**DOI**:10.11992/tis.201511026
**网络出版地址**:http://www.cnki.net/kcms/detail/23.1538.TP.20160315.1239.010.html

# 基于用户移动轨迹的个性化健康建议推荐方法

陈万志[1],林澍[1],王丽[2],李冬梅[2]
(1.辽宁工程技术大学 电子与信息工程学院,辽宁 葫芦岛 125105; 2. 渤海装备辽河重工有限公司,辽宁 盘锦 124010)

**摘 要**:随着移动智能终端的普及,移动医疗应用已成为当前研究的热点。针对移动医疗环境下个性化健康建议推荐问题,依据用户移动轨迹与职业类型间相似性特点,提出一种基于驻点区域特征向量与用户职业特征向量相结合的相似度计算方法,通过构建相似用户组的方式完成组内用户健康建议信息的共享,最终实现在节约医疗资源的基础上为海量用户提供个性化健康推荐服务的功能。算法测试与分析结果表明了方法的有效性和可实施性,在移动医疗大数据分析应用方面具有广阔的前景和实用价值。

**关键词**:移动医疗;大数据分析;移动轨迹;特征向量;个性化推荐
**中图分类号**: TP311 **文献标志码**:A **文章编号**:1673-4785(2016)02-0264-08

**中文引用格式**:陈万志,林澍,王丽,等. 基于用户移动轨迹的个性化健康建议推荐方法[J]. 智能系统学报, 2016, 11(2): 264-271.
**英文引用格式**:CHEN Wanzhi,LIN Shu,WANG Li, et al. Personalized recommendation algorithm of health advice based on the user's mobile trajectory[J]. CAAI transactions on intelligent systems, 2016, 11(2): 264-271.

# Personalized recommendation algorithm of health advice based on the user's mobile trajectory

CHEN Wanzhi[1], LIN Shu[1], WANG Li[2], LI Dongmei[2]
(1. School Electronics and Information Engineering, Liaoning Technical University, Huludao 125105, China; 2. China Petroleum Liaohe Equipment Company, Panjin 124010, China)

**Abstract**: Mobile medical applications have become a hotspot in research with the popularization of mobile intelligent terminals. In response to the problem of personalized recommendation of health advice in the mobile medical environment, a similarity calculation method based on stagnation region eigenvector and user occupation eigenvector was proposed according to the similarity characteristics of users´ mobile trajectory and occupation. The sharing of information about the suggestion of health in a group was completed by constructing similar user groups. Thus, personalized health recommendation services were provided for users with limited medical resources. Results showed the effectiveness and implementation of the algorithm, which has a broad application prospect and practical value in large data analysis and mobile medical application.

**Keywords**: mobile medical; large data analysis; mobile trajectory; feature vector; personalized recommendation

随着 GPS、移动基站、室内等定位技术与方法的发展和普及,越来越多的用户将其移动记录分享到朋友圈中实现社交行为,同时各种基于位置的服务已逐渐成为人们日常生产生活中不可或缺的要素。因此,对用户移动轨迹的分析与挖掘已成为当前行为分析与挖掘方面的研究热点。

用户的 GPS 轨迹序列记录的是其在真实物理世界中移动路线,在一定程度上蕴含着用户的个人意图、喜好以及行为模式。如何挖掘轨迹中的知识,实现从个体数据中挖掘出用户行为、意图、经验和生

**收稿日期**:2015-11-25. **网络出版日期**:2016-03-15.
**基金项目**:辽宁工程技术大学博士基金项目(2015-1147).
**通信作者**:陈万志. E-mail:chenwanzhi@lntu.edu.cn.

活模式[1-6],融合群体数据来发现热点地区和经典线路[7],甚至挖掘人和人之间的相关性[8-9]及个体在地域之间的活动模式[10-11]等等都具有十分重要的现实意义。特别是在云计算和大数据分析背景下,以海量用户轨迹数据分布式云存储为基础,对用户移动轨迹数据进行深度的分析和挖掘,完成"数据-信息-知识-智能"的计算过程,实现更深层次、更人性化、更有效的为用户提供基于位置的增值延伸性服务[12]。

现代社会人类复杂的社会交际和迁移活动使得人与社会、人与自然环境的关联复杂性突显,自然和社会环境均可影响人体健康,因此关注用户的身体和心理健康是同等重要的。通过用户移动行为轨迹及周边区域特征等组成的用户社会行为相关信息,研究如何在用户移动轨迹数据与健康体征信息密切关联的海量空间数据分布式云存储和计算架构的基础上实现为用户提供个性化的心理健康建议推荐信息服务是大数据分析领域性应用中心理疾病防治与个性化健康指导方面的主要难点问题。

