安全态势感知系统中 K-Means 算法的并行化研究

江佳希，谢颖华
（东华大学 信息科学与技术学院，上海 201620）

摘 要： 大数据环境下的网络安全事件层出不穷，安全态势感知系统的应用势在必行。通过挖掘日志数据并进行安全分析，可以实现对异常事件的追责与溯源，有效地减少网络安全事故的发生。针对传统 K-Means 算法时间开销大、执行效率低的问题，将改进 K-Means 算法在大数据计算框架 Hadoop 上实现并行化，来满足大数据下安全态势感知系统日志安全分析的需求。实验表明，改进后的算法在有效性和时间复杂度方面都优于传统算法。

关键词： Hadoop；安全态势；K-Means；数据挖掘
中图分类号： TP311　　**文献标识码：** A　　**DOI：** 10.19358/j.issn.2096-5133.2020.07.006
引用格式： 江佳希，谢颖华. 安全态势感知系统中 K-Means 算法的并行化研究[J]. 信息技术与网络安全，2020，39(7)：36-40，51.

Research on parallelization of K-Means algorithm in security situation awareness system

Jiang Jiaxi，Xie Yinghua
(School of Information Science and Technology，Donghua University，Shanghai 201620，China)

Abstract： With the emergence of network security events in a big data environment，the application of security situation awareness systems is imperative. By digging log data and performing security analysis，we can achieve accountability and traceability to abnormal events，and effectively reduce the occurrence of network security incidents. Aiming at the problems of large time overhead and low execution efficiency of the traditional K-Means algorithm，the security situation awareness system in this paper improves the K-Means algorithm to achieve parallelization on the big data computing framework Hadoop，and to meet the needs of log security analysis under big data. Experimental results show that the improved algorithm is superior to traditional algorithms in terms of effectiveness and time complexity.

Key words： Hadoop；security situation；K-Means；data mining

0 引言

随着大数据时代的来临，SQL 注入攻击、XSS 攻击等网络安全事件层见叠出，给网络安全带来了巨大的挑战。日志记录着设备运行状态，各种安全事件都会在系统中留下日志记录，通过对日志进行分析，可以挖掘重要信息，实时掌握网络安全状况，既可做到事前防护，又可做到事后追本溯源及责任追查。

本文设计的安全态势感知系统将采集到的日志文件送至分布式文件系统 HDFS 进行存储，在 Hadoop 架构上将改进的 K-Means 算法和 MapReduce 高效的并行计算能力相结合，对存储的日志进行聚类和分析。安全态势感知系统可以实时监控网络安全态势，实现日志分析追责，有效地减少网络安全事故的发生。系统采用高可用部署模式，具有可靠、易拓展、易维护以及可视化的特点[1]。

1 系统总体架构及工作原理

1.1 总体架构

基于 Hadoop 大数据架构的安全态势感知系统通过对安全日志的实时分析，可以对网络安全环境实现实时监测。针对现存威胁，系统以实时和统计的方式对安全告警进行展现，针对未知威胁，系统从外部威胁和脆弱性两个方面进行预警。

安全态势平台的总体软件架构由数据层、分析

层、展示层三部分组成。数据层处于整个系统框架的底层，主要完成的是数据源的采集与存储，将不同来源的日志切成大小相同的数据片后送到各个节点。分析层调用 Map 和 Reduce 函数库将任务进行拆分作并行计算，结合 K-Means 聚类算法进行聚类结果分析。在展示层，用户可以用 Hive 语句查询结构化的数据，对于非结构化的数据，可以用 MapReduce 进行分析[2]。系统架构图如图 1 所示。

