基于 DBSCAN 算法的 A 区犯罪预测*

赵传鑫

(中国人民公安大学 国家安全与反恐怖学院,北京 100038)

摘 要:为有效提升公安部门在实践工作中的犯罪预测能力,提出基于 DBSCAN 算法的 A 区犯罪预测方法。该方法采用了时空分析可视化技术和 DBSCAN 算法,对 A 区的犯罪数据进行分析。首先,对 A 区的犯罪数据进行描述性统计分析;然后,利用 DBSCAN 算法构建犯罪预测模型,并进行可视化处理;最后,通过对不同类型犯罪进行分析,预测犯罪热点,识别犯罪模式。实验结果表明,与传统的经验预测相比,该方法具有更好的预测效果,为公安机关打击犯罪和优化警力配置提供了决策依据。

关键词:DBSCAN;聚类分析;犯罪预测

中图分类号:TP309 **文献标识码**:A **DOI**:10.19358/j.issn.2096-5133.2020.07.013

引用格式:赵传鑫. 基于 DBSCAN 算法的 A 区犯罪预测[J].信息技术与网络安全,2020,39(7):72-77.

Crime prediction in area A based on DBSCAN clustering

Zhao Chuanxin

(School of National Security and Counter Terrorism, People's Public Security University of China, Beijing 100038, China)

Abstract: In order to effectively improve the crime prediction ability of the public security department in practical work, this paper proposed a crime prediction method in area A based on DBSCAN algorithm. This method used the spatio-temporal analysis visualization technology and DBSCAN algorithm to analyze the crime data in area A. Firstly, this method conducted descriptive statistical analysis of the crime data in Area A; then, it used the DBSCAN algorithm to build a crime prediction model and visualized it; finally, it predicted the hot spots of crime and identify crime patterns by analyzing different types of crime. The results show that compared with traditional empirical prediction, this method has a better prediction effect and provides a decision basis for public security organs to fight crime and optimize police force allocation.

Key words: DBSCAN; clustering analysis; crime prediction

0 引言

犯罪作为当今世界普遍存在的社会问题,对经济的发展和人们的生活有着重大的影响,因此预防犯罪是世界各国警察机构共同的目标。预测是预警预防的前提和首要环节,研究预防犯罪问题离不开犯罪预测[1]。根据 2018 年全国公安机关统计的犯罪数据,从犯罪类型结构上来看,传统犯罪呈下降趋势,但互联网相关的犯罪不断增加,新兴行业领域犯罪活动较为活跃,非法集资、“食药环”犯罪、未成年人暴力犯罪等受到社会高度关注,黑恶势力犯罪、毒品犯罪出现了新特点[2]。总体来看,盗抢骗、黄赌毒和经济犯罪成为当前的主要犯罪问题,重点打击此类犯罪,才能维护社会稳定和保障人民的安全感。通过准确实时的犯罪预测,可以帮助公安机关优化警力部署、提前制定预案,从而降低犯罪率。犯罪预测的过程包括数据的收集、模型的建立和对犯罪模式的识别。这不仅有助于公安机关判断犯罪热点地区,还可以发现犯罪高风险人员和易害群体。

犯罪是可以预测的。《美国统计学会会刊》上有文章提出,“最初的犯罪就像初震,接下来的犯罪就好比余震”[3]。把握犯罪的两个杠杆即犯罪成因和犯罪规律,就可以推出预测趋势,犯罪成因是犯罪运行的动力,犯罪规律是犯罪运行的规则,动力和

* 基金项目:中央高校基本科研业务费项目(2019JKF335)

规则确定了运行的方向。如果分析人员不是简单地对犯罪数据进行归纳总结，而是挖掘出海量犯罪数据背后的犯罪成因和犯罪规律，那么就可以准确地评估犯罪趋势并采取有力的措施。根据犯罪学相关理论，在一定区域中，犯罪地点并不是随机分布的，而是呈现出一些集中点，即“热点地区”[4]。掌握这些热点地区可能发生的犯罪类型，对于公安机关的决策部署具有重要的参考价值。此外，根据邻近重复效应，当一个盗窃犯在一个地点成功实施犯罪之后，他往往会在一周后再次潜入同一对象或者邻近对象家中作案，他的犯罪半径一般是在 2 公里以内[5]。案发地点附近的同类案件发生的概率较高，这同时给犯罪预测提供了理论支撑。