GPS 采集的数据是用户移动行为轨迹数据的主要来源,其中采样精度与采样频率对后续的分析有较大影响,存在干扰因素的数据直接用于用户数据挖掘时往往得到的不是预期的效果。对于采样误差的问题而言,一般地,民用 GPS 定位精度在米级,在某些道路稠密的地段的误差将使用户的当前位置映射到错误的道路上导致影响用户的定位与导航功能,比较成熟的地图匹配(map matching)的方法是将带有位置偏差的 GPS 轨迹映射到正确的道路上,从而实现导航质量的提高,但是如何实现采样误差 GPS 数据的用户社会行为分析与挖掘尚无相关的研究和方法。而对于采样频率低的问题而言,直接简单的提高采样率方法来处理由于实时获取 GPS 设备位置信息的通信和存储代价限制而无法实施。文献[2]针对这种低采样率的 GPS 轨迹提出了一种基于全局信息的匹配方法,通过分析"存疑点"周围"确定点"的位置信息与关系信息,从而确定"存疑点"可能出现的大致位置,这种处理方案可以实现各采样点关联的过渡性轨迹平滑和底图匹配,但如何在用户移动速度不稳或 POI(point of interest)信息点稠密等情况下有效地动态描述用户移动轨迹,进而挖掘其所处驻点区域的社会行为特性还无法得到解决。

采集用户移动轨迹时用户的交通出行方式可能是不同的,因此,若能够从用户的移动轨迹信息中挖掘出轨迹采集时用户的出行模式,则对用户的分类和信息推荐是有辅助意义的。当前基于 GPS 轨迹的交通工具判别主要依据轨迹序列中时间序列与位移距离计算得到的平均速度来实现,但在交通工具的速度不均衡,特别是城市交通状况的影响导致这种判断方法的识别精度小于 50%。另外用户在两次轨迹采集点的区间如变换交通方式,则使得同一段移动轨迹可能会由多种交通方式所构成,若不能每个位移区间的交通方式原子化,必然导致判断结果也会包含不可避免的错误。

文献[4-6]针对上述问题提出的解决方法是:采用一种有效的路线分割方法,其主要的思想是利用步行路段来分割轨迹;通过发现一些受交通状况影响不大的特征,如方向改变率等,并结合监督学习的方法来训练一个分类模型;采用一种后处理方法,从大量的线路中挖掘出一个隐含的地图,并分析了不同路段上各种交通工具的使用概率和交通工具之间的转移概率。因此,巧妙地利用了自然常识、地理限制和地图信息来修正错误的判别。

用户历史轨迹中出现的频繁模式反映了个人的生活习惯和行为规律,若能够从轨迹中推理挖掘出这些知识,服务提供商将会为用户提供更深入、更个性化的位置服务[3]。而要从轨迹中挖掘这些频繁模式首先要解决的问题是如何对个人的历史轨迹进行建模。可以通过算法检测出该用户停留过的有效位置,一个用户的历史轨迹就可以基于这些位置表达为一个停留位置序列,这样既可以挖掘出用户行为的重点,同时也大大减轻了数据处理量。更进一步讲,由于用户多次访问同一地点所产生的停留点由于 GPS 数据偏差原因可能不一定是完全一致的,因此直接对停留点进行对比并是不可行的,这就需要对从轨迹中提取出来的停留点进行聚类分析。将相近的停留点分配到同一个聚类中,此后再用各个停留点所归属的聚类来替换该停留点,进而将停留点序列进一步转化为相应的聚类序列,最终使得用户在不同时间段的历史轨迹可以进行对比。在用户历史轨迹的模型基础上,可采用 FP-growth、Closet+等算法来挖掘其中的频繁项集,并且这些频繁模式是可以相互组合和连接的,从而可进一步挖掘出一些表征了用户生活、行为规律的顺序模型。

综上所述,现有基于位置的服务一般都是直接通过用户提供的位置数据进行处理,缺乏对这些数据的进一步分析和挖掘,忽略了这些位置数据中蕴含的信息和知识。因此,研究以云存储和计算技术为基础,通过位置信息与地图 POI 兴趣点数据的融合实现对用户轨迹数据更深入的分析和挖掘得到更丰富的知识,最终达到更智能的为用户推送个性化健康建议信息服务的目的。

## 1 用户社会行为数据的相似分组算法

算法的主要思路是通过对采集得到的用户每日上下班 GPS 轨迹数据进行分析,按照用户工作时间、城市、地点、交通工具等用户特征进行个性化建模,并对此模型进行量化分析处理,计算用户在社会工作方面的相似程度,最终将具有相似工作环境和工作压力的用户构建为基于用户社会行为数据的相似用户组,进而实现相似用户组内的职场心理与健康指导信息共享,达到个性化推荐的目的。其工作流程如图 1 所示,首先通过用户的多条 GPS 轨迹数据结合地区 POI 数据库确定用户的工作类型,然后根据 GPS 轨迹的时间序列和用户使用交通工具情况折算出用户工作压力并分组。为了获得用户的工作类型并使得用户的历史位置具有可比较性,拟利用用户驻点区域特征向量描述 POI 数据与用户空间位置的语义联系。此外,还通过有限的 GPS 数据信息推断用户可能的收入和工作时间特征,并形成可表示用户工作压力的向量,最终通过用户工作类型和向量夹角实现对用户的分类。

