Computer Science and Application 计算机科学与应用, 2018, 8(8), 1153-1163
Published Online August 2018 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2018.88127

Hans汉斯

Network Self-Media Spatial Data Mining Based on Improved K-Means

Xiang Zheng, Xiaoji Lan*, Yu Zhong

School of Architectural and Surveying & Mapping Engineering, Jiangxi University of Science and Technology, Ganzhou Jiangxi
Email: *496439745@qq.com

Received: Jul. 17th, 2018; accepted: Jul. 30th, 2018; published: Aug. 6th, 2018

Abstract

Today, the per-capita holding rate of electronic media such as mobile phones and tablet computers is greater than one. The spread of online media has reached an unprecedented peak. Based on the Mahout data mining framework of Hadoop platform, this paper selects the K-means clustering analysis algorithm optimized by Canopy algorithm, clusters the data, and mines the network self-media tweets with much information to discover the micro: the hot words related to the current society and life contained in the Weibo data, and then through the ArcGIS, the kernel density analysis of the text clusters, and then the fishing grid rasterization analysis, so that the discrete cluster samples have the adjacency, enabling visualization visually see the main distribution of cluster topics, to study the habits of people's daily lives, to understand the preferences of a single individual, and to evaluate the social events, such as social and life related information.

Keywords

K-Means, Canopy, Micro-Blog, Cluster Analysis, Spatial Data Mining, ArcGIS

基于改进后K-Means下网络自媒体空间数据挖掘

郑　翔，兰小机*，钟　宇

江西理工大学建筑与测绘工程学院，江西　赣州
Email: *496439745@qq.com

收稿日期：2018年7月17日；录用日期：2018年7月30日；发布日期：2018年8月6日

*通讯作者。

文章引用: 郑翔, 兰小机, 钟宇. 基于改进后 K-Means 下网络自媒体空间数据挖掘[J]. 计算机科学与应用, 2018, 8(8): 1153-1163. DOI: 10.12677/csa.2018.88127

摘　要

在手机、平板电脑等电子媒介的人均持有率大于一的今天，网络自媒体的传播达到了前所未有的巅峰。本文通过基于Hadoop平台的mahout数据挖掘框架，选用经过Canopy算法优化后的K-means聚类分析算法，对数据进行聚类分析，对内涵众多信息的网络自媒体推文进行数据挖掘，以发现微博数据中蕴含的与当下社会和生活相关的热点词，后通过ArcGIS，对文本类簇进行核密度分析，再做渔网栅格化分析，使离散的类簇样本具备邻接性，能在可视化中直观地看到类簇主题的主要分布情况，以研究人们日常生活中的习惯、了解单一个人的喜好，以及对某个社会事件的评价等日常生活中隐含着关于社会和生活相关的信息。

关键词

K-means，Canopy，微博，聚类分析，空间数据挖掘，ArcGIS

1. 引言

在互联网技术快速发展的今天，网络自媒体在近些年得到了爆炸式的增长，它以病毒传播的形式快速地渗透到了各行各业中，给予了所有人展示自我和了解他人的途径[1]。新浪微博作为国内大型网络自媒体平台之一，它具有庞大的用户基础，以及由这基数庞大的用户群体所产生的与个人生活或社会现象等与各行各业相关的大量信息[2]；随着 Web2.0 时代的发展成熟，微博除了基于常规数据的数据挖掘以外，还有大量的包含有经纬度位置属性的数据[3]；这些空间位置数据可以很好地将我们的信息挖掘结果通过各大地图的前端 API 很直观地展示出来，让我们能很好地发现各种个人生活或社会现象等与各行各业、个人相关的话题或者感兴趣的商品等事物的空间分布规律等的空间信息，因此基于微博的数据挖掘研究是十分有价值的科研方向。

空间数据挖掘与知识发现(SDMKD, Spatial Data Mining and Knowledge Discovery)是数据挖掘和知识发现的分支学科，它通过对空间数据集进行一系列的处理，最终得到空间特征规则、空间聚类规则以及空间分布规律等能够直观展现空间实体的信息。最早开始关注、了解空间数据挖掘这一领域的人，是李德仁院士，他曾经在二十世纪末期召开的国际地理信息系统学术会议上，由他首次提出空间数据挖掘和知识发现理论，且研究并提出了空间数据挖掘和知识发现的理论框架[4] [5] [6] [7]。在现存的空间数据库里蕴含着巨量的信息，其中包括山高、河宽等可以使用地理信息系统的查询工具发现的浅层信息[8]；但除了浅层信息以外还有很多深层次的，如空间分类规则、空间偏差[9] [10] [11] [12]等信息则难以利用地理信息系统的查询方法来获取，只能通过运算或者挖掘等手段才能够发现这些信息。