基于以上几点，本文提出一种基于 DBSCAN 算法的犯罪预测模型，该模型在 MATLAB 中构建，对 A 区的犯罪数据进行密度聚类，可以获得较为客观的结果。该实验遵循数据分析中的步骤，包括数据收集与预处理、数据可视化和构建犯罪预测模型。在数据采集与预处理阶段，数据来自于 A 区公安局提供的犯罪数据，并进行了脱敏处理。数据可视化阶段生成了三维的犯罪时空分布图。最后，在构建模型阶段，本文对犯罪类型、犯罪时间、经度、纬度进行密度聚类分析，对警方的警力部署和决策执行具有参考价值。

1 相关研究

传统的犯罪分析通常是基于经验的判断，存在主观盲目性、数据量少、方式单一等特点，容易发生错误的判断。而大数据时代的犯罪预测，以现代犯罪学理论为基础，结合人口统计、社会环境和犯罪案件等要素获得预测结果，研究人员据此在不同领域通过多种方式进行数据分析。通过数据建模获取数据之间的关系，让数据发声，具有科学性。文献[6]提出了基于随机森林的犯罪预测模型，通过对犯罪嫌疑人的基本属性进行分析，判断其犯罪可能性。文献[7]依据犯罪类型、位置、日期、纬度和经度 5 个属性的犯罪数据，利用 K 近邻算法对伦敦的犯罪情况进行分析，获得了不同犯罪类型的高发区域和高发月份。文献[8]通过测量误差来评估人工神经网络(ANN)、支持向量回归(SVR)、随机森林(RF)和梯度树增强(GTB)四种人工智能技术预测犯罪的能力，其研究结果认为 GTB 的犯罪预测能力更强。文献[9]根据 Facebook 和人口普查的人口变量，采用各种回归模型来预测犯罪率，结果发现与媒体和消费相关的因素对犯罪率的预测超越了人口因素。文献[10]利用基于萤火虫的模糊认知地图神经网络对犯罪活动相关特征进行了有效预测，预测过程采用了改进的联想神经网络方法。

在实践运用方面，犯罪预测对降低犯罪率有明显的效果。洛杉矶警局在大数据警务方面走在世界前列，它在 2011 年与当地高校合作开发了 Pred Pol 预测性软件，采集分析了 80 年来 1 300 万起犯罪案件，通过犯罪预测使相关区域的犯罪率降低了 36个百分点[11]。Blue Crush 犯罪预测分析系统在 2005 年投入使用之后，不停地分析犯罪模式和变化趋势，使得美国孟菲斯市成为更加安全的城市[12]。美国马里兰州和宾夕法尼亚州使用的一款犯罪预测软件为司法提供了巨大的帮助，它能够预测犯罪假释或者缓刑期间的犯罪可能性，也能为法庭假释条款和审判提供参考性意见[13]。英国警方开发了一款名为 NDAS 的系统，将有心理健康问题而可能实施暴力犯罪的人群锁定为高风险人群，为他们提供医疗帮助和心理咨询，从而降低犯罪的可能性[9]。

2 DBSCAN 算法简介

DBSCAN(Density-Based Spatial Clustering of Application with Noise)聚类算法已广泛应用于道路交通[14]、图像处理[15]、金融风险评估[16]等领域，是一种比较有代表性的基于密度的聚类算法。它将簇定义为密度相连的点的最大集合，并可在噪声的空间中发现任意形状的聚类，其目标就是找到密度相连对象的最大集合。

有关 DBSCAN 聚类的几个基本概念：

(1)邻域：以给定对象 p 为圆心，半径为 r 的圆形区域，称为 p 的邻域。

(2)核心对象：给定对象 p，其邻域内的样本点数 ≥ MinPts，则称 p 为核心对象，如图 1 所示，设定 MinPts=3，可以看到 p 的邻域内样本点数为 7 个，所以 p 为核心对象，可以理解为点 p 的密度比较大。

(3)边界点：非核心对象，如图 2 所示，p 的邻域内样本数小于 MinPts。

(4)噪音点：不与任何密度区域相连，如图 3 所示，p 点是孤立于其他点的。

(5)密度可达：如图 4 所示，o 在 p 的邻域内，从 p 到 o 是直接密度可达，而 q 对象的邻域内不包括 p，但是包括 o，这样 p-o-q，称 p 到 q 是密度可达的。

