大数据框架下基于相似度计算模型的用户位置检测

梁广荣[1]
LIANG Guangrong

摘 要 针对现有用户位置检测算法存在的定位检测误判率较高的不足，在大数据框架下设计了一种基于相似度计算模型的检测算法。在Hadoop大数据平台上布置各核心组件，平台能够提供可追溯的SQL查询功能，在大数据框架下提取网络用户移动轨迹的五要素特征，同步构建多尺度的相似度模型，分别计算用户移动点群分布范围的相似度和距离相似度。为提升模型应对大规模用户位置数据集的能力，通过构建深度卷积网络模型用于训练复杂的位置数据集，提升对用户位置检测的精度。仿真结果显示：提升检测算法的用户位置检测误判率仅为2%，优于两种传统的用户检测算法而且随着测试集规模的增加，检测误判率指标未出现提升趋势，表明算法的稳定性良好。

关键词 大数据框架；相似度计算模型；距离相似度；误判率

doi：10.3969/j.issn.1672-9528.2025.08.012

0 引言

伴随着计算机科学技术、互联网技术、信息通信技术的快速发展，社交网络已经成为人与人之间通信交流的最重要方式之一[1-2]。与传统的社交模式相比，社交网络具有巨大的优势，例如，数据传递快捷、信息共享方便、能够实现消息的群发、通信的过程可以追溯等[3-4]。当前，基于5G移动网络和各类移动平台的社交方式已经成为人们工作[5]、学习和生活的最主要方式、近年来随着大数据技术的进一步发展，移动通信逐渐转换为视频通信[6]，进一步拉近了人与人之间的距离。在大数据框架下的移动互联网发展具有一个显著的趋势，社交活动与用户位置及用户的移动轨迹之间相融合，很多社交平台也提供了位置检测服务功能[7]。在开放的大数据框架和互联网环境下，每个网络用户都可以被视为一个可移动的信息收发节点，用户通过授权提供自己的位置信息，能够形成一个具有较为稳定拓扑结构的信息网络。

随着网络的发展一些不法分子开始通过网络攻击服务器或窃取用户的数据，通过用户位置检测也能够合理保护用户的数据安全、隐私安全，并有效打击网络不法分子[8]。现有的社交网络用户位置信息检测多基于节点之间的数据流定位用户的位置，文献[9]提出一种基于话单数据的用户位置检测方法，通过分析用户ID信息和浏览轨迹等活跃的数据流，以事后追溯的方式获取用户的位置。该种方案具有一定的滞后性且用户位置的定位准确率较低；文献[10]提出一种轨迹k匿名的定位方法，根据用户的k条活动轨迹来构建用户的大致活动区域，通过不断地细化和迭代推测出社交用户的大概位置。该种方法能够从一定程度上保护用户的隐私安全，但依然存在定位精度不高的缺点。针对现有用户位置检测方法存在的不足，本文在大数据框架下设计了一种相似度模型以海量的用户位置检索数据为基础，计算位置数据之间的相似度进而较为准确地判断出用户的当前位置。

1 大数据框架的构建与节点组件部署

针对用户的网络位置检测，依赖于对海量与用户位置相关大数据的分析才能实现，因此本文设计的用户位置定位与检测模型基于Hadoop平台来实现[11]。从狭义的视角来分析Hadoop平台是一种集成分布式计算、存储和资源调度的大数据处理平台。在现有的网络环境下将目标用户作为大数据平台的一个节点，通过获取该目标用户的节点信息，来判断目标用户的大概位置范围。在每个用户节点上部署Hadoop大数据平台的组件，具体包括Flume组件、Sqoop组件、Spark组件、Hive组件等。

1.1 大数据框架的Flume架构

该组件用于采集和传输用户的行为活动日志信息[12]，并将相关的数据信息通过网络传输到后台上位机，与其他节点共享，Flume组件架构如图1所示。

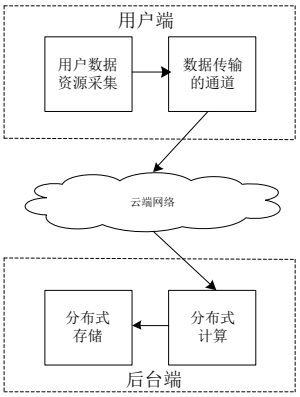

图1 大数据框架的Flume组件

1. 广东培正学院 广东广州 510830

1.2 大数据框架的 Sqoop 组件

Sqoop 组件用于不同类型数据库之间的数据高效传输，并与 Hive 组件之间进行数据的双向传递及数据共享。Sqoop 本质上是一种升级版的 MapReduce 工具，能够智能化对不同类型的数据进行分类，并解决数据格式不一致带来的问题。

