doi:10.11887/j.cn.201703013 http://journal.nudt.edu.cn

日常交互中朋友关系强度度量方法*

史殿习,杨若松,莫晓赟,李 寒,赵邦辉
(国防科技大学 计算机学院,湖南 长沙 410073)

摘 要:针对如何度量日常生活中人们之间的关系强度问题展开研究,提出一个从日常轨迹、语义位置以及语义标签三个层次度量朋友之间关系强度的层级模型FRSHV。采用动态时间规整模型通过计算朋友之间的空间距离来度量其日常轨迹之间的相似度,进而使用轨迹序列熵值对用户每天轨迹的相似度进行加权处理,将其作为朋友之间的关系强度;采用主题模型隐含狄利克雷分布分别计算朋友之间的基于语义位置和语义标签的行为模式的相似性,将其作为朋友之间的关系强度;采用集成学习的思想对三个层次的度量结果进行投票,以投票结果作为最终的朋友之间的关系强度。在公开数据集上对FRSHV模型的有效性进行实验验证,结果表明该模型能够有效地度量朋友之间的关系强度。

关键词:关系强度;轨迹相似度;动态时间校正;熵;潜狄利克雷分布;投票
中图分类号:TP391 **文献标志码**:A **文章编号**:1001-2486(2017)03-077-08

Measuring method for friend relationship strength in daily communication

SHI Dianxi, YANG Ruosong, MO Xiaoyun, LI Han, ZHAO Banghui
(College of Computer, National University of Defense Technology, Changsha 410073, China)

Abstract: The FRSHV(friend relationship strength hierarchy vote), a hierarchical model, was proposed to measure the friend relationship strength by user's daily moving track, semantic positions and the corresponding semantic labels. The daily track similarity was measured by dynamic time warping model using the spatial distance between friends, and the results were then weighed by the entropy of track series. The similarities of friend's behavior patterns were inferred by latent Dirichlet allocation topic model, respectively using semantic positions and the corresponding semantic labels. Finally, these three similarity results were voted for the ultimate relationship strength. The FRSHV was evaluated by using an open dataset and the results proved the validity of the model in inferring friend's relationship strength.

Key words: relationship strength; trajectory similarity; dynamic time warping; entropy; latent Dirichlet allocation; vote

目前,内嵌了各种各样传感器的智能手机已经成为人们日常生活中集通信、计算及感知于一体的移动平台,通过内嵌的各种传感器如GPS、加速度、麦克风等可以随时随地感知和获取人们自身及其周围环境的各种信息。通过智能手机所收集的各种数据研究人们之间的日常交互行为和人们之间的社会关系成为普适计算领域当中一个重点研究的问题。文献[1]基于手机所收集的各种数据推理人们之间的社会交互关系以及群组的活动韵律,从而洞察个人和组织的行为模式;文献[2]研究分析了家庭和朋友圈对个体行为在社交方面的影响;文献[3]研究了在校学生的日常活动、交互情况、精神健康与学业成绩之间的关系;文献[4]则从多渠道、细粒度地收集反映在校学生日常活动和交互情况的各种数据,从多个层面真实、全面地反映学生日常活动以及他们之间的交互行为和交互关系。但是,这些研究重点关注的是人们之间的日常交互行为和交互关系。关系强度度量的是人们之间的亲密程度,通过关系强度可以更好地了解人们之间关系的强弱,进而了解人们之间的亲密程度,从而可以更好地预测社会关系的演变以及社交结构的变化、促进信息传播以及传染疾病的预防与控制等。

社会关系强度理论始于文献[5]中对于弱关系的研究,将弱关系和强关系的测量分为四个维度,即交往人员之间的互动频率、感情的投入程度、关系亲密程度和在互惠互利上的交换程度;文献[6]对这四个维度做了相关指标化;文献[7]认

* 收稿日期:2016-02-10
基金项目:国家自然科学基金资助项目(61202117,91118008)
作者简介:史殿习(1966—),男,山东龙口人,教授,博士,博士生导师,E-mail:dxshi@nudt.edu.cn

为关系强度涉及关系的数量以及交往的频率。随着关系强度研究领域的不断发展,以互动频率、联系次数、亲密程度为关系强度核心测量指标的主流研究观点[8]逐渐形成。但是,如何度量社会网络中人们之间的关系强度一直是社交网络关系分析中的一个难点问题。