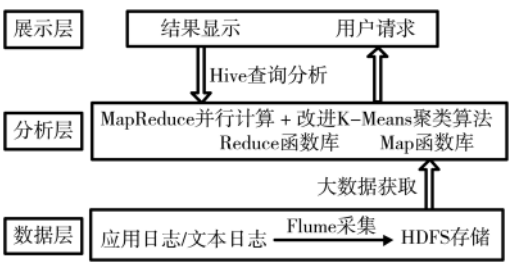

图 1 系统架构图

1.2 系统工作原理

安全态势感知系统包含四个模块，分别是数据采集、数据处理、安全分析以及态势呈现模块。系统的功能模块图如图 2 所示。

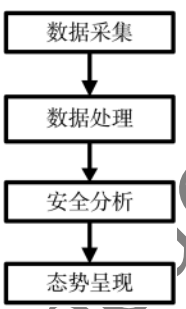

图 2 功能模块图

数据采集模块用 Flume 采集器收集系统日志，利用 web+ssh 获取采集程序的状态，将不同数据来源的日志数据进行高效收集、聚合、移动，最后存储到一个中心化数据存储系统 HDFS 中，实现各种网络设备、安全设备、服务器、应用服务等设备的日志采集。在 HDFS 底层，对日志进行切片，一般默认切片大小为 64 MB，存放至不同的数据节点。数据处理模块采用 kafka 集群，对采集到的数据进行预处理，预处理过程中首先对数据进行清洗过滤，通过指定过滤规则，将重复数据、噪声点以及不合理的数据从数据集中去除，然后对异构原始数据作格式统一处理，最后补齐字段、做数据标准化等处理，并将标准数据加载到 HDFS 系统中。安全分析模块是安全态势感知系统的核心，通过制定风险模型，对预处理后的数据集进行聚类和分析，将并行化编程模型 MapReduce 与改进的 K-Means 聚类算法相结合，完成对日志的聚类分析并形成追责证据链，可以识别恶意扫描、SQL 注入攻击、XSS 攻击等威胁事件在内的多类安全事件。态势呈现模块呈现将安全数据按各种场景分析之后得出的结果，利用各类图表直观地展现数据分析结果[3]。

2 K-Means 算法及其改进

2.1 K-Means 算法

K 均值聚类算法(K-Means Clustering Algorithm)由 James MacQueen 在 1967 年提出，基于相似度和距离这两个方面的技术已经非常成熟，而且由于实现简单、收敛速度快，K-Means 算法常用于日志挖掘。K-Means 算法的主要思想是将用户与服务器交互的各维数据作为特征进行聚类，对一部分非数值类型的特征需要进行转化和标记，使其适用于与质心之间的距离的计算[4]。传统的 K-Means 算法将用户的数据集 O 和 K 值作为输入，随机自动选取 K 个聚类中心，通过距离计算将数据点划分到离它最近的聚类中心所在的簇，直至收敛，输出 K 个簇。算法描述如下：

(1)输入 K 值和数据集 O。

(2)从数据集 O 中选取 K 个值作为初始聚类中心，记为 $X_1, X_1, \cdots, X_K$，选取方式为随机选取。

(3)将已经被选为聚类中心的点从数据集中去除，新的数据集记为 O_1。

(4)对于 O_1 中的数据点，计算其到 K 个初始聚类中心的欧式距离，按照就近原则将数据点划分到距离最近的簇中。

(5)K 个簇初步划分完成后，计算簇内均值，记为新的聚类中心，重新计算数据点到中心点的距离，进行新一轮的划分，直至通过准则函数判断聚类中心点收敛，聚类停止，否则迭代更新中心点。

(6)输出 K 个类簇。

2.2 改进 K-Means 算法

安全态势感知系统要实现实时告警和威胁预测功能，要求算法具有可伸缩性，不管是处理小数据集还是大数据集，都要具有高效的特点。同时，安全态势感知系统要求算法准确率高，对于干扰点能准确识别并去除[5]。

传统的 K-Means 算法采用单机部署，执行效率低下，且 K-Means 的性能与聚类中心的位置关系很密切，初始聚类中心的选择极大程度地影响聚类效果。由于传统 K-Means 算法在选取初始聚类中心的随意性，假如将孤立点选为聚类中心，聚类准确性会大打折扣，同时算法循环迭代次数也会增多，且

少量的孤立点会使得类内均值计算出现大的偏差。

针对 K-Means 在数据处理方面的缺陷，本文提出一种改进 K-Means 算法。新算法在去除孤立点干扰、寻找合适的聚类中心、判断中心点收敛的方式等方面做出了改进，改善了传统 K-Means 算法耗时长、执行效率低等问题[6]。改进后的算法流程图如图 3 所示。