1.3 Spark 组件

该组件是 Hadoop 大数据平台核心数据处理模块，具体功能包括为各种数据分析和处理算法提供平台，实时处理结构化的数据和多源异构的数据，兼容各种机器学习和图学习硬件和软件算法，Spark 组件是进行相似度计算和用户位置检测的系统平台。

1.4 Hive 组件

该组件的功能是为 MapReduce 计算和 Spark 组件计算提供 SQL 代码支持，并为每位授权用户提供 SQL 查询的功能，Hive 组件的学习成本和使用成本均较低，且能够兼容多种不同的网络接口，以更好地服务用户。构建大数据框架的意义在于为每个用户节点都提供了较强的数据采集和分析处理的能力，更便于通过数据流提高用户位置的可追溯性，实现对用户位置的精确定位和检测。

2 基于相似度模型的用户位置定位与检测

2.1 用户浏览轨迹数据特征的提取

基于大数据框架能够提取到判断用户 i 当前 t 时刻移动轨迹 l_t 的五元组要素：

$$l_t = f\left(I, T_1, T_2, T_3, C\right) \tag{1}$$

式中：f 表示对用户移动轨迹定位的函数；I 表示用户 ID 信息；T_1、T_2、T_3 分别表示用户的数据请求时间、到达时间和离开时间；C 表示距离此用户最近的基站。

通过对 5 个元素的逐一遍历和信息更新，同时利用用户节点的大数据框架过滤空白的字段和不活跃的数据，得到 t 时刻较为合理的移动轨迹数据集。计算用户 i 从当前 t 时刻到 t+1 时刻的移动距离 d：

$$d = \frac{R}{180} \cdot \arccos\left(\theta\right) \tag{2}$$

式中：R 表示地球半径；θ 表示 t 时刻位置与 t+1 时刻位置形成夹角的余弦值：

$$\theta = \sin\alpha \sin\beta + \cos\alpha \cos\beta \cdot \cos\left(\alpha - \beta\right) \tag{3}$$

式中：α 和 β 分别表示 t 时刻位置与 t+1 时刻位置形成经度值和维度值夹角。

通过对用户 i 一天不同时间段位置轨迹的识别，可以定义出用户移动的关键位置或偏好位置。

2.2 多尺度相似度计算模型的构建

在大数据框架下基于空间特征数据和属性特征数据两个视角来构建多尺度相似度计算模型。先采用德洛内三角网定义空间内用户 i 的位置变化关系，即用户 i 从当前 t 时刻位置 A，移动到下两个时刻的位置 B、位置 C，3 个点位置连线形成一个三角形，三条边分别为 AB、BC 和 AC，假定用户 i 移动的点数量为 N，则在用户 i 整个移动过程中任意三点都可以形成一个三角形，在用户的移动周期内所有两点之间的边长均值$\overline{a}$表示为：

$$\overline{a} = \frac{\sum_{j=1}^{N} a_j}{N} \tag{4}$$

式中：a_j 表示任意相邻两点间的第 j 条边，则 a_j 的约束条件为：

$$\tau_{a_j} = \overline{a} + \frac{2\sum_{i=1}^{N} V\left(a_j\right)}{N} \tag{5}$$

式中：V 表示 a_j 的方差函数，通过方差约束剔除方差波动值过大或过小的边。

在获取到用户 i 区域内全部活动轨迹后，计算边长的变异值并删除整体层次上的长边，根据方差函数优选三角形以获得相对更优的聚类结果。定义用户 i 移动点群分布范围的相似度$\sigma_{s_p s_q}$（s_p 和 s_q 分别由不同且不重叠的 3 个顶点构成的三角形）：

$$\sigma_{s_p s_q} = 1 - \frac{\left|s_p - s_q\right|}{\max\left(s_p s_q\right)} \tag{6}$$