通过智能手机可以随时随地地获取位置、通话记录、短信、微信等体现人们之间日常交互和社会关系的各种信息,而人们之间的交互频率、时间、位置、地点、距离以及轨迹相似性等信息能够直接体现人们之间的交互关系以及关系强度,因为关系密切的人们之间更愿意面对面地进行交流,如朋友之间就会经常进行面对面的交流(如聚会、一起游览等)。通过对这些信息的分析处理,可以更好地度量朋友之间的关系强度。为了方便描述,本文将分析处理的对象称为用户,并认为用户和陌生人之间的关系强度应该为零如果他们互不认识。虽然对一个用户来说,即使其与一些陌生人并不认识,其也可能会经常与之在一些地方同时出现,但本文只考虑用户和其好友之间的关系强度。本文设想能够在一定程度上反映两个朋友之间的关系,而非完整全面地度量两个用户之间的关系;认为使用手机上所有传感器的全部数据能够精确地分析朋友之间的关系强度。轨迹数据是手机传感器数据非常重要的组成部分,本文主要研究如何只使用轨迹数据度量朋友之间的亲密程度。文献[9]认为用户之间的关系强度与用户共同出现的时间和共同出现的位置相关,提出了一个基于 GPS 轨迹数据的层级模型,根据用户的 GPS 轨迹来度量用户之间的关系强度,并在仿真数据集上进行了实验验证。

本文在文献[9]的基础上,针对如何度量日常生活中人们之间的关系强度问题展开研究,提出了一个可以对 GPS 数据和基站数据进行处理,从日常轨迹、语义位置以及语义标签三个层次度量用户与朋友之间关系强度的层级(Friend Relationship Strength Hierarchy Vote, FRSHV)模型。该模型采用动态时间校正(Dynamic Time Warping, DTW)模型通过计算用户与朋友之间的空间距离来度量其轨迹之间的相似度,进而使用轨迹序列熵值对用户每天轨迹的相似度进行加权处理,并将其作为用户与其朋友之间的关系强度;采用主题模型潜狄利克雷分布(Latent Dirichlet Allocation, LDA)分别计算用户与朋友之间的基于语义位置和语义标签的行为模式的相似性,将其作为用户与朋友之间的关系强度;采用集成学习的思想对三个层次的度量结果进行投票,以投票结果作为最终的用户与朋友之间的关系强度。

1 关系强度度量方法

通过对社会心理学相关研究成果的分析,认为人们之间的关系强度与他们之间的轨迹相似性以及日常行为的相似性密切相关,因此,为了有效地度量人们之间的关系强度,本研究从人们之间的日常轨迹和日常行为这两个角度出发,提出采用不同计算方法来计算人们之间的关系强度。

1.1 基于 DTW 模型的计算方法

空间距离能够直观反映人们之间在物理世界中的距离,空间距离非常接近的用户在现实生活中会有更多的面对面的交互,从而增强两个人之间的关系强度。根据社会心理学的研究成果,文献[10]在一个大型住宅区研究了接近性效应(接近性效应指两个人住得越近越可能是朋友),结果表明人们居住得越近,不管这种近是物理距离还是功能性距离,人们越容易称为朋友。文献[11]用实验证实了单纯接触效应,即熟悉性能够促进好感,实验结果表明接触频率越高喜欢程度越强。

DTW 是 Itakura 等于 1987 年[12]提出的一种距离度量方法。将用户的轨迹数据看作一个时间序列,因此同样可以使用 DTW 方法度量轨迹的相似度,并且将轨迹相似度作为人们之间的关系强度。通过深入分析 DTW 算法可知,序列的长度越长,则距离可能越大。因此,采用文献[13]中的三种归一化方法对 DTW 的计算结果进行进一步的处理和优化,即采用 DTW 结果除以最优变形路径的长度、DTW 结果除以两个序列中较短序列的长度以及 DTW 结果除以两个序列中较长序列的长度三种方法对 DTW 计算结果进行归一化,以便获得最优结果。

1.2 基于序列熵值加权的计算方法

通过日常生活体验很容易发现,如果两个人在晚上等休息时间经常一起出去,则其关系可能更亲密,因而他们之间的轨迹越可能相似。因此,可以使用熵值来度量用户每天活动的多样性,若某天活动越多样,则该天轨迹的相似度对总体轨迹的相似度贡献越大,进而对人们之间的关系强度贡献越大。

计算轨迹序列的熵值是为了对 DTW 计算结果进行加权,因为用户每天的轨迹序列的相似度对其总体相似度的贡献是不一样的,如果某一天

用户的轨迹序列的熵值越大,则这一天对总的相似度贡献越大。因此,使用用户每天轨迹序列熵值对用户与朋友之间每天的轨迹相似度进行加权,能够更真实地反应用户与朋友之间的关系强度(计算过程见第 2.2 节)。

1.3　基于主题模型 LDA 的计算方法

在日常生活当中,人们之间尤其是好友之间的行为模式之间具有一定的相似性,如经常在某些时间段(晚上)去一些地方(餐馆)等。基于位置的用户行为模式一方面能够反映用户在物理层次的相遇,另一方面能够在一定程度上体现用户的相似性,前文已经从社会心理学的角度阐述了相遇次数与用户关系强度的关系。文献[14]认为人们倾向于喜欢在态度、兴趣、价值观、背景和人格上与其相似的人,因此,在日常生活当中行为相似的人更可能成为朋友,而根据社会心理学的研究成果,用户的相似性对用户的关系强度也有一定的影响。为此,在通过基于用户轨迹度量用户之间关系强度的基础上,可进一步通过基于位置的用户日常行为来度量用户之间的关系强度。