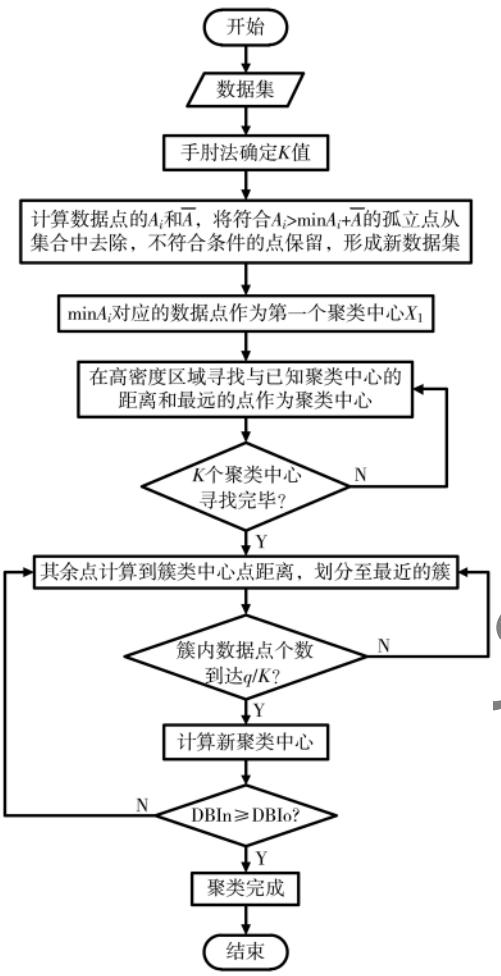

图 3 改进 K-Means 算法流程图

改进 K-Means 算法描述如下：

(1)设数据集中数据量为 m，定义 A_i 为数据点到数据集中其他点的距离之和，计算公式如下：

$$A_i=\sum_{j=1}^{m}\sqrt{\sum_{n=1}^{M}(x_{in}-x_{jn})^2} \tag{1}$$

计算数据集中所有点距离和的均值$\overline{A}$，定义如下：

$$\overline{A}=\sum_{i=1}^{m}\frac{A_i}{m} \tag{2}$$

扫描数据集，寻找符合 $A_i>\min A_i+\overline{A}$ 的数据点，由于数据点距离和远远大于均值，将其认为是孤立点，从数据集中去除。$\min A_i$ 为距离和的最小值，数据点到数据集中其他点的距离和越小，证明该点周围聚集的数据点越多，即密度越大。记去除孤立点后得到的新的数据集为 P，新数据集中的数据量记为 q。

(2)令 $K=\sqrt{q}$ 为聚类数。从集合 P 中找出 $\min A_i$ 对应的点作为第一个聚类中心 X_1，同时将 X_1 从 P 集合中去除。

(3)将符合$A_i<\overline{A}-\min A_i$ 且距 X_1 最远的数据点 X_2 作为第二个聚类中心，确保 X_2 和 X_1 来自于两个不同的簇，同时将 X_2 从集合 P 中去除。此处将符合 $A_i<\overline{A}-\min A_i$ 的数据点定义为高密度点。

(4)在高密度区域，即符合 $A_i<\overline{A}-\min A_i$ 条件的数据集中寻找与前两个数据点距离之和最远的数据点 X_3 作为第三个聚类中心，并将 X_3 从 P 集合中删除。

(5)依次类推，在高密度区域寻找与已确定的簇中心距离之和最大的数据点，同时将这些点从集合 P 中去除，直到 K 个聚类中心搜寻完毕。

(6)对于 P 中剩余的数据点，分别计算其到 K 个中心的欧式距离，根据就近划分原则将数据点都划分到离它最近的簇中。

(7)当簇内数据点个数达到阈值 $\frac{q}{K}$ 时，该簇初次划分完毕，原本应该划分到该簇的数据点划分到离它次近中心点的簇中。

(8)各个饱和类簇分别计算簇内各个维度的算术平均值，形成新的聚类中心。

(9)计算新一轮迭代和上一轮迭代的 DBI 指标值，分别用 DBIn 和 DBIo 表示，若 DBIn<DBIo，则开始新一轮的迭代，直到 DBIn>DBIo 为止。DBI 计算方式如下：