在用于 i 的移动区域范围内，其顶点（停留点）的分布密度 ρ_N 是影响到$\sigma_{s_p s_q}$值的重要依据，顶点密度相似度计算为：

$$\rho_N = \frac{\rho_s}{\sum_{j=1}^{N}\left(\frac{1}{s_j}\right)} \tag{7}$$

各顶点之间的距离也是度量相似度的重要指标之一，顶点之间距离能够反映出用户移动范围的离散程度。用全部顶点的最大距离作为区内目标的最大约束，多尺度下用户 i 从 K 点到 H 点距离关系相似度 L_{KH} 表示为：

$$L_{KH} = 1 - \frac{\left|L_K - L_H\right|}{\max\left(L_K L_H\right)} \tag{8}$$

通过综合考虑用户移动点群分布范围的相似度$\sigma_{s_p s_q}$和距离关系相似度 L_{KH}，综合评估用户当前的精确位置。

2.3 用户位置相似性匹配与检测的实现

本文基于德洛内三角网定义了与用户相关的节点之间的关系，在目标网络用户移动轨迹相似性匹配方面，根据用户移动节点位置距离和点群分布范围的相似度预估准确的位置。根据临近原则从当前的三角形移动范围拓展到邻近的三

角形移动范围，直到全部的节点匹配完毕。当相似度计算模型面对海量用户多终端的位置检测需求时，计算性能会有所降低，为进一步提升相似度计算模型的数据处理能力，引入了深度卷积网络模型，用于分析和处理多用户的位置变化信息，并优选出最佳的点群分布范围的相似度$\sigma_{s_p s_q}$和距离关系相似度 L_{KH}。深度卷积网络模型在结构上包括了输入层、卷积层、池化层、全连接层和输出层，如图 2 所示。

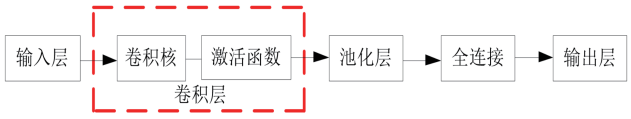

图 2 用于相似度计算的深度卷积网络模型

深度卷积网络模型是一种具有局部连接、权重共享等特性的深度神经网络，卷积过程就是一个资源匹配和加速计算的过程，在高效学习、训练获得精确的模型外，还可以实现更高效的运算。对于大数据框架下多用户位置的跟踪、定位和检测，应用深度卷积网络模型强大的数据特征提取能力，能够提高对用户位置的定位检测精度。在卷积层卷积核的选择上，根据输入用户数据集的规模采用 3×3 的卷积核或 1×1 的卷积核。本文在模型激活函数方面采用了光滑连续的 sigmoid 函数，以更好地匹配用户位置信息检测的数据集复杂度：

$$s\left(x_t\right)=\frac{1}{1+e^{x_t}} \tag{9}$$

式中：x_t 表示 t 输入深度卷积网络的数据集由用户的当前的位置信息、顶点分布相似度、距离分布相似度加权构成；$s(x_t)$ 表示 sigmoid 激活函数。

池化层的主要作用是对训练用户位置数据的降采样处理，本文所采用的池化操作是平均池化方法，即将训练完毕特征数据集划分为不同区域分别对比识别，不同区域的位置特征数据集不重叠，在均值化处理后输出结果到全连接层。深度卷积网络的全连接层核心功能是将与用户当前位置特征信息进行系统化整合，也是一种位置信息的融合处理，通过与前面一层的所有人工神经元节点的逐个连接，再对其加权求和处理，最后得到的关键特征信息经过全连接层的整合，转换成可直观识别的位置特征向量。

3 实验结果与分析

3.1 实验环境搭建与参数设置

首先在实验室虚拟环境下构建一个包括 20 个节点的大数据网络实验环境，选定 5 个虚拟网络用户，确保用户始终处于移动的状态，该 5 个用户的位置具有随机性，可能在整个拓扑结构中任一个节点，也可能多个用户在同一个节点，通过对各用户之间的通信数据的采集和分析，判断各网络用户在不同时刻的位置。节点的拓扑关系设计，如图 3 所示。