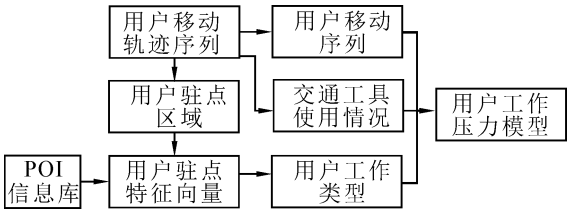



**图 1 算法的工作流程**

**Fig.1 Algorithm of workflow**

### 1.1 用户轨迹处理

用户的移动轨迹信息通常是 GPS 移动终端采集并处理得到,其中包含采样点的坐标信息、采集时间以及位移速度等,属于瞬时离散型数据,在每次采集后采集设备均有一段静默期;在实际采集过程中由于环境复杂导致所采集的时间和坐标具有一定的不准确性,数据往往有一定的时间和坐标偏差。因此用户移动轨迹数据用于判断用户驻点区域特征是要充分考虑定位偏差与精度对结果的影响,移动终端采集的用户日常上下班的 GPS 轨迹数据采用下述方法处理得到用户的工作轨迹信息。其中重要的定义包括:

GPS 轨迹: $GPS_t$ 是一系列与时间相关的 GPS 轨迹点的序列 $GPS_t = (P_1, P_2, P_3, \cdots, P_n)$,其中 GPS 轨迹点 $P_i = (x, y, t)\ (1 \leqslant i \leqslant n)$,其中 $(x, y)$ 分别表示采集数据的经度和纬度、$t$ 表示采集数据的时间且满足条件 $P_i \cdot t < P_{i+1} \cdot t (1 \leqslant i \leqslant n-1)$。

驻留区域:GPS 驻留区域 Sz 指的是一组在一定时间内,相邻或相近的 GPS 轨迹点的集合 $Sz = (P_i, P_{i+1}, \cdots, P_j)$,满足条件 $Dist(P_i, P_k) \leqslant \theta_d$, $Dist(P_i, P_j) > \theta_d$, $Int(P_i, P_j) \geqslant \theta_t$,其中 $\theta_d$ 为驻留区域直径,$\theta_t$ 为驻留区域时长,$Dist(P_i, P_j)$ 为 $(P_i, P_j)$ 两点间的欧氏距离,$Int(P_i, P_j)$ 为 $P_i \cdot t$ 与 $P_j \cdot t$ 时间间隔,$i \leqslant k < j$。

用户驻点:用户驻点 $Sp = (x, y, t_{in}, t_{out})$ 指的是驻留区域的几何中心,其中 $Sp \cdot x = \sum_{k=j}^{i} P_k \cdot x / |P|$, $Sp \cdot y = \sum_{k=j}^{i} P_k \cdot y / |P|$, $Sp \cdot t_{in} = P_i \cdot t$, $Sp \cdot t_{out} = P_j \cdot t$ 且 $P_k \in Sz$。

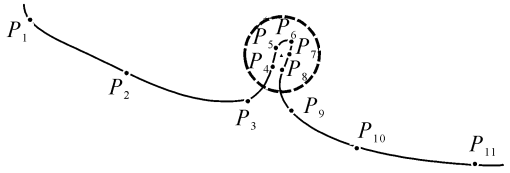



**图 2 一组 GPS 轨迹**

**Fig.2 A GPS track**

如图 2 中 $\{P_1, P_2, \cdots, P_{11}\}$ 为智能终端采集得到的一组 GPS 轨迹,则 $4 \leqslant k < 9$ 时有 $Dist(P_4, P_k) \leqslant \theta_d$, $Dist(P_4, P_9) > \theta_d$, $Int(P_4, P_8) \geqslant \theta_t$,故驻留区域应从 $\{P_4, P_5, P_6, P_7, P_8\}$ 计算,并得到图中以三角号表示的用户驻点 Sp。驻留区域直径和驻留区域时长的取值应该根据用户所在地区的环境特点进行设置,如当用户活动范围位于城市中心时,驻留区域直径应设置在 200 m 左右为宜,而当用户活动范围位于城乡结合部或者远离城市时,驻留区域直径应设置在 500 m 左右为宜,与此同时在大多数情况下驻留区域时长应保证大于半小时,以上设置有利于更加准确地从众多轨迹信息中找到可描述用户工作地点的用户驻点。

通过以上定义可知,从用户移动终端采集的 GPS 数据中提取到一些关键信息,通过计算用户从一个驻点到另一个驻点的时间差得知用户的行程时间(如上班时间),也可以通过计算用户在驻点内的驻留时间差得知用户驻留时间(如工作时间)。当然用户驻点与现实生活中具体地点的关系仍需要进一步确认。

### 1.2 用户在驻点区域的行为建模

用户驻点可以粗略地表明用户每次移动的起止地点以及移动的起止时间,结合 POI 数据可以进一步得到用户移动起止地点的详细信息,为驻留区域语义提取提供数据基础。每条 POI 数据内容包含信息点名称、类别、经度和纬度及其他说明等相关地理信息。

用户驻点与 POI 数据往往不能简单地通过距离计算的方式建立关系。由于驻留区域直径和驻留区域时长的不当选取或实际生活中突发的事件,导致用户在移动过程中的某些位置上停留了一段时间,产生了非目的地或出发地的驻点,如十字路口或车站等;考虑到 GPS 定位误差和城市中密集分布的信息点,通过 POI 数据识别用户在驻点处访问的确切地点成了不可能完成的任务。一个 GPS 采集点数据可以具有 10 m 或以上的位置偏差,而在其周围可以有多种的 POI 数据,而距离用户驻点最近的 POI 数据所代表的信息点可能不是用户真实访问的地方,如在有些地方餐厅、商场和电影院重叠在同一建筑物内,实际应用如图 3 所示。