由于云计算的迅速崛起，为我们在解决机器学习中的聚类问题时面临的复杂、大量的迭代计算提供了出色的解决方案；其中在众多的分布式计算框架中，开源框架 Hadoop 以其稳定的性能和廉价的成本被众多企业和科研机构所青睐，与传统并行框架相比，它具有高效、高可用、易部署等特点；apache 组织在该平台基础上开发了一个针对机器学习算法的计算框架——mahout；本文将以使用 mahout 加上 Hadoop 组成的平台为基础：Hadoop 生态中的 HDFS 为数据存储系统；Hadoop 生态中的 MapReduce 为分布式计算框架；然后选用 Canopy 算法优化后的 K-means 聚类分析算法，利用搭载在 Hadoop 集群上的 Mahout

数据挖掘框架来实现并行的聚类算法操作[13]，最后，使用可视化分析的手段，将带有主题的类簇展示在地图上，用这种更直观的方式来分析这些微博数据所蕴含的信息，以研究网络舆论中隐含着关于社会和生活中相关的信息，为社会的和谐、稳定的发展提供支持。

2. 数据与方法

2.1. 基于微博的数据获取

爬虫程序是一种快速、高效、轻量级的互联网数据抓取工具或手段。本文所使用的微博数据均是通过利用 Scrapy 框架编写的爬虫脚本所获取，它是由 Python 开发的一个快速、高层次的屏幕抓取和 Web 抓取框架，用于抓取 Web 站点并从页面中提取结构化的数据，它的用途广泛，可应用于数据挖掘、监测和自动化测试。本文通过 Scrapy 项目对微博数据进行抓取，数据内容为：用户 id、微博正文、经纬度三项，一共获取 700 万条数据，其中包括经纬度数据为空、经纬度不全的噪点数据。

2.2. 数据处理与分析

因 MongoDB 的数据结构与架构松散的 JSON 对象相似，且操作简单、易上手，在分布式存储上也表现的非常出色，故本文将通过 Scrapy 项目爬取的数据全都存储在 MongoDB 数据库中，在此数据的基础上进行预处理、聚类分析、相似度评价和可视化分析，具体方法如下：

1) 微博正文预处理：本文将微博用户看成一个空间实体，将其推发的微博正文和其经纬度位置数据为属性，将使用 IK 分词器加停用词表和 ext 表(额外自定义词库)，并使用加权的 TF-IDF 算法对微博正文分词结果进行加权，后使用 java 语言的 IO 流操作读入数据，并处理输出，最终得到囊括 11 万条微博经纬度完整的文博数据集。因为微博正文中的大量表情会对聚类结果产生影响，在爬取的过程中则将这部分表情符号替换，表示为“[表情]”，如：“微笑”的表情表示为[“微笑”]。除此之外，文本中还有一部分类似于“#、@”之类的特殊符号，同样会对结果产生影响，故将其剔除，剔除采用正则表达式执行。

2) 基于 Hadoop 与 Mahout 并行框架下的聚类分析：本文将微博用户看成一个空间实体，将微博正文和其经纬度位置数据当做其属性，在 hadoop 平台上对其微博正文属性进行中文聚类，并通过增删停用词，添加新词，去除高权重词等手段来优化聚类质量，最终发掘出微博数据集中的热点人物、商品、话题等信息以及各个热点信息所携带的关键字(例如：巴黎这一热点词伴随的关键字为：品味、艺术、旅行等)。

3) 基于文本相似的用户相似度评价：用户间的相似性通常用于好友推荐功能，传统的用户相似性度量通常是对用户兴趣、关注列表等微博用户的个人信息和关系网来计算得到[14] [15]，而在本文中则是考虑用户之间微博正文的相似程度和空间位置上是否邻近来判定用户是否相似；两个用户的微博正文相似表明其兴趣方向相似，而地理位置相近的用户可能具有相似的地域文化背景；在优秀的聚类结果中，同一个类簇中的各个样本涉及的主题都是类似的，对文本进行向量化后，两个向量之间的距离如果越接近，它们之间文本的相似程度就会越高，主题相似性自然也会越高；将相近的文本向量提取出来，使用微博中携带的空间位置信息评价其地域背景的相关性，则能够得到地域相近、兴趣话题相同的相似用户。

4) 使用空间数据挖掘中的数据可视化方法进行可视化分析：对 3、4 步中产生的聚类结果使用具备可视化功能的软件或者开放 API 进行可视化，以直观的发现各个类所对应的关键词或者关键话题的空间分布规律和关键词所包括的大致内容。本文实验的大致流程如图 1 所示。