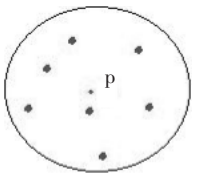

图 1 核心对象

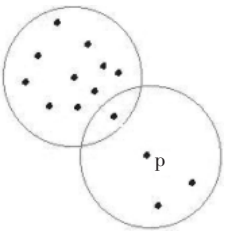

图 2 边界点

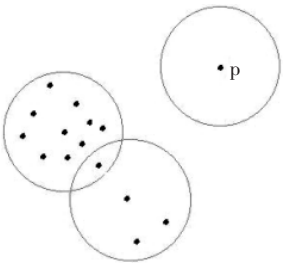

图 3 噪音点

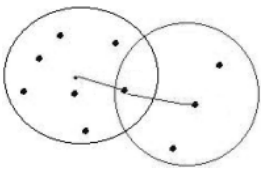

图 4 两点密度可达

(6)密度相连:如图 5 所示,q 和 p 是密度可达的,q 和 t 也是密度可达的,则 p 和 t 是密度相连的。

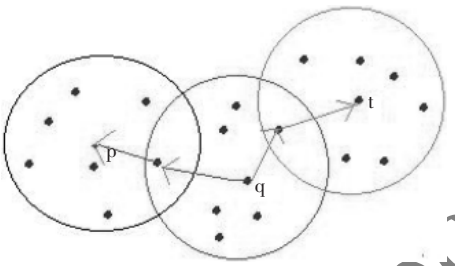

图 5 两点密度相连

DBSCAN 算法伪代码如下:

```
标记所有对象为 unvisited
while unvisited 元素个数>0:
    随机选择一个 unvisited 对象 p:
      标记 p 为 visited
      if p 的邻域至少有 MinPts 个对象:
          创建一个新簇 C,并把 p 添加到 C
          N={p 的邻域中的对象集合}
            For p′ in N:
                if p′==unvisited:
                    标记 p′为 visited
                    if p′的邻域至少有 MinPts 个对象:
                        把这些对象添加到 N 中
                    把 p′添加到 C
          属于簇 C 的样本点归位
      else:
          标记 p 为噪声
```

3 实验与分析

本文运用 DBSCAN 算法进行犯罪预测,是综合考虑了 DBSCAN 算法的特点和犯罪学相关理论的。既然犯罪在时间上和空间上存在一定的规律,那么就可以通过 DBSCAN 算法来发现这些数据之间的关联性。高密度的区域犯罪的可能性大,低密度的区域犯罪的可能性则较小,因此,在聚类得到的不同的簇中,可以获取不同的犯罪可能性。本实验分为三个步骤:(1)对收集到的 A 区公安机关的犯罪数据进行预处理,筛选与实验相关的数据并进行脱敏处理;(2)对数据进行描述性统计分析,并将结果可视化;(3)运用 DBSCAN 算法进行聚类分析,获取 A 区时空不同犯罪可能性。

3.1 数据预处理

本文所使用的数据集是从 A 区公安机关获得的可靠、真实数据。该数据集共包含 8 个属性,实验应用了犯罪类型、地点、日期、纬度和经度 5 个属性进行犯罪预测的研究。选取其中的 2018 年的数据进行一年的系统分析。在处理数据时,删除了其中缺失和存疑的数据,获得的有效数据为 1 804 条。在对时间的处理上,案件发生的时间取值为月、日,如 2018 年 3 月 2 日,取值为 3.02,这样有助于后期的 DBSCAN 实验。

3.2 数据可视化

通过初步的描述性统计分析,可以从直观上得到犯罪类型与犯罪时间的关系。在 A 区 2018 年发生的 1 804 件犯罪案件中,主要犯罪类型有盗窃类、打架斗殴类、毒品类、赌博类、妨害公务类、故意伤害类、交通肇事类、卖淫类、敲诈勒索类、危险驾驶类、性侵类、寻衅滋事类和诈骗类 13 个大类。从犯罪类型的角度分析,如表 1 所示,其中盗窃类、诈骗类和危险驾驶类是三大犯罪案件,分别占到了该区域全年总犯罪数量的 39.1%、33.8%和 9.6%,当地公安机关应该对这三类犯罪加以重视。从犯罪发生时间的角度分析,5 月、6 月和 7 月是犯罪高发时间段。盗窃类犯罪的高发时间段为 1 月、4 月、6 月和 10 月,诈骗类犯罪的高发时间段为 5 月、6 月和 7 月,危险驾驶罪的高发时间段为 5 月和 6 月。综合犯罪类型和犯罪时间的特点,公安机关应当在 5 月、6 月和 7 月加强对盗窃类、诈骗类和危险驾驶类犯罪的管控力度,从而有效控制 A 区的犯罪总数。