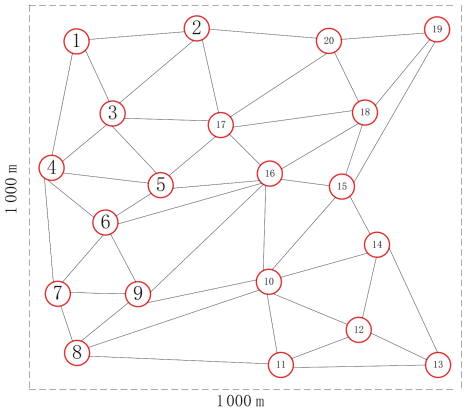

图 3 实验的阶段拓扑结构设计

在 1 000 m×1 000 m 的仿真区域内，被测试的 5 个虚拟用户的通信数据均被记录和采集，形成训练集和测试集，其中训练集的数据用户训练各模型的参数，测试集的数据用于检测用户的位置，训练集和测试的数据比例为 4:1，具体的训练集和测试集的分布情况，如表 1 所示。

表 1 用户在各节点的通信数据集分布情况

序号	数据集	数据量
1	训练集	34 120
2	测试集	8 530
合计	—	42 650

仿真实验的硬件环境和软件的设置如下：CPU Intel Corei7 14700 KF，最高主频 5.6 GHz，RAM 16 GB，ROM 2 TB，软件操作系统选择开源性和兼容性更好的 LINUX 系统；其中深度卷积网络的隐含层数为 4 层，学习率为 0.001，最大迭代次数为 200 次。其他参数均按照最优结果匹配，使多种算法下仿真系统的功能达到相对的最优状态。

3.2 实验数据对比

在图 3 的仿真区域内分别选定 10 个时间点，5 个用户采用随机的方式停留在某一个节点，系统会记录用户的停留时间并做好停留记录，但用户的移动方向和下一个停留的位置随机，用户在移动的过程中保持通信，通信数据被系统采集并记录，分别基于本文相似度位置检测算法，文献 [9] 算法及文献 [10] 算法通过用户的通信信息预估和判断用户在当前时刻的位置，用户 1~5 在 10 个时刻的位置信息统计，如表 2 所示。

针对不同用户位置检测算法位置检测的误判率 ξ_i 计算过程为：

$$\xi_i=\frac{m_1}{m}\times 100\% \tag{10}$$

式中：m 表示用户 i 经过设定的时间点的次数，本文 m 取值为 10；m_1 表示错误检测用户位置的次数。

表 2 用户在不同时间点的位置统计

节点序号	用户 1	用户 2	用户 3	用户 4	用户 5
时间点 1	1	2	20	19	7
时间点 2	3	17	19	18	5
时间点 3	5	5	18	15	16
时间点 4	6	16	15	16	15
时间点 5	9	10	12	17	14
时间点 6	10	14	10	3	12
时间点 7	11	12	9	4	10
时间点 8	8	11	8	6	9
时间点 9	7	8	11	7	6
时间点 10	4	9	13	9	5

各个算法关于 5 个用户误判率 ξ_i 的统计结果如表 3 所示。

表 3 用户在不同时间点的位置误判率统计

单位：%

节点序号	用户 1	用户 2	用户 3	用户 4	用户 5	均值
本文算法	0	0	0	0	10	2
文献 [9]	10	20	10	20	0	12
文献 [10]	20	0	10	10	20	12

误判率的统计结果显示：本文算法在针对 5 个虚拟用户的误判率均值为 2%，证明仅出现过 1 次位置误判；而两种传统的用户检测算法的误判率均值均为 12%，在针对 10 个时间点的监测中分别出现了 6 次误检。扩大样本的检测范围并针对表 1 中训练集和测试集进行位置误判的定位和检测，5 个虚拟用户的移动轨迹更加复杂且统计的样本数据增加，针对训练集和测试集的位置误判率统计结果如图 4~5 所示。