3. 结果与讨论

3.1. Canopy 优化 K-means 聚类

Mahout 的 K-means 聚类算法中随机生成的初始中心点集会使聚类结果产生随机误差，改变这一现状

的最好方式就是使用合适的初始中心点来替换随机产生的点集，既使用 Canopy 近似聚类来生成初始的点中心；本文使用 mahout 中封装的 Canopy 聚类脚本命令来运行 Canopy 聚类，其命令如下：

该命令其他参数可输入 mahout Canopy-help 查看。

```
./mahout canopy –i[输入目录]  -o[输出目录]  -dm[距离测度方式]  -t1[外圈阈值 t1]  -t2[内圈阈值 t2]
```

在经过多次试验，最佳参数选择为：距离测度选择欧式距离，针对本文的样本集，t1 和 t2 的值在经过测试可以选择在 0 到 100 之间，过大的 t1 和 t2 会使样本集只有一个 Canopy，过小时 Canopy 则会很多，在多次测试后得到 Canopy 的最佳取值分别为 83 和 42，共生成 32 个 Canopy；Canopy 聚类的结果是 sequencefile 格式的文件，可以使用 seqdumper 命令将其转换成文本进行查看，如图 2 所示。

图中 key 为 Canopy 的 ID，value 为 Canopy 值，因为该值为向量形式，须使用 vectordump 查看，该处显示为向量所在地址的哈希值；针对文章中样本集选择的不同 Canopy 算法参数生成的 Canopy 数如表 1 所示。

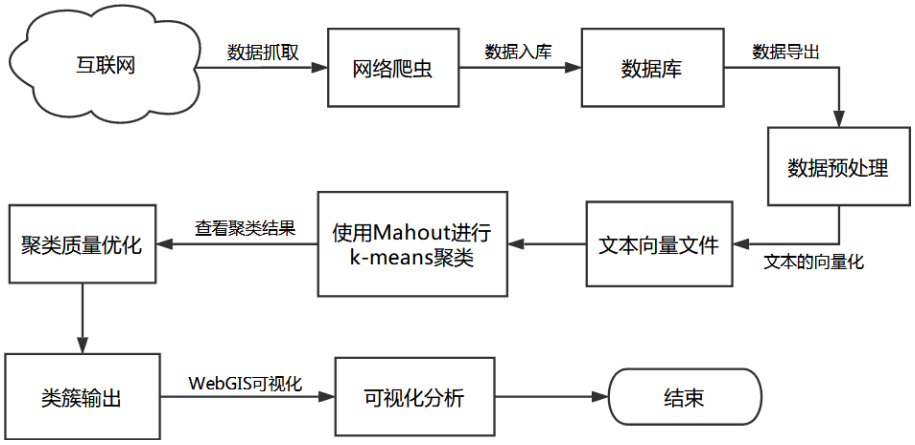


Figure 1. Experimental flow chart
图 1. 实验流程图

```
Key: C-10: Value: org.apache.mahout.clustering.iterator.ClusterWritable@2d34798d
Key: C-11: Value: org.apache.mahout.clustering.iterator.ClusterWritable@2d34798d
Key: C-12: Value: org.apache.mahout.clustering.iterator.ClusterWritable@2d34798d
Key: C-13: Value: org.apache.mahout.clustering.iterator.ClusterWritable@2d34798d
Key: C-14: Value: org.apache.mahout.clustering.iterator.ClusterWritable@2d34798d
Key: C-15: Value: org.apache.mahout.clustering.iterator.ClusterWritable@2d34798d
Key: C-16: Value: org.apache.mahout.clustering.iterator.ClusterWritable@2d34798d
Key: C-17: Value: org.apache.mahout.clustering.iterator.ClusterWritable@2d34798d
Key: C-18: Value: org.apache.mahout.clustering.iterator.ClusterWritable@2d34798d
Key: C-19: Value: org.apache.mahout.clustering.iterator.ClusterWritable@2d34798d
Key: C-20: Value: org.apache.mahout.clustering.iterator.ClusterWritable@2d34798d
Key: C-21: Value: org.apache.mahout.clustering.iterator.ClusterWritable@2d34798d
Key: C-22: Value: org.apache.mahout.clustering.iterator.ClusterWritable@2d34798d
Key: C-23: Value: org.apache.mahout.clustering.iterator.ClusterWritable@2d34798d
Key: C-24: Value: org.apache.mahout.clustering.iterator.ClusterWritable@2d34798d
Key: C-25: Value: org.apache.mahout.clustering.iterator.ClusterWritable@2d34798d
Key: C-26: Value: org.apache.mahout.clustering.iterator.ClusterWritable@2d34798d
Key: C-27: Value: org.apache.mahout.clustering.iterator.ClusterWritable@2d34798d
Key: C-28: Value: org.apache.mahout.clustering.iterator.ClusterWritable@2d34798d
Key: C-29: Value: org.apache.mahout.clustering.iterator.ClusterWritable@2d34798d
Key: C-30: Value: org.apache.mahout.clustering.iterator.ClusterWritable@2d34798d
Key: C-31: Value: org.apache.mahout.clustering.iterator.ClusterWritable@2d34798d
Count: 32
```