表 1 不同犯罪类型发生次数

案件类别	1 月	2 月	3 月	4 月	5 月	6 月	7 月	8 月	9 月	10 月	11 月	12 月	总计
盗窃类	74	41	43	70	59	71	64	58	68	76	49	32	705
打架斗殴类	1	3	2	2	1	0	0	0	2	1	0	2	14
毒品类	6	5	6	4	2	4	0	5	3	1	1	0	37
赌博类	5	2	4	3	4	3	3	1	4	3	2	2	36
妨害公务类	0	1	1	2	0	0	0	1	3	0	1	0	9
故意伤害类	1	0	1	6	2	1	1	5	0	3	3	5	28
交通肇事类	7	1	6	3	9	3	3	4	2	4	1	6	49
卖淫类	1	2	1	2	2	0	0	0	0	0	0	0	8
敲诈勒索类	0	0	2	0	0	1	2	0	0	0	0	0	5
危险驾驶类	9	9	11	16	20	24	13	9	14	17	15	17	174
性侵类	2	0	0	1	3	0	2	1	3	1	2	0	15
寻衅滋事类	1	4	2	3	4	3	5	3	2	2	1	0	30
诈骗类	48	18	48	44	70	70	89	56	30	46	47	44	610
其他类	12	4	7	11	7	7	10	9	4	9	2	2	84
总计	167	90	134	167	183	187	192	152	135	163	124	110	1 804

可视化技术可以将统计结果进行呈现。将不同的犯罪类型以时间和空间两个维度在三维图中显示出来。如图 6 所示，盗窃罪在该区的时空分布最广，显然经济犯罪成为严重影响该区人民生活安全感的犯罪类型。盗窃罪在一个特定的时空区域内较为集中，该区域纬度范围：30°80′N~30°85′N；经度范围：120°20′E~120°30′E；时间范围：6 月~10 月。因此，研究人员可以根据一个地区在某个时间段经常发生的犯罪类型，采取相应的预防措施。比如在 A 区 b 镇的 4 月~6 月盗窃案件频发，8 月~10 月危险驾驶案件频发，那么可以在 4 月~6 月加强社区巡逻，8 月~10 月在道路上分配更多的交警，同时注重宣传安全驾驶。

3.3 基于 DBSCAN 算法的犯罪预测模型

在前期的描述性统计中可发现犯罪类型和犯罪时间的相关规律，但是要想进行犯罪预测，还需要考虑更多的因素。本节构建基于 DBSCAN 算法的犯罪预测模型，将犯罪类型、犯罪地点、犯罪时间进