LDA[15]是一个针对离散数据集合的产生式概率模型。文献[16]最先使用 LDA 主题模型发现用户的行为模式。在使用 LDA 模型发现用户基于位置的行为模式基础上,本研究进一步使用 LDA 主题模型来度量用户之间的关系强度,其核心思想如下:将每个用户每天去过的位置(语义位置或语义标签)序列视为一个句子,每个用户所有天的位置序列视为一篇文档,对所有用户所有天的位置序列使用 LDA 主题模型训练得到若干个主题。在计算两个用户之间的关系强度时,将这两个用户同一天的数据按固定长度的时间片划分;对于每个时间片内用户去过的位置,用训练好的 LDA 主题模型推断这些位置对应的主题分布;以同一时间片内,两个用户分别去过的位置对应的主题分布的余弦相似度作为这两个用户之间的关系强度(计算过程见第 2.2 节)。

2　关系强度度量模型框架

要真实全面地反映人们之间的关系强度,需要从不同角度和不同层次对人们之间的关系强度进行度量,为此,提出了一个层次化的、对用户与朋友之间的关系强度进行度量并对度量结果进行投票的模型 FRSHV,其框架结构如图 1 所示。FRSHV 模型是一个三层的、能够对通过 GPS 和基站的位置数据进行处理的度量模型,其从轨迹、语义位置以及语义标签三个层次对用户与朋友之间的关系强度进行度量,并使用集成学习的思想对三个层次度量结果进行投票,最终以投票结果作为用户与朋友之间的关系强度。

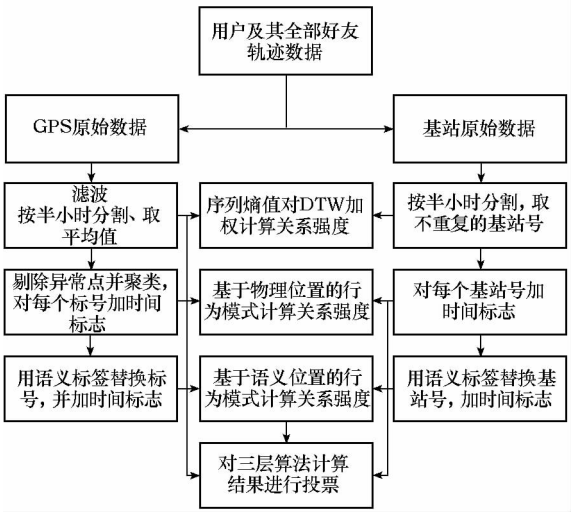

图 1　FRSHV 模型框架

Fig. 1　FRSHV model framework

在 FRSHV 模型当中,第一层度量主要针对用户的轨迹序列数据,根据不同用户轨迹序列的相似度来度量用户与朋友之间的关系强度;第二层度量主要针对用户的语义位置序列数据,考虑用户个人的基于位置的行为模式,如经常在什么时间出现在哪些位置等,根据不同用户行为模式的相似度来度量用户与朋友之间的关系强度;第三层度量主要针对用户的语义标签序列数据,物理上不同的位置可能拥有相同的语义标签,“办公室”“家”等语义概念在每个用户轨迹中都可能出现,而这些语义概念在原始数据中会表现为不同的基站号和区域号或不同的 GPS 经纬度,因此用户的语义标签数据更能体现用户群体的日常习惯,因此本层考虑的行为模式更倾向于群体的行为模式,从而根据不同用户在群体中表现出的行为模式来度量用户与朋友之间的关系强度。

2.1　GPS 及基站位置数据处理

在日常生活中,用户的位置既可以通过智能手机内嵌的 GPS 传感器获取,又可以通过对用户所处区域内的通信基站进行定位获取,基站定位更有利于用户隐私的保护。为了满足不同用户的不同需求,FRSHV 模型能够同时对 GPS 位置数据和基站位置数据进行处理。设用户集合为 U,其中 n 表示用户个数,D_i 表示用户 u_i 采集数据的日期的集合,m_i 表示用户 u_i 采集数据的总天数。F_i 表示用户 u_i 的全部朋友组成的集合,其中 f_i 表示

用户 u_i 的好友的个数。所有用户所有天的轨迹数据的集合为 *Trace*,其中 $Trace_i$ 表示用户 u_i 所有天采集的轨迹序列的集合,$Trace_{i,k}$ 表示用户 u_i 在 k 这一天的轨迹序列,$n_{i,k}$ 表示用户 u_i 在 k 这一天采集的轨迹数据的条数。对 GPS 和基站表示的用户轨迹序列进行预处理时,使用以下三种做法分别构造三层算法的输入。