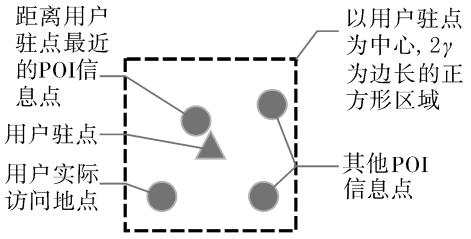



**图 3　用户驻点与 POI 信息点**

**Fig.3　Users stagnation and POI information point**

为了解决用户驻点与 POI 数据的内容的关联问题,将用户驻点所代表的 POI 数据的驻点区域表示为

$$[s\cdot x-\gamma, s\cdot x+\gamma]\times[s\cdot y-\gamma, s\cdot y+\gamma]$$

式中: $\gamma$ 是一个与 GPS 相关的参数定位误差,则可采用 TF-IDF 的统计方法构建了一组特征向量表示每个驻点区域所代表的兴趣点内容,用以评估字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加,但同时会随着它在语料库中出现的频率成反比下降。同理,应用 TF-IDF 算法时把词的类别和用户驻点区域中兴趣点作为文档处理。直观地说,如果同一类兴趣点发生在同一个地区的频率较高,这该种类是该地区的典型代表。此外,有些 POI 类型(如“博物馆”和“公园”)由于在城市建设的数量较少且分散,因此此种类型的兴趣点在城市中出现的概率很少,而一些常见功能的兴趣点(如“餐馆”)可以遍布在城市的任何角落。对于个人而言,由于存在一定的生活规律和做事目的,在一定的生活规律条件下,每日访问的地点可能有所不同但目的应该相近或相同,因此在判断哪些兴趣点是用户真正的移动目的地时应重点分析那些经常存在于用户驻点区域的 POI 类型。综上,需要考虑一个 POI 类别在一个区域发生频率和其逆向文件频率两种因素,故引入特征向量的定义

特征向量:用一组向量 $\boldsymbol{f}_r=(w_1,w_2,\cdots,w_n)$ 表示用户驻点区域中兴趣点内容的集合,有

$$w_i=\frac{n_i}{N}\times\log\frac{|R|}{|r|} \tag{1}$$

式中:$n_i$ 是该区域属于第 $i$ 种类别的 POI 的数量和,$N$ 是位于该区域的 POI 的总量,$R$ 为用户驻点区域总数,$r$ 表示出现第 $i$ 种 POI 类别的用户驻点区域数量。式(1)的第 1 部分表示一个类别 $i$ 发生频率,第 2 部分表示一个类别 $i$ 在该用户整个驻点区域中 POI 类型总数 $|R|$ 的逆向文件频率。

根据式(1)用一个特征向量代表一个驻点区域,虽然还不能确切地确定用户访问的地点,但此特征向量在一定程度上可以代表用户位置的语义含义,即该区域中具有哪些突出的 POI 类别,进而可以推断该区域的功能类别。

为了进一步推断出用户在驻留区域中的动作行为(如工作、用餐、访友、居住等),还需要将该区域的功能类别与用户在该区域的驻留时间相结合,并以 24 h 为周期,对每天用户的驻点区域进行比较,即可筛选出用户日常工作的区域和用户临时停留的区域。考虑到目前存在着一些在非固定地点工作的用户,对于其中大多数用户来说,其所从事的工作性质往往相同,因此在对用户工作地点的查找方法上,使用基于特征向量的比较方式要比使用基于坐标位置的比较方式更具说服力。

利用余弦相似性原理可以对前述 TF-IDF 计算方法产生的特征向量进行相似度比较,其原理为计算求得两组向量的夹角,并得出夹角对应的余弦值,用来表征这两个向量的相似性。夹角越小,余弦值越接近于 1,它们的方向更加吻合,特征越相似。其计算式为

$$\cos\theta=\frac{\sum_1^n(f_a\cdot w_i\times f_b\cdot w_i)}{\sqrt{\sum_1^n f_a\cdot w_i^2}\times\sqrt{\sum_1^n f_b\cdot w_i^2}} \tag{2}$$

由式(2)可以从用户驻点含义层面对同一用户出现的地区进行比较判断哪些区域可能是用户的工作区域,同时也需要从时间层面对同一用户的上班规律进行比较,进一步确定哪些区域是用户的工作区域。国内采用的标准工作时间制度是指职工每日工作 8 h,每周工作 40 h 的工时制度,不同地区、不同职业会导致工作时间有所变化,但大部分工作每日的作息时间相差不大,尤其是上班时间。与此同时也应考虑到采用轮休或倒班制度的工作,因此在计算时间相似性时法定工作日与节假日是不区分

的。在判断某些相似驻点区域是否为用户工作区域的主要依据是用户进入到达该驻点的时间序列是否可以收敛到一个或几个时间点上,若可以找到进入驻点时间一定或偏差不大且驻点特征向量相似的驻点,则此驻点区域可认为是用户工作区域,此驻点特征向量包含用户的工作信息。