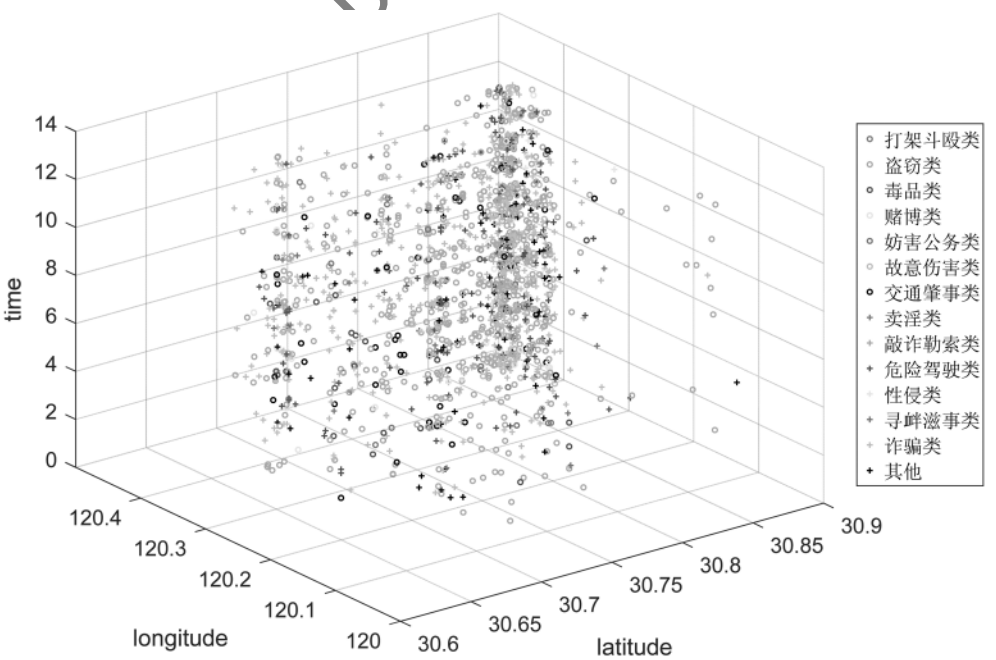

图 6 2018 年 A 区犯罪类型时空分布图

行聚类分析。

具体的算法聚类过程如下：

(1)对 eps 和 min_samples 的各种参数组合进行拟合计算，在噪声比较小的情况下，选取合适的参数 Eps=0.71，MinPts=10；

(2)输入预处理好的数据集 D={犯罪类型，时间，纬度，经度}；

(3)从数据集 D 中随机抽取一个未被处理的对象 p，且在它的 Eps 近邻满足密度阈值要求时把该对象 p 称为核心对象；

(4)遍历整个数据集，找到所有从对象 p 密度可达对象，形成一个新的簇；

(5)通过密度相连产生最终的簇结果；

(6)重复执行步骤(4)和(5)，直到数据集中所有对象都为“已处理”状态。

实验结果如图 7 所示，犯罪数据聚类后形成了 5 个簇(图中分别用 *，o，+，x，☆进行标出)，说明这 5 个簇中的点在时空上具有密切的联系。密集程度越高的地方，说明该时间段该区域的犯罪风险就越高，应当分配更多的警力[16]。可见，A 区在犯罪时空上存在着 5 个犯罪热点，并且这 5 个犯罪热点主要集中在一块区域之内：纬度为 30°85′N~30°90′N，经度为 120°40′E~120°50′E。公安机关分配警力应当以该区域为核心向周边扩散。

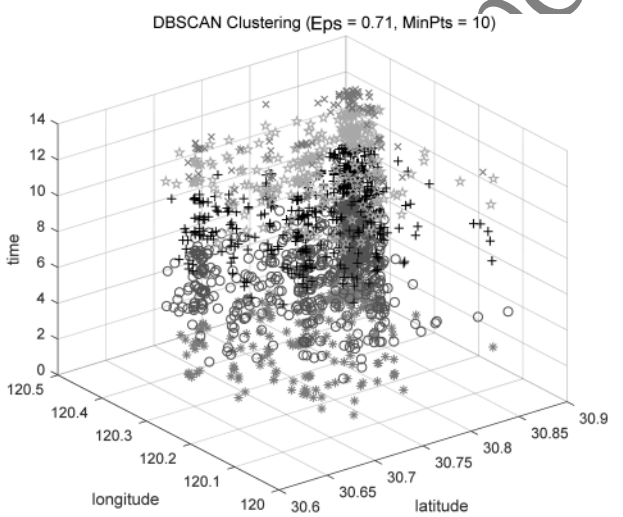

图 7 A 区 DBSCAN 聚类分析犯罪风险等级结果

如果一个地区发生了连续犯罪，就可以通过密度聚类的方式找出实际住所周围的排斥区和衰减区，从而得出犯罪嫌疑人可能的住所分布，有助于警方缩小案件调查范围。此外，该模型还可以实时监测犯罪动态，更新犯罪热点；对长期犯罪问题进行跟踪调查研究；通过密度的变化来评估治安防控的效果。

4 结束语

在当前全国公安机关积极推进“智慧警务”的趋势之下，犯罪预测必将成为其中的重要内容。本文采用了三维可视化技术和 DBSCAN 聚类算法来分析不同犯罪类型在 A 区的时空分布，研究结果显示盗窃罪分布最广，在特定的时空范围较为集中；该区在时空上存在 5 个犯罪热点，应当以 5 个犯罪热点为中心来部署警力。本文所构建的犯罪预测模型旨在为基层民警管控辖区犯罪趋势和做好犯罪预防提供技术支撑，有助于公安机关优化警力配置和提前制定预案，具有现实意义。由于收集到的犯罪数据的局限，本文在进行犯罪预测时没有纳入足够多的因素，下一步将考虑将犯罪造成的损失、地区经济发展水平以及人流量等要素纳入到犯罪预测的指标中，同时改进优化算法，提高犯罪预测的准确性和客观性。

参考文献

[1] 冯树梁.中国预防犯罪研究回眸与前瞻——为纪念改革开放 40 周年而作[J].犯罪与改造研究，2019(2)：12-22.