2.1.1 轨迹数据处理

1)GPS 位置数据处理。首先,对每个用户每天的数据 $Trace_{i,k}$ 进行滤波,目的是减少数据噪声;而后对滤波后的数据按半小时进行划分,将用户 u_i 的每天数据 $Trace_{i,k}$ 按时间均分为 48 份,$Sep\_trace_{i,k,s}$ 表示第 i 个用户第 k 天第 s 份数据;对 $Sep\_trace_{i,k,s}$ 按经纬度计算平均值,并将用户 i 在第 k 天新的轨迹序列表示为 $Ntrace_{i,k}$,将用户 i 所有天采集的数据 $Ntrace_i$ 作为用户 u_i 使用第一层算法计算其与全部好友关系强度的输入。

2)基站位置数据处理。对每个用户每天的数据按半小时进行划分,即将用户 u_i 第 k 天的数据 $Trace_{i,k}$ 按时间均分为 48 份,$Sep\_trace_{i,k,s}$ 表示第 i 个用户第 k 天第 s 份数据;对每半个小时内数据计算依次不重复的基站号序列;再将每天 48 份数据重新拼成一个序列 $Ntrace_{i,k}$(表示用户 i 在 k 这一天采集的全部的数据),目的是对每天轨迹序列降维,以降低计算的复杂度,将用户 i 所有天的数据 $Ntrace_i$ 作为用户 u_i 使用第一层算法的输入。

2.1.2 语义位置数据处理

1)GPS 位置数据处理。采用文献[17]中的聚类方法对所有用户的轨迹数据进行聚类,得到全部语义位置序列 *Loc*。通过聚类得到用户 u_i 在第 k 天的语义位置序列 $Ltrace_{i,k}$;用户 u_i 的全部语义位置序列表示为 $Ltrace_i$,所有用户的所有语义位置序列表示为 *Ltrace*,对序列 *Ltrace* 添加对应的时间标记后记为 *LLtrace*,训练对应的 LDA 主题模型并记为 $LLDA(K)$,K 表示主题个数。对每个用户每天的数据按半个小时进行划分,即将用户 u_i 的每天数据 $Ltrace_{i,k}$ 按时间均分为 48 份,$Sep\_trace_{i,k,s}$ 表示第 i 个用户第 k 天第 s 份数据;对每份数据计算不重复出现的语义位置,并对每个位置加上时间标记。用户 u_i 在第 k 天第 s 时间段语义位置序列表示为 $Tltrace_{i,k,s}$,将用户 i 所有天的语义位置序列 $Tltrace_i$ 作为用户 u_i 使用第二层算法计算其与全部好友关系强度的输入。

2)基站位置数据处理。将每一个基站视为一个语义位置,即 *Ltrace* = *Trace*,其余处理与 GPS 位置数据处理完全相同。

2.1.3 语义标签数据处理

1)GPS 位置数据处理。对前文得到的序列 *Loc* 中每一个语义位置采用文献[17]中的方法标记其语义标签,标记语义标签后,用户 u_i 第 k 天的语义标签序列表示为 $Strace_{i,k}$,用户 u_i 的全部语义标签序列表示为 $Strace_i$,所有用户的所有语义标签序列表示为 *Strace*,对序列 *Strace* 添加对应的时间标记后记为 *SStrace*,训练对应的 LDA 主题模型并记为 $SLDA(K)$,K 表示主题个数。对每个用户每天的数据按半个小时进行划分,即将用户 u_i 的每天数据 $Strace_{i,k}$ 按时间均分为 48 份,$Sep\_trace_{i,k,s}$ 表示第 i 个用户第 k 天第 s 份数据;对每份数据计算不重复出现的语义标签,并对每个位置对应的语义标签加上时间标记。用户 u_i 在第 k 天第 s 时间段内的语义位置序列表示为 $Tstrace_{i,k,s}$,将用户 i 所有天的语义标签序列 $Tstrace_i$ 作为用户 u_i 使用第三层算法计算其与全部好友关系强度的输入。

2)基站位置数据处理。计算每一个基站对应的语义标签,其余处理与 GPS 数据处理完全相同。

2.2 关系强度计算

计算每一个用户 u_i 与其每一个朋友 u_k($u_k \in F_i$)之间的关系强度,并对 F_i 中的每一个朋友,按照其与 u_i 的关系强度大小降序排列,使此序列中任意两个朋友与 u_i 的关系强弱顺序尽可能与实际情况一致。

1)基于 DTW 及序列熵值加权计算用户之间的关系强度。对用户 u_i 的每一个好友 u_k,利用 2.1.1 节中得到的 $Ntrace_i$ 和 $Ntrace_k$ 计算其轨迹序列相似度。$Ntrace_{i,a}$ 表示用户 u_i 在第 a 天的数据,其中 $a \in D_i$;$Ntrace_{k,b}$ 表示用户 u_k 在第 b 天的数据,其中 $b \in D_k$。$DTW(Ntrace_{i,a}, Ntrace_{k,b})$ 表示用户 u_i 在 a 这一天的轨迹和用户 u_k 在 b 这一天的轨迹的相似度,$Entropy(Ntrace_{i,a})$ 表示用户 u_i 在 a 这一天的轨迹序列的熵值。用户 u_i 和用户 u_k 的基于轨迹序列的关系强度计算方法见式(1)。DTW 计算的是距离,距离越小相似度越大,即该公式值越小,则两个用户关系强度越大。

$$Ent\_Dtw(u_i,u_k) = \sum_{a\in D_i, b\in D_k} S(a,b)\frac{DTW(Ntrace_{i,a},Ntrace_{k,b})}{Entropy(Ntrace_{i,a})} \quad (1)$$