## 2 用户社会行为数据的相似度比较

基于用户社会行为数据的相似度用于衡量用户工作职业和工作环境等与用户职场活动相关的信息,包含用户工作时间、城市、工作类型、通勤工具使用情况等,通过对这些信息的处理从用户经济条件、工作类型、工作压力等方面对用户进行建模,进而通过相似度比较方法实现相似用户聚类。

1)用户经济条件

用户经济条件主要依赖于用户的收入和支出,可以间接通过用户所在城市和用户乘坐通勤工具的情况进行分析,通过城市的平均收入可以简单区分用户的收入等级;文献[13-14]指出在城市中若弱化用户住所与用单位间距离影响,用户的平均收入情况与用户通勤所选用的交通工具情况服从线性分布,因此考察用户乘坐通勤工具可以进一步地划分用户收入的等级,使得用户收入等级明确化。通过国内近 5 年的国内城市收入排名统计,将用户按照所在城市的不同进行收入的划分,进而使得不同城市间用户的收入情况得以比较。通过城市中乘坐不同通勤工具的人口比例进一步对收入进行划分,并最终计算出用户的经济条件指数。其计算公式为

$$\mathrm{Ueci} = \frac{A_i}{1\,000} \times Vr_{\mathrm{type}_i} \tag{3}$$

式中:$A_i$为平均收入, 通勤工具比例系数 $Vr_{\mathrm{type}_i}$ 根据用户乘坐通勤工具可分为以下 3 种类型:

①私家车或公务轿车:

$$Vr_{\mathrm{type}_1} = \frac{\frac{Vr_1}{2} + Vr_2 + Vr_3}{Vr_1 + Vr_2 + Vr_3} \tag{4}$$

②单位通勤车:

$$Vr_{\mathrm{type}_2} = \frac{\frac{Vr_2}{2} + Vr_3}{Vr_1 + Vr_2 + Vr_3} \tag{5}$$

③公共交通工具或步行:

$$Vr_{\mathrm{type}_3} = \frac{\frac{Vr_3}{2}}{Vr_1 + Vr_2 + Vr_3} \tag{6}$$

通过式(3)~(6)可线性的将用户所在城市和所选通勤工具映射为用户经济条件指数,依据通勤工具类型可将用户的收入分为 3 个等级并核算成数值,并采用各地平均收入作为区域计算的因子更恰当地表明不同城市间的收入差。用户经济条件指数相关信息的实例关系如表 1 所示。

**表 1 城市收入与通勤工具**

**Table 1 Urban income and vehicle**

| 城市等级 | 对应城市或地区 | 平均收入/元/月 | 通勤工具比例系数 |
|---|---|---|---|
| 1 | 一线城市,如北京、上海、广州、深圳、南京等 | 6 000 左右 | 18∶39∶43 |
| 2 | 二线城市,如宁波、福州、厦门、长沙、大连等 | 4 000 左右 | 27∶36∶37 |
| 3 | 三线城市,如海口、佛山、泉州、东莞、南宁等 | 3 000 左右 | 24∶40∶36 |
| 4 | 四线城市,如唐山、秦皇岛、邯郸、保定、廊坊等 | 2 000 左右 | 16∶45∶39 |
| 5 | 其他城市 | 1 200 左右 | 8∶52∶40 |

2)用户工作类型

用户工作类型可由用户工作区域驻点的特征向量中处理得到。由于城市建设规划时常将功能相近的建筑建于相近区域中,如商业区、居民区、工业区等,因此实际处理而得的特征向量所包含的具有信息点较多的 POI 类型,其功能类型往往是相同或相互辅助的关系,其反映到现实生活中的结果就是在此工作的人群具有相似的工作习惯和作息时间,即同一工作类型。用户驻点区域特征向量的趋势向量 POI 类型总体概括如表 2 所示。

**表 2 工作类型与平均工作时长**

**Table 2 Type of work and average hours worked**

| POI 类型 | 工作类型 | 平均工作时长/h |
|---|---|---|
| 医疗、维修服务 | 应急服务业 | 9 |
| 餐饮、汽车服务、生活服务、购物、休闲娱乐、旅游景点 | 普通服务业 | 8.5 |
| 行政地标、政府机构 | 政府机关 | 8 |
| 公司企业、金融 | 企业 | 10 |
| 工厂 | 工厂 | 8 |
| 教育 | 教育 | 7 |

3)用户加班时长

用户加班时长等于平均工作时长除以城市平均工作时长,将用户的工作时长与该城市的平均工作时长或该用户所在工作类型的平均工作时长作比较,可计算出用户每天加班情况,这在一定程度上反映出用户的工作压力。用户加班时长与用户加班指

数 Uoti 之间的对应关系如表 3 所示。

4)基于用户社会行为数据的相似度比较

综合上述 3 个方面分析,由于在工作类型相同时用户的工作压力和工作时间具有可比性并能反映用户在其行业中的压力情况,首先将用户按照工作类型进行分类,并根据利用余弦相似性原理计算各分类中通过经济条件指数和加班指数形成的向量之间的夹角,最终确定哪些同类工作的用户相似,进而形成相似用户组。即

$$\cos\omega = \frac{\mathrm{Ueci}_1 \times \mathrm{Ueci}_2 + \mathrm{Uoti}_1 \times \mathrm{Uoti}_2}{\sqrt{\mathrm{Ueci}_1^2 + \mathrm{Uoti}_1^2} \times \sqrt{\mathrm{Ueci}_2^2 + \mathrm{Uoti}_2^2}} \quad (7)$$