Figure 2. Seqdumper output Canopy
图 2. Seqdumper 输出 Canopy

将 Canopy 聚类结果的 clusters-0-final 文件作为 K-means 聚类的-c 参数输入，则可得到经过 Canopy 优化的 K-means(文章后续内容称为 C-means 算法)聚类结果，其详细的结果参数如表 2 所示。

以上参数是忽略了 Canopy 算法的运行时间的结果，在经过 Canopy 优化以后的 K-means 聚类质量得到一定提高，簇间距离更加明显，使得类簇间的区别变大，能够降低一个话题多个类簇情况出现的几率。使用 clusterdump 命令查看聚类结果，得到的部分结果如表 3 所示。

表中的类簇 id 为 mahout 随机分配，从聚类结果来看，大部分类别都带着较为明确的话题，而使用 K-means 亦或 Canopy + K-means 对类簇的话题并不会产生太大的影响，簇间距离的扩大能够有效的将类簇区分开，可是簇内密度的小幅变化难以对类簇话题产生影响。但是样本簇内距离的大小可以用作微博用户相似性的评价指标，为向用户推荐好友做决策。

使用聚类结果对各个类簇进行话题的总结。话题总结的最好方式是依据类簇中的高权重词建立决策树，可以将决策树的输出作为话题，对类簇中的样本所属用户进行广告推送，或者对类簇中的样本所属用户进行簇内距离计算，将一定阈值内的用户看做相似用户进行好友推荐。

K-means 和 C-means 算法在算法收敛速度、簇间最大距离、簇间最小距离的对比，如图 3。

3.2. 文本相似性计算

针对聚类结果中聚类样本所在的 clusteredPoints 目录，以其中单个样本为目标，计算目标样本所在类

Table 1. Canopy number corresponding to Canopy algorithm parameters
表 1. Canopy 算法的参数对应的 Canopy 个数

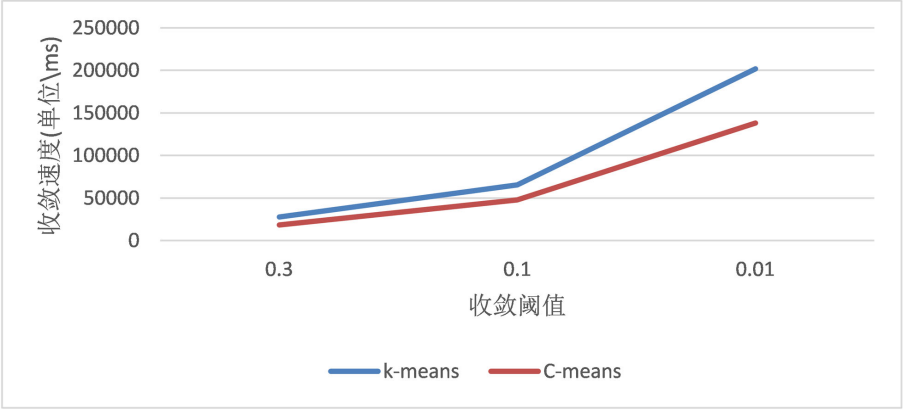
T1 (欧式距离)	T2 (欧式距离)	收敛速度	Canopy 个数
100	50	3537 ms	1
85	40	3686 ms	11
83	42	6164 ms	32
70	35	184,462 ms	113

Table 2. Canopy optimized K-means clustering parameters
表 2. Canopy 优化的 K-means 聚类参数

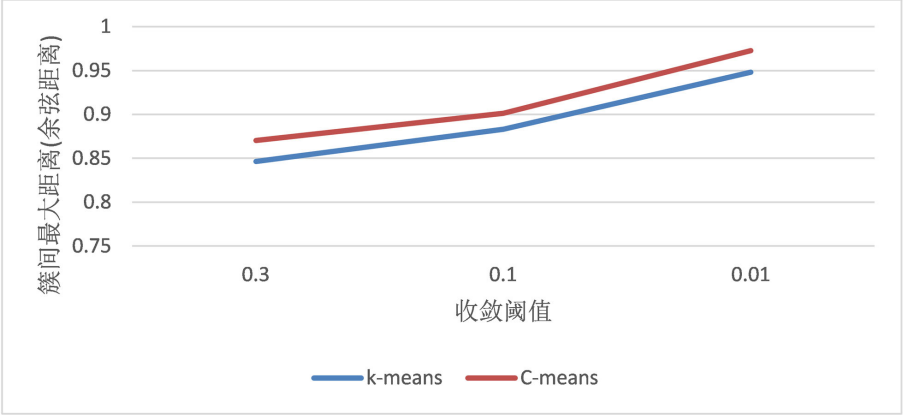
收敛系数	0.3	0.1	0.01
收敛速度(C-means)	18,430 ms	47,904 ms	138,076 ms
迭代次数(C-means)	2	3	9
簇间最大距离(C-means)	0.870353	0.901472	0.972938
簇间最小距离(C-means)	0.437512	0.487238	0.520349