其中,若 $a=b$,则 $S(a,b)=1$;若 $a\neq b$,则 $S(a,b)=0$。

2)基于主题模型计算用户之间的关系强度。$Tltrace_i$ 表示用户 u_i 根据第 2.1.2 节得到的语义位置序列,$Tltrace_k$ 表示用户 u_k 根据第 2.1.2 节得到的语义位置序列。$LLDA(K).inf(Tltrace_{i,a,p})$ 表示对 $Ttrace_{i,a,p}$ 推断得到的主题分布,通常表示为 K 维的向量,其中 K 表示主题的个数。基于用户语义位置的行为模式的关系强度计算方法见式(2),其中 cos 表示余弦相似度。

$$LocLDA(u_i,u_k) = \sum_{a\in D_i,b\in D_i} S(a,b)\sum_{p=q=1}^{48} T(a,p,b,q)\cdot cos\,(LLDA(K).inf(Tltrace_{i,a,p}),LLDA(K).inf(Tltrace_{k,b,q})) \tag{2}$$

其中,若用户 u_i 在 a 这一天第 p 个时间段和用户 u_k 在 b 这一天第 q 个时间段数据均存在,则 $T(a,p,b,q)=1$,否则 $T(a,p,b,q)=0$。

基于用户语义标签的行为模式的关系强度计算公式与基于语义位置的关系强度计算公式相似,见式(3)。

$$SemLDA(u_i,u_k) = \sum_{a\in D_i,b\in D_i} S(a,b)\sum_{p=q=1}^{48} T(a,p,b,q)\cdot cos\,(SLDA(K).inf(Tstrace_{i,a,p}),SLDA(K).inf(Tstrace_{k,b,q})) \tag{3}$$

本研究更关注的是用户和好友 A 的关系强度大于或小于用户与好友 B 的关系强度,因此实际计算结果应为用户与其全部好友按关系强度降序排列得到的好友序列。对于用户 u_i,对其全部好友 F_i 中的每一个朋友 u_k 使用 $Ent\_Dtw(u_i,u_k)$ 计算用户 u_i 和用户 u_k 之间的关系强度,对 F_i 中的每一个朋友按照计算得到的关系强度降序排列得到 $E_i=\{u_{d_1},\cdots,u_{d_{fi}}\}$;在此基础上,使用 $LocLDA(u_i,u_k)$ 计算用户 u_i 和用户 u_k 之间的关系强度,并对 F_i 中的每一个朋友按照计算得到的关系强度降序排列得到 $L_i=\{u_{l_1},\cdots,u_{l_{fi}}\}$;最后使用 $SemLDA(u_i,u_k)$ 计算用户 u_i 和用户 u_k 之间的关系强度,并对 F_i 中的每一个朋友按照计算得到的关系强度降序排列得到 $S_i=\{u_{s_1},\cdots,u_{s_{fi}}\}$。

2.3　结果投票

采用集成学习的思想对三个层次的计算结果 E_i,L_i,S_i 进行投票,投票规则为:对于与用户 u_i 关系第 k 强的好友 u_{v_k}($k\geqslant 1$ 且 $k\leqslant f_i$),使用三个层次对应的方法分别计算得到 u_{d_k},u_{l_k} 和 u_{s_k}。若这三个用户都不相同,则认为 $u_{v_k}=u_{d_k}$,若某个用户比如 $u_{l_k}=u_{s_k}$ 出现两次及以上,则认为 $u_{v_k}=u_{l_k}$,最终以 $V_i=\{u_{v_1},\cdots,u_{v_{fi}}\}$ 作为投票结果。

3　数据集及评估方法

3.1　移动数据集

在实验验证过程中,使用 MIT 媒体实验室采集的 The Reality Mining Data 数据集[1]。实验中使用到的信息主要包括每个用户每天由基站号组成的轨迹序列、所有用户之间的朋友关系以及各个用户的调查问卷,同时数据集中还提供了每个基站号和区域号对应的位置的语义标签。数据集中采集的位置信息是基站信息。虽然基站定位方式的精确度比 GPS 定位方式的低,但其更有利于用户隐私的保护,这也是选择此数据集进行实验的主要原因之一。

在对数据集的分析过程中发现朋友关系信息表中存在如下问题:①部分用户自己和自己是好朋友,另外一部分用户自己和自己不是好朋友;②某用户和另一个用户是好朋友,而另一个用户和该用户不是好朋友。用户之间的好友关系应该满足反自反和对称。经过这样处理后,得到好友数大于 1 的用户共有 34 个,剔除只有一个好友的用户。在后面的实验中,使用这 34 个用户及其全部朋友的数据来对 FRSHV 模型进行验证。