式中: $\cos\omega$ 可描述用户的工作压力情况,当用户收入较高且加班时间较少时该值会趋近于 1,即表示用户工作压力较小;而当用户收入较低且加班时间较长时该值会趋近于 0,表示用户工作压力较大。

**表 3　用户加班时长与 Uoti 的对应关系**

**Table 3　Relationship with the user overtime and Uoti**

| 用户加班工作时长/h | 用户加班指数 |
|---|---|
| 小于 0.5 | 0 |
| 0.5-1 | 1 |
| 1-2 | 2 |
| 2-3 | 3 |
| 大于 3 | 4 |

## 3　算法测试与结果分析

基于用户社会行为数据的相似分组算法测试采用微软亚洲研究院的 Geolife Data v1.3[15] 用户轨迹数据集,其涵盖 182 个用户在 2007 年 4 月—2012 年 8 月间的部分出行轨迹记录,其中不但包括 GPS 轨迹信息,如标记点时间序列、采集点纬度、经度和高度的信息序列,而且数据集中还包括 73 个用户由不同的 GPS 采集设备记录并有多种采样频率的17 621 条轨迹记录,其中 91.5% 的轨迹记录的采集频率比较密集,采集时间间隔每 1~5 s 或每行进 5~10 m 一次,并标记了出行交通工具,如驾车、乘公共汽车、骑自行车和步行。为了能够与 POI 底图数据相对应,选取 Geolife 数据集中北京地区的轨迹信息,与百度地图 API 导出北京地区的 POI 数据相对应。由于 Geolife 数据集中没有包含用户实际的工作压力情况,因此无法对算法得出的分组结果进行量化分析,但其中的有些部分可与现有文献资料中的算法对比实现量化分析,因此算法的测试与结果分析分成两部分进行。

### 3.1　用户工作类型分组测试与分析

首先在进行算法测试前将标记有出行交通工具的相关用户轨迹进行整理,利用人工判别的方式根据每个用户的出行轨迹起点与终点判断出用户可能的大致职业类型,并记录此职业信息为用户属性;其次对这些标记有出行交通工具的轨迹进行基于用户社会行为数据的相似分组计算,计算时根据文献[16-17]的研究结论并考虑到所在城市的特点将算法中的相关系数分别设置为:驻留区域直径 $\theta_d$ = 200 m ,驻留区域时长 $\theta_t$ = 30 min ,GPS 定位误差参数 $\gamma$ = 200 m ,依据用户驻点区域相似度 $\cos\theta$ 的阈值进行用户驻点区域相似度的用户分组;最后将分组结果与人工判别标记的职业类型相比较,通过分类的准确率来评价此部分算法的优劣程度。算法的处理流程如图 4 所示。

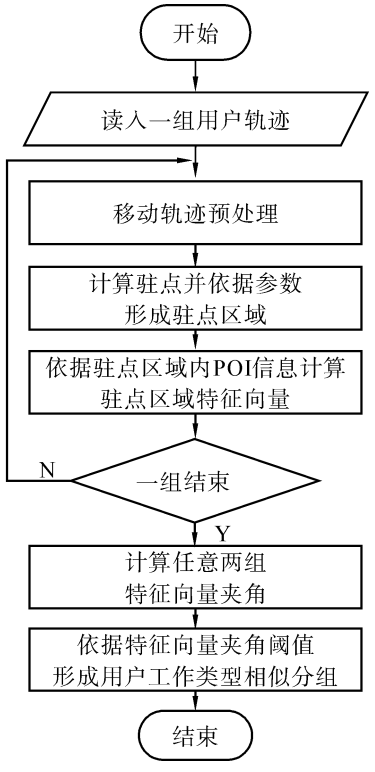



**图 4　用户工作类型分组流程图**

**Fig.4　The user type grouping flow chart**

为测出用户驻点区域相似度 $\cos\theta$ 阈值的最佳值,在其他条件不变的情况下仅改变 $\cos\theta$ 阈值,最终得出的用户工作类型分组准确率如图 5。由结果可知对于本次使用的测试数据而言,当 $\cos\theta$ = 0.85 时所得到的用户工作类型分组准确率达到最高。

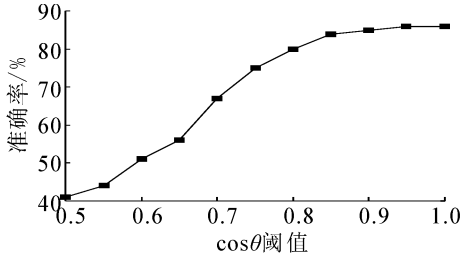



**图 5　cos θ 与分组准确率的关系**

**Fig.5　Relations between the packet accuracy rate and cos θ**

在此基础上,令 $\cos\theta = 0.85$ 时通过文献[18]的 HGSM 算法、文献[19]的 SLH-MTM-L2 算法($2^{nd}$ layer)与本文方法在分类准确度和运行时间两方面进行比较,结果如表 4 所示。由结果可知本文提出的用户工作类型分组方法基本可以保证准确的用户分组,且运算耗时较少,在用户记录项繁多且需要实时计算的情况下优势明显。