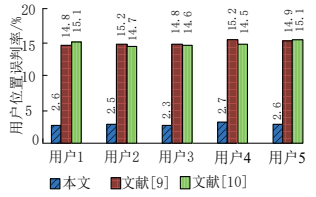

图 4 针对训练集的用户位置故障检测

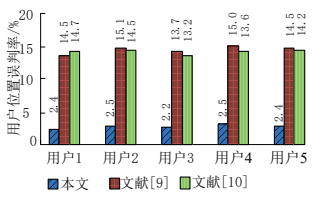

图 5 针对测试集的用户位置故障检测

从针对更复杂的训练集和测试集数据分析结果可知，本文基于相似度的用户位置检测优势更加明显，误判率指标值未超过 3%，而文献 [9] 和文献 [10] 两种传统检测算法的误判率指标未低于 13%，由此可以证明本文大数据框架下的相似度计算模型在网络用户位置检测方面具有较为明显的优势。

4 结语

在开放的互联网环境下通过对网络用户的位置定位和检测，能够更好地实现与用户的在线通信，并为用户提供点对点服务。通过准确掌握用户的当前位置，还能够为用户提供远端的数据云端存储和数据保护，确保网络用户的数据安全。本文在大数据框架下，设计了一种基于相似度计算模型的检测算法，根据用户的移动轨迹确定当前的空间特征和属性特征，同步获取点群移动相似度特征和距离相似度特征。针对大规模用户位置数据集和通信数据集的海量数据环境，本文在大数据框架下还引入了神经卷积网络模型，用于训练当前的用户位置数据，有效提取网络用户移动的轨迹特征。实际的仿真结果显示，大数据框架下基于相似度的位置检测误判率更低，能够更加准确地检测用户位置，为网络用户提供安全和快捷的定位服务。

参考文献：

[1] 张志然 . 位置社交网络中用户签到特征分析及推荐方法研究 [J]. 测绘学报 , 2023, 52(6):1039.

[2] 王晓丹 , 王子乔 , 金山海 . 社交网络中用户签到行为位置泄露风险预警 [J]. 计算机仿真 , 2023, 40(9):401-405.

[3] 高嘉媛 , 熊伟 , 陈荣 , 等 . 融合文本主题和社交关系的社交网络用户住所位置推测方法 [J]. 地球信息科学学报 , 2024, 26(2):488-498.

[4] 宗传玉 , 李箬竹 , 夏秀峰 . 基于位置社交网络的用户社区和属性位置簇搜索 [J]. 计算机应用研究 , 2023, 40(9): 2657-2662.

[5] 周琳 , 肖玉芝 , 刘鹏 , 等 . 基于节点多关系的社团挖掘算法及其应用 [J]. 计算机应用 , 2023, 43(5):1489-1496.

[6] 牛淑芬 , 戈鹏 , 宋蜜 , 等 . 移动社交网络中基于属性加密的隐私保护方案 [J]. 电子与信息学报 , 2023, 45(3): 847-855.

[7] 谭振江 , 马瑀浓 , 姜楠 , 等 . 基于移动社交网络的位置隐私保护研究 [J]. 吉林师范大学学报 : 自然科学版 , 2023, 44(1): 123-131.

[8] 谭郁松 , 王伟 , 蹇松雷 , 等 . 基于异常保持的弱监督学习网络入侵检测模型 [J]. 计算机工程与科学 , 2024, 46(5): 801-809.

[9] 钟其柱 . 基于机器学习分析 VoLTE 视频通话质量的研究及应用 [J]. 电信科学 , 2020, 36(3):156-165.

[10] 黄宴委 , 林涛 , 黄文超 , 等 . 一种快速有限时间收敛的轨迹跟踪引导律 [J]. 控制理论与应用 , 2023, 40(6):965-976.

[11] 刘洋 , 王剑 , 唐明 , 等 . 基于 Hadoop 分布式计算的混合神经网络负荷分类模型 [J]. 科学技术与工程 , 2023, 23(4): 1549-1556.

[12] 程思强 , 李晓戈 , 李显亮 . 基于日志多特征融合的无监督异常检测算法 [J]. 小型微型计算机系统 , 2023, 44(12): 2727-2733.

【作者简介】

梁广荣（1972—），男，广东湛江人，本科，副教授，研究方向：软件工程。

（收稿日期：2025-03-27 修回日期：2025-08-04）