3.2　评估方法与基准

根据前文提到的社会心理学的一些研究成果,态度、兴趣、价值观、背景和人格等方面更相似的人关系更亲密,尤其是对生活在一起的一个群体来说,如果在这些方面类似并且对某些问题的看法相似,则其关系可能就更加紧密。在现实生活当中,通常通过问卷调查方式来获得这些方面的信息,问卷调查结果是这些方面的一种真实体现和反映,因此,问卷调查结果越相似的用户关系越亲密,为此,本文以数据集中问卷调查回答结果的相似性作为朋友之间真实的关系强度。

经过对数据集中问卷调查的仔细分析发现问卷调查中的所有问题基本上可以分为两类:第一类问题可以用“是”或“否”来回答,另一类问题答案多选,但是每个选项按顺序呈现强度增强、次数增加或者次数减少。为了计算用户与朋友之间的真实的关系强度,针对这两类问题,采用不同的评分方法。针对第一类问题当中的每一个问题,如果两个朋友的答案相同,则评分为 1,否则评分为 0;针对第二类问题当中的每一个问题,如果两个朋友的答案越接近,则评分越高,并且将评分归一化到 0 ~ 1 之间,使得每个问题在总的关系强度评分中占有相同的权重。在完成对所有问题评分基

础上,对所有评分进行累加求和,以此作为两个朋友之间的关系强度。依次对每个用户及其所有朋友按上述方法计算其与每个朋友之间的关系强度,并对其所有朋友的评分按降序排列,得到一个用户与其所有朋友之间的关系强度序列,以此序列作为该用户与其朋友之间真实的关系强度。在此基础上,将使用 FRSHV 模型计算出来的用户与朋友之间的关系强度序列与真实的关系强度序列进行对比,验证 FRSHV 模型的有效性。

为了度量使用 FRSHV 模型计算出来的用户与朋友之间关系强度序列 V_i 与真实的关系强度序列 G_i 的一致性,提出一种基于逆序对数的有序序列一致性度量方法。设 A 为一个有 N 个数字的有序集($N>1$),且所有数字均不相同,如果存在正整数 i,j,使得 $1\leqslant i<j\leqslant N$,而 $A[i]>A[j]$,则称 $<A[i],A[j]>$ 为 A 的一个逆序对。A 中全部的逆序对的个数称为逆序对数。我们把序列 G_i 作为有序集,来计算序列 V_i 的逆序对数。设该用户共有 f_i 个好友,若逆序对数为 0,说明实验结果与实际结果完全一致;若逆序对数为 $\frac{f_i(f_i-1)}{2}$,则说明实验结果恰好是实际结果的逆序。提出的有序序列一致性度量公式见式(4),其中 f_i 为用户 u_i 的全部好友的个数,K_i 为 V_i 相对于 G_i 的逆序对数。对每个用户可计算得到一个一致性评分,在此基础上,对所有用户的一致性评分取平均值,以此作为模型 FRSHV 对朋友关系强度度量有效程度的度量,见式(5)。

$$score(u_i)=1-\frac{K_i}{f_i(f_i-1)/2} \tag{4}$$

$$score=\frac{1}{n}\sum_{i=1}^{n}score(u_i) \tag{5}$$

4 实验验证及分析

实验环境为 windows 7 64 位,4 核,3.2 GHz 主频,8 G 内存,使用 Python 编码实现。

因为用户之间的物理距离难以直接确定,所以以基站之间的距离作为用户之间的物理距离。采取如下方法来定义基站之间的距离:将每天用户手机连接过的基站视为一条基站序列,对于基站 A 和 B,从所有用户所有天的基站序列中找到同时出现 A 和 B 的序列,计算每个序列中 A 和 B 中间不同的基站号的个数,取最小值加 1 作为基站 A 和基站 B 之间的距离。若通过上述方法能够计算出两个基站之间的距离,则称这两个基站之间的距离存在。若 A 和 B 从未在同一个基站序列中出现过,则定义 A 和 B 之间的距离为所有两个基站之间最大距离的 K 倍(K 为一个正实数参数,在后面实验中能够看到该参数对实验结果的影响)。

4.1 基于轨迹相似性计算用户之间的关系强度

通过上文对基站距离的定义,使用 DTW 以及归一化后的 DTW 计算第一层用户之间的相似度,一致性评分可通过式(4)和式(5)计算得到,上文论述到使用参数 K 定义两个不存在距离的基站的距离,不同的参数 K 以及不同方法对结果的影响见图 2。通过观察实验结果发现,取不同的 K 值会产生不同的结果,当值取得很大时候,意味着如果这两基站之间不存在距离,在实际处理过程中则认为两个基站之间距离比较大,这样会得到更理想的结果。