**表 4 对比实验结果**

**Table 4 Contrast test results**

| 算法 | 用户分组率/% | 分组准确度/% | 计算耗时/ms |
|---|---|---|---|
| HGSM | 93.15 | 94.12 | 1 275 |
| SLH-MTM-L2 | 94.52 | 95.65 | 2 765 |
| 本文算法 | 94.52 | 95.65 | 843 |

### 3.2 用户社会行为数据的相似度分组测试与分析

首先利用前节的用户工作类型分组数据及最优化的 $\cos\theta$ 值计算出各工作类型分组中经济条件指数和压力指数,其次计算出基于不同的用户社会行为数据的相似度 $\cos\omega$ 值,并根据 $\omega$ 角度将用户分为三大类,即处于 $0 \leqslant \omega < 30°$ 的用户工作压力较轻;处于 $30° \leqslant \omega < 60°$ 的用户工作压力适中;处于 $60° \leqslant \omega < 90°$ 的用户工作压力较重。算法的处理流程如图 6 所示。表 5 给出了测试中 6 种工作类型用户的工作压力分布情况。

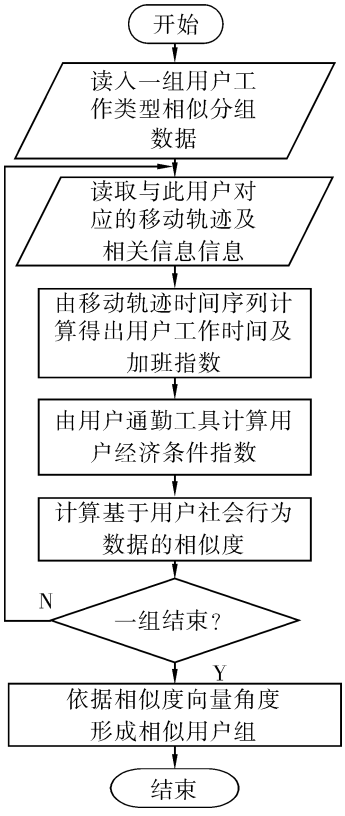



**图 6 用户工作压力分组流程图**

**Fig.6 User working pressure grouping flow chart**

**表 5 工作压力分布**

**Table 5 Working pressure distribution**

| 工作类型 | 用户人数 | 工作压力较轻人数 | 工作压力适中人数 | 工作压力较大人数 |
|---|---|---|---|---|
| 应急服务业 | 13 | 8 | 3 | 2 |
| 普通服务业 | 10 | 6 | 3 | 1 |
| 政府机关 | 8 | 7 | 1 | 0 |
| 企业 | 27 | 13 | 6 | 8 |
| 工厂 | 6 | 2 | 4 | 0 |
| 教育 | 5 | 5 | 0 | 0 |

在对上述用户进行分类时可采用两种分类粒度:一是无视工作类型分类,适合在形成相似用户组后组内推送与用户工作类型无关的信息;二是参考工作类型分类,适合在形成相似用户组后组内推送与用户工作类型有关的信息。

以无视工作类型的分类粒度为例,在 3 类用户中随机选取 5 个用户,由专家结合实际情况给出的工作心理建议,对于不同用户最终的推送信息是:

工作压力较轻用户的健康建议分别为:

1)需要不断树立新的目标,让自己再拼搏一次;

2)轻松工作之余可开辟“第二战场”,让时间变得有意义;

3)您还可以做得更好,赚更多的钞票。

工作压力适中的 3 位用户的健康建议分别为:

1)请您保持健康的心态,保持现在的工作步调;

2)车间工作会枯燥,建议休息时看看书调节心情;

3)你可以选择默默无闻,也可以选择惊天动地。

工作压力较重的用户的健康建议分别为:

1)请正确地认识自己,不要让办不到的事压垮你;

2)学会休息,学会说“不”,学会满足,避免成为一个完美主义者;

3)工作狂,请多陪陪家人,顺便放松自己。

通过以上推荐建议内容可知本分组算法的效果良好,适当地将工作压力程度相似的用户分组,相似分组用户间对于工作压力辅导健康建议具有可相互推送性,从而在功能性和可实施性的角度完成了用户分组的测试。

## 4 结束语

基于位置信息的服务已融入到人们的生产生活中,各种健康养生推荐服务已成为研究热点。针对有限医疗资源条件下如何实现用户的个性化健康建议推荐服务问题,本文提出了一种利用已记录的用户移动轨迹数据信息实现移动医疗环境下的个性化推荐算法,将已有的心理健康建议推荐给具有相似工作环境与工作压力的用户,提高了推荐效率,大幅缩减了由人工填写心理健康建议的代价。算法测试

与分析结果表明技术方案的有效性和可实施性。但原型系统实际应用过程中诸多亟待解决的问题还需进一步深入研究,如用户工作类型如何更精确化地定位;如何引入用户其他行为信息因素优化推送对象等。

## 参考文献:


[1]谢幸, 郑宇. 基于地理信息的用户行为理解[J]. 计算机学会通讯, 2008, 4(10): 13-21.
XIE Xing, ZHENG Yu. User behavior understanding based on geographic information[J]. Computer society news-letter, 2008, 4(10): 13-21.