[2] 靳高风，守佳丽，林晞楠.中国犯罪形势分析与预测(2018—2019)[J].中国人民公安大学学报(社会科学版)，2019，35(3)：1-11.

[3] 董青岭.预测性警务：大数据犯罪预防[J].中国投资，2018(23)：18-19.

[4] 杨学锋.热点警务的犯罪学理论基础及实践评估[J].中国人民公安大学学报(社会科学版)，2018，34(3)：33-39.

[5] 王欣.美国的大数据警务应用[J].现代世界警察，2018(4)：60-63.

[6] 卢睿，李林瑛.基于随机森林的犯罪预测模型[J].中国刑警学院学报，2019(3)：108-112.

[7] 王娟，兰月新，吴春颖，等.时空分析和 K 近邻算法在犯罪分析中的应用研究[J].福建电脑，2019，35(7)：35-37.

[8] KHAIRUDDIN A R，ALWEE R，HARUN H.Comparative study on artificial intelligence techniques in crime forecasting[J].Applied Mechanics and Materials，2019，892：94-100.

[9] FATEHKIA M，O′BRIEN D，WEBER I.Correlated

impulses : using Facebook interests to improve predictions of crime rates in urban areas[J/OL]. Plos One , 2019 , 14(2)[2020-02-08]. https : //doi.org/10.1371/journal.pone.0211350.

[10] ALTAMEEM T , AMOON M. Crime activities prediction using hybridization of firefly optimization technique and fuzzy cognitive map neural networks[J]. Neural Computing and Applications , 2019 , 31(5) : 1263-1273.

[11] 李忠东. 预测犯罪[J]. 检察风云 , 2019(5) : 32-33.

[12] 阎耀军 , 张明. 犯罪预测时空定位信息管理系统的构建[J]. 中国人民公安大学学报(社会科学版) , 2013 , 29(4) : 73-80.

[13] 陈刚. 大数据时代犯罪新趋势及侦查新思路[J]. 理论探索 , 2018(5) : 109-114.

[14] 冯慧芳 , 杨振娟. 基于时空相似度聚类的热点载客路径挖掘[J]. 交通运输系统工程与信息 , 2019 , 19(5) : 94-100.

[15] 刘梦迪 , 潘晓 , 高珊珊 , 等. 结合 DBSCAN 聚类的室内场景分割[J]. 计算机辅助设计与图形学学报 , 2019 , 31(7) : 1183-1193.

[16] 杨瑞成 , 吕强 , 杨静. 基于圆形邻域孤立点挖掘算法的企业信用风险失真度研究[J]. 数学的实践与认识 , 2012(4) : 96-103.

(收稿日期 : 2020-04-13)

作者简介 :

赵传鑫(1995-) , 男 , 硕士研究生 , 主要研究方向 : 侦查讯问、公安情报。

(上接第 66 页)

on graphs[J]. Journal of Theoretical Biology , 2006 , 243(1) : 86-97.

[12] 胡裕靖 , 高阳 , 安波. 不完美信息扩展式博弈中在线虚拟遗憾最小化[J]. 计算机研究与发展 , 2014 , 51(10) : 2160-2170.

[13] LI D D , MA J. How the government's punishment and individual's sensitivity affect the rumor spreading in online social networks[J]. Physica A : Statistical Mechanics and its Applications , 2017 , 469 : 284-292.

[14] OHTSUKI H , NOWAK M. Evolutionary stability on graphs[J]. Journal of Theoretical Biology , 2008 , 251(4) : 698-707.

[15] PICKHARDT R , GOTTRON T , ANSGAR S. Efficient graph models for retrieving top-k news feeds from ego networks[C]. ASE/IEEE International Conference on Social Computing , 2012.

(收稿日期 : 2020-03-24)

作者简介 :

臧正功(1994-) , 男 , 硕士研究生 , 主要研究方向 : 社交网络、博弈论。

丁菁(1972-) , 男 , 博士 , 讲师 , 主要研究方向 : 车载自组织网、分布式计算。

版权声明