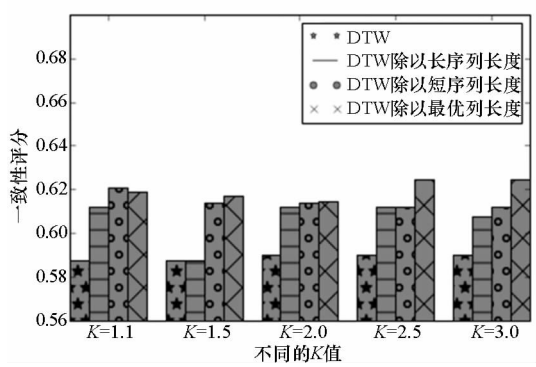

图 2 参数 K 对一致性评分结果的影响

Fig. 2 Influences of K on the consistency

在上一个实验的基础上,对 DTW 方法以及归一化的 DTW 方法使用序列熵值加权,对应 2.2 节的 E_i,并以编辑距离[18]计算的结果作为基准,一致性评分的实验结果见图 3。

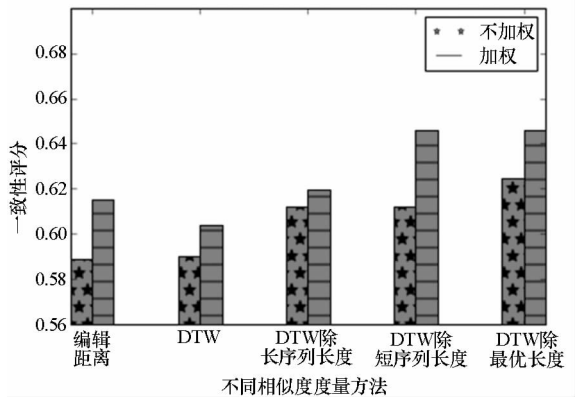

图 3 用熵值加权前后结果对比($K=2.5$)

Fig. 3 Comparison of weighted consistency and non-weighted consistency ($K=2.5$)

4.2 基于语义位置相似性计算用户之间的关系强度

在计算关系强度的过程中,使用 LDA 模型进

行推断。因为推断过程进行随机初始化,使得 LDA 模型的每次执行结果不一定完全相同,因此,在实验中,针对每个不同的参数值(即主题个数)执行 10 次,并将每次计算获得的 L_i 与 G_i 进行一致性评分。对所有用户按式(5)计算最终的一致性评分,进而取这 10 个一致性评分的中位数作为该参数对应的一致性评分,如图 4 所示。

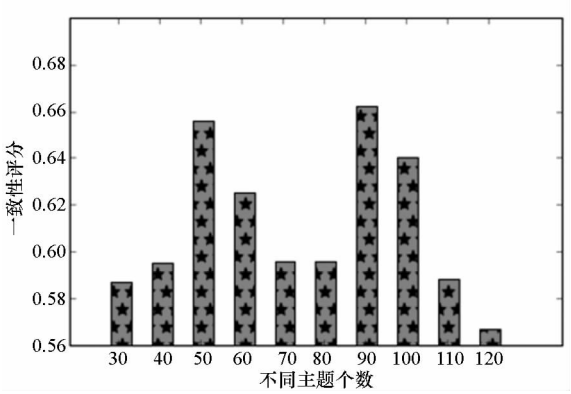

图 4　使用语义位置时主题个数及对应的一致性评分实验结果

Fig. 4　Influence of topic numbers to consistency using semantic location

4.3　基于语义标签相似性计算用户之间的关系强度

数据集中提供了基站号和区域号对应的位置的语义标签[1],对所有语义标签加上时间标记,将每个带时间标记的语义标签视为单词,每天的语义标签序列视为句子,每个用户所有语义标签序列视为文档,使用所有用户的全部文档对 LDA 模型进行训练,其实验过程与上面的基于语义位置的实验过程一样,最后对对应的 2.2 节所示的 S_i 进行一致性评分。图 5 展示了在主题个数取不同值时所对应的一致性评分结果。

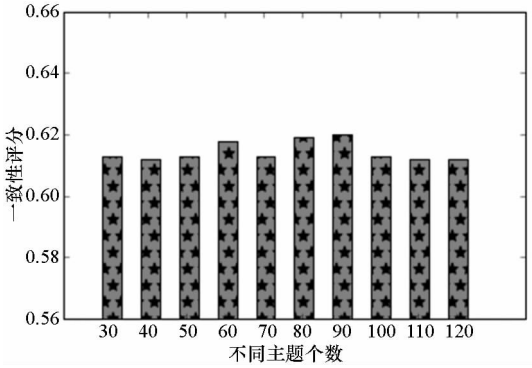

图 5　使用语义标签时主题个数和对应一致性评分的实验结果

Fig. 5　Influence of topic numbers to consistency using semantic label

语义标签有实际含义,以主题个数 75 为例,通过观察 LDA 模型学习到的主题,发现该模型学习得到了 3 个主题,如表 1 所示,主题 1 表示的是晚上在实验室或教室,主题 2 表示早上和晚上在家,主题 3 表示上午在实验室。

