J]. Perspect Psychol SCI, 2011, 6(1): 6-8.

[9] ZHANG D, VASILAKOS A V, XIONG H. Predicting location using mobile phone calls[J]. ACM Sigcomm Computer Communication Review, 2012, 42(4): 295-296.

[10] PANG J, ZHANG Y. Exploring communities for effective location prediction[C]//International World Wide Web Conference. Florence: ACM, 2015: 87-88.

[11] TANG J, CHANG Y, LIU H. Mining social media with social theories:a survey[J]. ACM Sigkdd Explorations Newsletter, 2014, 15(2): 20-29.

[12] CHO E, MYERS S A, LESKOVEC J. Friendship and mobility: User movement in location-based social networks[C]//Proceedings of the 17th ACM SIGKDD International Conference On Knowledge Discovery And Data Mining. [S.l.]: ACM, 2011: 1082-1090.

[13] EAGLE N, PENTLAND A, LAZER D. Inferring friendship network structure by using mobile phone data[J]. Proceedings of the National Academy of Sciences of the United States of America, 2009, 106(36): 15274-15278.

[14] CRANDALL D, BACKSTROM L, COSLEY D, et al. Inferring social ties from geographic coincidences[J]. Proceedings of the National Academy of Sciences of the United States of America, 2010, 107(52): 22436-22441.

[15] PHAM H, SHAHABI C, LIU Y. EBM: an entropy-based model to infer social strength from spatiotemporal data [C]//ACM SIGMOD International Conference on Management of Data. [S.l.]: ACM, 2013: 265-276.

[16] WANG H, LI Z, LEE W C. PGT: Measuring mobility relationship using personal, global and temporal factors[C]//International Conference on Data Mining. Atlantic: IEEE Computer Society, 2014: 570-579.

[17] ZHANG Y, PANG J. Distance and friendship: a distance-based model for link prediction in social networks[M]. [S.l.]: Springer International Publishing, 2015.

[18] VALVERDE-REBAZA J, ROCHE M, PONCELET P, et al. Exploiting social and mobility patterns for friendship prediction in location-based social networks[C]// International Conference on Pattern Recognition. Cancún, Mexico: IEEE, 2016: 2526-2531.

[19] CHENG R, PANG J, ZHANG Y. Inferring friendship from check-in data of location-based social networks[C]// IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. [S.l.]: IEEE, 2015: 1284-1291.

[20] HE H, GARCIA E A. Learning from imbalanced data[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263-1284.

编 辑 叶 芳