Table 3. Canopy + K-means clustering results
表 3. Canopy + K-means 聚类结果

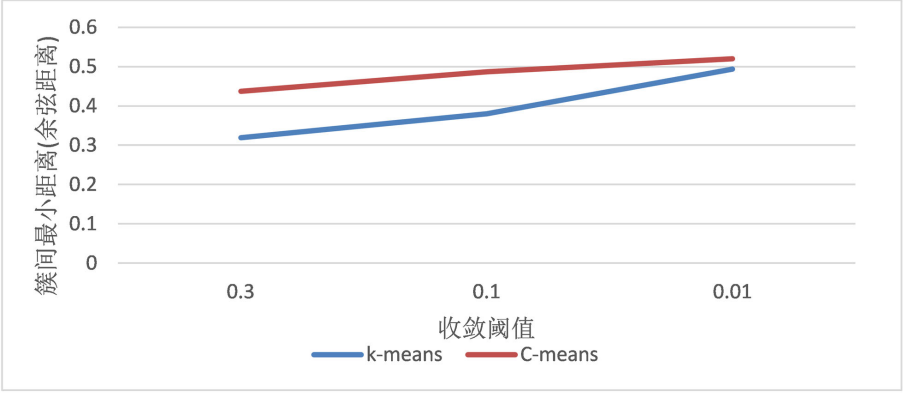
类簇 id	簇内 10 个最高权重词(top terms)
20136	普宁 机芯 表 机械 钢 表带 石英 机械 男士 镜面
19658	效果 反馈 皮肤 好了 面膜 补水 肌肤 毛孔 改善 产品
21179	商圈 北京 成都 吃 杭州 花园 漳州 苏州 江滨 香港
4738	运动 金牌 夺金 脚力 步 勋章 收获 跑步 挑战 行走
17938	琼海 十里 昆明 桃花 三世 三生 电视剧 卫视 大年 剧



(a)



(b)



(c)

Figure 3. (a) Convergence rate comparison between K-means and C-means; (b) Maximum distance between K-means and C-means algorithm clusters; (c) Comparison of minimum distance between K-means and C-means algorithm clusters

图 3. (a) K-means 与 C-means 算法收敛速度对比；(b) K-means 与 C-means 算法簇间最大距离对比；(c) K-means 与 C-means 算法簇间最小距离对比

簇中与其距离相近的样本点，将它们作为目标样本的潜在近似用户，然后通过经纬度过滤掉空间位置上距离较远的近似用户，剩下的就是与目标用户为中心，在地理位置和兴趣主题上都较为相似的用户，将它们看成一个相似簇，则可以在该簇中的样本所对应的用户之间进行好友推荐的操作。伪代码如下：

计算样本相似度伪代码：

```
Clustering Similar_User(ClusteredPoints , UserVector , T)
// UserVector 为目标样本，类簇中样本点与目标样本间距离小于 T 则判定为相似
    foreach (point in ClusteredPoints)
        distance = CalcDistance(UserVector , point)//距离计算
        if (distance <= T)
            PointList<Vector>.add(point)//加入相似簇
    end foreach
    return PointList
end Clustering Similar_User
```

其中距离计算使用的是余弦距离，T 值的选择应该较小，本文实验使用的 T 值是 0.3，结果查看使用 clusterdumper 类以目标样本为中心点，相似簇为类簇，dictionary.file-0 为映射输入；部分内容如图 4 所示：

其中 distance 为样本与目标间的余弦距离，后边为样本分词后的每个词在对应维度上的值。距离测度方式可按照好友推荐的精度进行调节；当聚类结果优秀时可对整个类簇中的样本进行好友推荐的操作。在计算出文本相似的用户以后使用针对这些微博用户使用欧式距离来计算它们之间的地域邻近性，就可以得到文本相似与地域邻近的相似用户。