表 1　LDA 模型学习到的不同主题示例

Tab. 1　Some topics of LDA learned

主题 1	主题 2	主题 3
Tech sq_47	home_14	Media lab_17
Tech sq_46	home_15	Media lab_16
Tech sq_40	home_8	Media lab_20
Tech sq_38	home_6	Media lab_18
Tech sq_39	home_0	Media lab_19
Tech sq_42	home_44	Tech sq_17

4.4　对计算结果进行投票

上面的实验分别描述了层级模型 FRSHV 每一层的实验结果,在此基础上,使用前面描述的投票规则对三层中每层最好的实验结果进行投票,三层结果投票的实验结果见图 6。

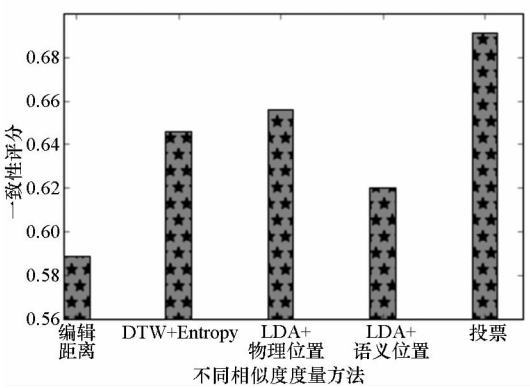

图 6　投票结果及分别只使用一种方法的结果对比

Fig. 6　Results comparison between vote method and simple method

通过实验结果可以发现,使用投票方法后可以更好地度量用户之间的关系强度,基于投票的方法比编辑距离一致性评分高出近 10%。

5　结论

针对如何度量日常生活中人们之间的关系强度问题展开研究,提出了一个从日常轨迹、语义位置以及语义标签三个层次度量用户与朋友之间关系强度的层级模型 FRSHV。采用基站数据对该模型进行了验证,观察实验结果发现基于投票的方法比编辑距离一致性评分高出近 10%。下一步我们将对相关度量方法进行进一步的优化,利

用更多的消息如通话记录、短信等,进而对多种数据进行融合来度量用户之间的关系强度。

参考文献(References)

[1] Eagle N, Pentland A. Reality mining: sensing complex social systems [J]. Personal and Ubiquitous Computing, 2006, 10(4): 255 – 268.

[2] Aharony N, Pan W, Ip C, et al. Social fMRI: investigating and shaping social mechanisms in the real world [J]. Pervasive and Mobile Computing, 2011, 7(6): 643 – 659.

[3] Wang R, Chen F L, Chen Z Y, et al. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones [C]//Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing, 2014: 3 – 14.

[4] Stopczynski A, Sekara V, Sapiezynski P, et al. Measuring large-scale social networks with high resolution [J]. PloS One, 2014, 9(4): e95978.

[5] Granovetter M S. The strength of weak ties [J]. American Journal of Sociology, 1973, 78(6): 1360 – 1380.

[6] Wegner D M. The illusion of conscious will [M]. Cambridge, MA, US: MIT Press, 2002.

[7] Burrows R, Nettleton S, Pleace N, et al. Virtual community care? Social policy and the emergence of computer mediated social support [J]. Information, Communication & Society, 2000, 3(1): 95 – 121.

[8] Petróczi A, Nepusz T, Bazsó F. Measuring tie-strength in virtual social networks [J]. Connections, 2007, 27(2): 39 – 52.

[9] Ma C, Cao J N, Yang L, et al. Effective social relationship measurement based on user trajectory analysis [J]. Journal of Ambient Intelligence and Humanized Computing, 2014, 5(1): 39 – 50.

[10] Festinger L. A theory of cognitive dissonance [M]. CA, USA: Stanford University Press, 1962.

[11] Zajonc R B. Attitudinal effects of mere exposure [J]. Journal of Personality and Social Psychology, 1968, 9(2p2): 1 – 27.

[12] Itakura F, Umezaki T. Distance measure for speech recognition based on the smoothed group delay spectrum [C]//Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 1987: 1257 – 1260.

[13] Ratanamahatana C A, Keogh E. Everything you know about dynamic time warping is wrong [C]//Proceedings of Third Workshop on Mining Temporal and Sequential Data, 2004.

[14] Singelis T M. The measurement of independent and interdependent self-construals [J]. Personality and Social Psychology Bulletin, 1994, 20(5): 580 – 591.

[15] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3: 993 – 1022.

[16] Farrahi K, Gatica-Perez D. What did you do today? Discovering daily routines from large-scale mobile data [C]//Proceedings of the 16th ACM International Conference on Multimedia, 2008: 849 – 852.

[17] Yang R S, Shi D X. SASLL: a system annotating semantic label of location [C]//Proceedings of the 7th International Symposium on UbiCom Frontiers Innovative Research, Systems and Technologies, 2015.

[18] Levenshtein V I. Binary codes capable of correcting deletions, insertions, and reversals [C]//Proceedings of Soviet Physics Doklady, 1966, 10(8): 707 – 710.