[2]ZHANG Chengyang, ZHENG Yu, XIE Xing. Map-matching for low-sampling-rate GPS trajectories[C]//Proceedings of ACM SIGSPATIAL Conference on Geographical Information Systems. Seattle, Washington, USA, 2009: 213-221.

[3]YE Yang, ZHENG Yu, CHEN Yukun, et al. Mining individual life pattern based on location history[C]//Proceedings of the International Conference on Mobile Data Management. Taipei, China, 2009: 36-39.

[4]LOU Yin, ZHANG Chengyang, ZHENG Yu, et al. Map- matching for low-sampling-rate GPS trajectories[C]//Proceedings of ACM SIGSPATIAL Conference on Geographical Information Systems. Seattle, Washington, USA, 2009: 69-102.

[5]YE Yang, ZHENG Yu, CHEN Yukun, et al. Mining individual life pattern based on location history[C]//Proceedings of the International Conference on Mobile Data Management. Taipei, China, 2009: 46-50.

[6]ZHENG Yu, LIU Like, WANG Longhao, et al. Learning transportation modes from raw GPS data for geographic application on the web[C]//Proceedings of the 17th International Conference on World Wide Web. Beijing, China, 2008: 45-49.

[7]ZHENG Yu, LI Quannan, CHEN Yukun, et al. Understanding mobility based on GPS data[C]//Proceedings of ACM Conference on Ubiquitous Computing. Seoul, Korea, 2008: 26-31.

[8]ZHENG Yu, CHEN Yukun, LI Quannan, et al. Understanding transportation modes based on GPS data for Web applications[J]. ACM transactions on the Web, 2010, 4(1): 1-36.

[9]ZHENG Yu, ZHANG Lizhu, XIE Xing, et al. Mining interesting locations and travel sequences from GPS trajectories[C]//Proceedings of International Conference on World Wild Web. Madrid, Spain, 2009: 121-125.

[10]LI Quannan, ZHENG Yu, CHEN Yukun, et al. Mining user similarity based on location history[C]//Proceedings of ACM SIGSPATIAL Conference on Geographical Information Systems. Irvine, CA, USA, 2008: 127-131.

[11]ZHENG Yu, ZHANG Lizhu, XIE Xing. Recommending friends and locations based on individual location history[J]. ACM transactions on the Web, 2009, 3(2): 16-21.

[12]ZHENG Yu, ZHANG Lizhu, XIE Xing. Mining correlation between locations using human location history[C]//Proceedings of ACM SIGSPATIAL Conference on Geographical Information Systems. Seattle, Washington, USA, 2009: 145-151.

[13]王德起, 许菲菲. 基于问卷调查的北京市居民通勤状况分析[J]. 城市发展研究, 2010, 17(12): 98-105.
WANG Deqi, XU Feifei. A study on the commuting problems in Beijing-based on the investigation to the citizens of Beijing[J]. Urban development studies, 2010, 17(12): 98-105.

[14]贾晓朋, 孟斌, 张媛媛. 北京市不同社区居民通勤行为分析[J]. 地域研究与开发, 2015, 34(1): 55-59.
JIA Xiaopeng, MENG Bin, ZHANG Yuanyuan. Analysis of the residents commuting behavior in different communities in Beijing city[J]. Areal research and development, 2015, 34(1): 55-59.

[15]Microsoft Research Asia. GeoLife data set[DB/OL]. Beijing: Microsoft Research Asia, 2012. (2012-08-09)[2015]. http://research.microsoft.com/en-us/downloads/b16d359d-d164-469e-9fd4-daa38f2b2e13/default.aspx.

[16]LI Quannan, ZHENG Yu, XIE Xing, et al. Mining user similarity based on location history[C]//Proceeding of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. New York, NY, USA, 2008: 1-10.

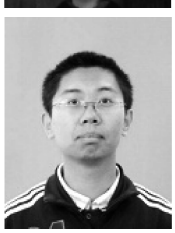
[17]ZHENG Yu, ZHANG Lizhu, MA Zhengxin, et al. Recommending friends and locations based on individual location history[J]. ACM transactions on the Web, 2008, 5(1): 5-44.

[18]XIAO Xiangye, ZHENG Yu, LUO Qiong, et al. Inferring social ties between users with human location history[J]. Journal of ambient intelligence and humanized computing, 2014, 5(1): 3-19.

[19]GIANNOTTI F, NANNI M, PEDRESCHI D, et al. Trajectory pattern mining[C]//Proceedings of the 13rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining. San Jose, CA, USA, 2007: 330-339.


**作者简介:**


陈万志,男,1977 年生,副教授,博士计算机学会会员,主要研究方向为人工智能、计算机过程控制、物联网应用、WebGIS 等。



林澍,男,1990 年生,硕士研究生,主要研究方向为人工智能、物联网应用。