3.3. 基于聚类结果的热点分析——以商圈主题类簇为例

ArcGIS 是一个功能强大的 GIS 桌面应用系统，本文中聚类生成的类簇里包含着几千甚至上万条与类簇主题相对应的微博，这些微博在空间上呈现一种“特殊的离散”状态，然而这些离散样本中其实隐含着类簇主题的分布规律；这些隐含的规律难以用肉眼识别，也难以使用简单的计算方式得到，因此我们需要选择一些强大的工具来帮我们发现这其中的规律；ArcGIS 强大的空间分析工具可以帮助我们来完成这个复杂的计算过程，使用 ArcGIS 中的核密度分析、渔网栅格化分析来对样本类簇进行分析，将样本中隐含的规律以直观的方式展现在我们眼前。

3.3.1. 核密度分析

核密度分析，是一种对输出象元附近或者邻域内的点要素或线要素的密度进行计算的分析方法，本文针对的是样本点进行核密度分析，所以主要介绍点的核密度分析过程。

点的核密度分析其关键所在是计算核表面的值；核表面是笼罩在点要素之上的平滑曲面，一个点要素对应的核表面的最大值是在该要素的正上方，核表面的值随着与点要素的距离增大而减小；核密度分析的过程是计算叠加在输出象元中心点之上的所有核表面值的和，在 ArcGIS 中核表面与底部平面相交组成的空间的体积记录为 Population 要素，此要素在为空值时体积为 1。单点核密度计算公式如式 1 所示：

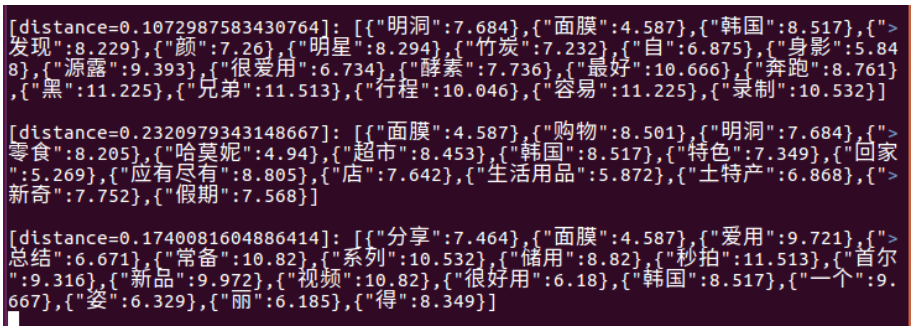


Figure 4. Viewing similar clusters
图 4. 查看相似簇

$$p(x)=\frac{1}{nh}\sum_{i=1}^{n}\left\{K\left[\frac{d(x,x_i)}{h}\right]\right\} \quad \text{(公式 1)}$$

式中的 K()为核函数，h 为阈值距离，n 为样本数量；在本小节的实验中对商圈类簇进行核密度分析，针对北京市中心城区使用核密度分析挖掘出该区域的热门商圈，实验数据只保留微博正文的签到信息，如表 4 所示。

实验大致流程如下：

1) 将样本点导入 ArcGIS 中，从全国县级矢量地图中选出北京市中心城区的矢量要素，建立新图层。

2) 使用北京城区图层为范围裁剪样本点，去除不在范围内的样本点，得到结果 4 千多个北京市中心城区样本点，如图 5 所示。

3) 使用 ArcToolBox 中的核密度分析工具来分析，输入数据为裁剪后剩余的样本点；处理后结果如图 6 所示。

经过核密度分析，可以明显的看出北京市中心城区的核心商圈分布情况，其中热点最高的是北京 CBD 核心区、工体、建外 SOHO、交大商圈。

3.3.2. 渔网栅格化分析

离散的位置数据点呈现在地图上时很难发现他们之间的空间信息，而商圈对周围区域的辐射影响使

Table 4. Experimental data
表 4. 实验数据

微博 ID	经纬度	签到位置
5a0d643dd56d0f10648a90f6	116.45749, 39.91401	北京 CBD
5a0d6740d56d0f10648a9730	116.46146, 39.90721	北京建外 SOHO
5a0d656fd56d0f10648a91d2	116.44234, 39.93321	北京工体

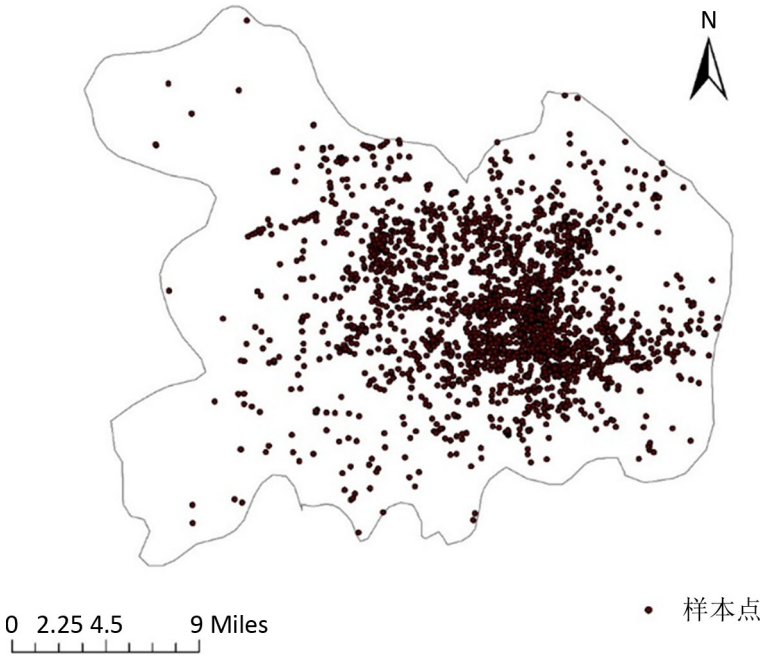


Figure 5. Sample points
图 5. 样本点

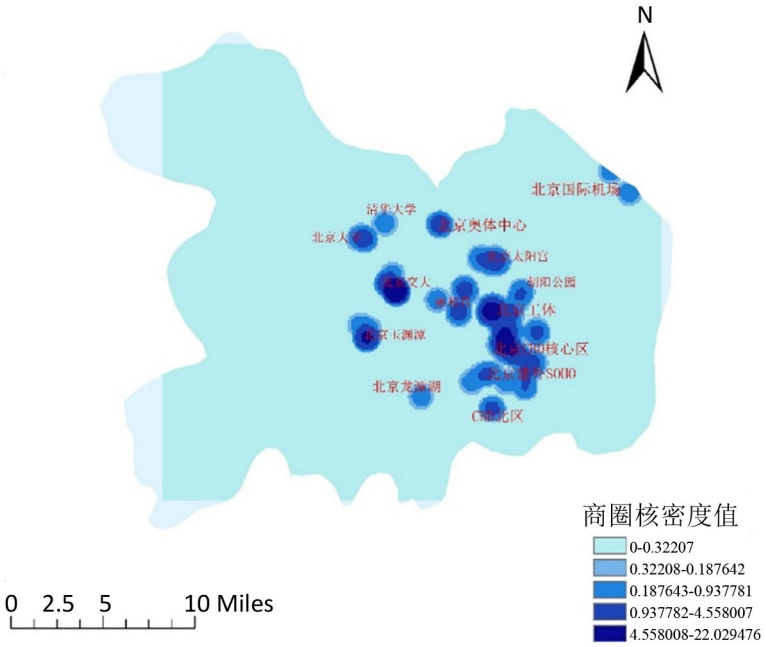


Figure 6. Nuclear density analysis
图 6. 核密度分析

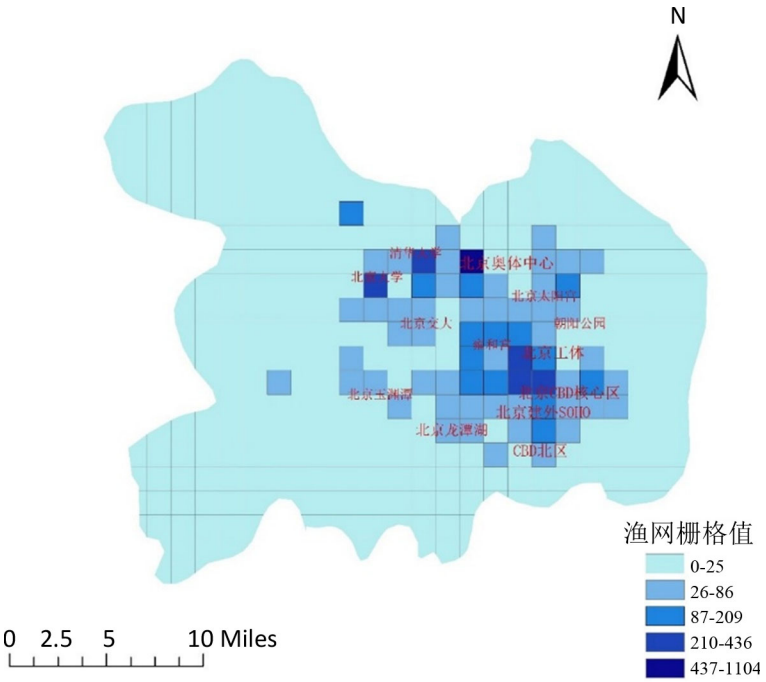


Figure 7. Rasterized circle hotspot map
图 7. 栅格化商圈热点图

得商圈的分布通常呈现临近性，因此选择使用栅格格网来覆盖样本点，以使空间上离散的样本点变成空间上相互临近的栅格格网。

每一个栅格格网的热度受其涵盖的点的数量的影响，大致流程如下：

1) 仍以北京市中心城区矢量图为底图，裁剪后的样本点为输入数据。

2) 使用 ArcToolBox 中的 fishnet 工具创建栅格格网，依据样本分布特征建立对应大小的渔网栅格。

3) 计算每个渔网栅格的属性值为栅格单元中包含的样本点个数；得到的结果如图 7 所示。

如图 6 所示，栅格化处理后的商圈的热度与核密度分析的结果相类似，但是在栅格化分析的结果中，可以较清晰的看出商圈的临近性，以及对商圈周边区域的影响，在两个商圈之间的过渡区域呈现热度下降，但是受附近商圈的影响仍然有一定的人气。

4. 结论

面向网络自媒体的空间数据挖掘，经过 Canopy 算法优化之后的 k-means 算法会比常规的 k-means 算法的聚类质量有较明显的提高，但是对于文本集所涉及的主题不会产生显著影响，但从簇间距离可以看出，Canopy 优化后的 k-means 能够提升类簇间的距离，从而降低类簇间的相似性，使产生的类簇更有代表性；而针对聚类结果的用户间相似性度量首先筛选出文本相似的用户，再在此基础上筛选出空间位置相邻的用户，通过文本和地域背景的双重比较来判别用户之间的兴趣和地域文化的相似程度；使用 ArcGIS 做可视化分析可以很好地呈现出城市热点商圈所在地区，而且，除了商圈主题的类簇，其余类簇同样可以使用此类分析方法来发现话题密集区域或热点区域的位置。

因为数据的来源是使用网络爬虫随机爬取的，所以爬取到的数据较为片面，不会有很强的主题针对性，且通过聚类挖掘出的信息也只是象征着样本集内涉及用户的热门话题，而不一定是代表当今社会的真实热门话题和热搜关键字；不是针对某一特定主题进行的数据爬取使聚类结果较为宽泛，但在当今的大数据环境下，本文的研究也能够为巨量的离散的随机数据的数据挖掘提供有用的参考依据。

基金项目

国家自然科学基金资助项目(4156010389)。

参考文献

[1] 杨桂满. 自媒体时代中学思想政治教育的策略研究[D]: [硕士学位论文]. 大连: 辽宁师范大学, 2016.

[2] 陈月华, 王雪. 自媒体环境中的传播暴力研究——以微电影为例[J]. 当代电影, 2013(5): 131-134.

[3] 杨飞, 江南, 李响, 张晶, 戴兵. 基于多策略的微博位置数据获取方法研究[J]. 测绘科学技术学报, 2016, 33(2): 201-207.

[4] 李德仁, 王树良, 李德毅, 王新洲. 论空间数据挖掘和知识发现的理论与方法[J]. 武汉大学学报(信息科学版), 2002, 27(3): 221-233.

[5] 朱红春. 数字高程模型（DEM）空间数据挖掘研究[D]: [硕士学位论文]. 西安: 西北大学, 2003.

[6] 王丽鲲. 基于社交媒体地理数据挖掘的游客时空行为分析[D]: [硕士学位论文]. 上海: 上海师范大学, 2017.

[7] 李德仁, 程涛. 从 GIS 数据库中发现知识[J]. 测绘学报, 1995(1): 37-44.

[8] 周海燕. 空间数据挖掘的研究[D]: [博士学位论文]. 郑州: 中国人民解放军信息工程大学, 2003.

[9] 毕硕本, 耿焕同, 闾国年. 国内空间数据挖掘研究进展与技术体系探讨[J]. 地理信息世界, 2008, 6(1): 21-27.

[10] 徐胜华, 刘纪平, 胡明远. 空间数据挖掘与发展趋势探讨[J]. 地理与地理信息科学, 2008, 24(3): 24-27.

[11] 李际平, 房晓娜, 封尧, 孙华, 曹小玉, 赵春燕, 李建军. 基于加权 Voronoi 图的林木竞争指数[J]. 北京林业大学学报, 2015, 37(3): 61-68.

[12] 张彩彩. Voronoi 图的改进及其在林分空间结构优化中的应用[D]: [硕士学位论文]. 长沙: 中南林业科技大学, 2015.

[13] 徐明. 基于 Hadoop 的空间数据挖掘研究[D]: [硕士学位论文]. 西安: 陕西师范大学, 2014.

[14] 戴振民. 基于微博用户相似度的社交圈挖掘算法研究[D]: [硕士学位论文]. 武汉: 华中科技大学, 2016.

[15] Highland, F. and Hart, C. (2016) Unsupervised Learning of Patterns Using Multilayer Reverberating Configurations of Polychronous Wavefront Computation. *Procedia Computer Science*, **95**, 175-184. https://doi.org/10.1016/j.procs.2016.09.310