$$\mathrm{DBI}=\frac{1}{K}\sum_{i=1}^{k}\max\frac{S_i+S_j}{d(i,j)} \tag{3}$$

其中，S_i 代表第 i 个聚类中数据点与聚类中心的标准误差，S_i 越小代表类内相似度越高。$d(i,j)$代表聚类中心之间的欧式距离，$d(i,j)$越大代表类间距离越大，类间相似度越低。DBI 可以描述聚类效果，DBI 的值越小，代表类内相似度高，且类间相似度低，聚类效果越好[7]。

2.3 改进算法的并行化实现

MapReduce 编程模型由谷歌提出，可以对大规模的数据集作并行运算。改进后的 K-Means 算法的迭代过程主要有两个步骤：(1)将数据集中的数据点进行分类；(2)计算中心点。由于数据点运算时完全独立，互不干扰，且计算中所用的公共变量不多，可采用 MapReduce 并行计算编程模型实现，一次迭代为一个任务，由 Map 函数、Combine 函数、Reduce 函数实现，Map 函数主要用于计算距离，对数据集中的数据点进行分类，Combine 函数对 Map 函数输出的键对值进行合并预处理，Reduce 函数以 Combine 函数的输出作为输入，负责重新计算中心点，判断是否符合阈值条件，不符合则开始新一轮的迭代[8]。实现过程如下：

(1)数据初始化，将日志文件进行 n 维切片，将大小相同的数据块存入 HDFS。

(2)将切片和副本送至对应的数据节点。

(3)节点调用 Map 函数提取出数据块，计算距离和，选择初始聚类中心，计算点到聚类中心的距离，得到 K 个聚类中心。

(4)前一阶段函数 Map 的结果进入 combine 阶段进行归并处理，为 Reduce 函数提供标准化输入。

(5)Reduce 函数对输入的候选聚类中心计算新一轮迭代和上一轮迭代的 DBI 指标值，若 DBIn<DBIo 则开始新一轮的迭代，直到 DBIn>DBIo 为止，判定中心点收敛，任务结束，否则，更新中心点并且开始新一轮的迭代。

MapReduce 并行化流程图如图 4 所示。

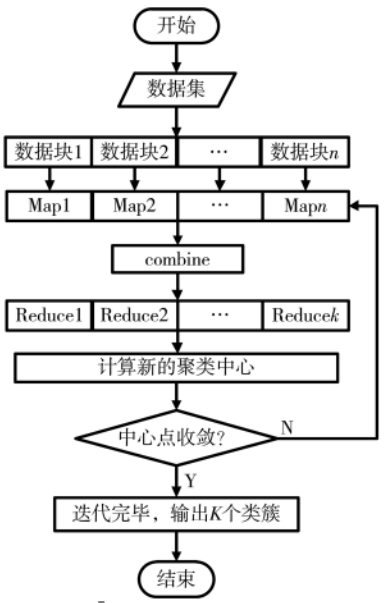

图 4 MapReduce 并行化处理流程

3 实验结果与分析

为了体现改进 K-Means 算法的有效性和准确性，本文采用了传统的 K-Means 算法和改进后的并行 K-Means 算法对同一组数据集进行聚类。实验数据来自于某安全态势感知网站 90 天的日志文件，总计 4.5 GB 大小。日志在处理之前，需先经过标准化处理。例如，原始日志格式为：time:2019-12-15_13:42:30;danger_degree:1;breaking_sighn:0;event:[50193]某某入侵攻击;src_addr:192.168.10.244;src_port: 138;dst_addr:192.168.10.255;dst_port:138;proto:UDP.NETBIOSDGM;user:xiaoming。

经过标准化、过滤、补齐、关联标签等流程后，日志数据如表 1 所示[9]。

表 1 预处理后的日志数据

字段	值	字段说明
COLLECT_TIME	2019-12-15_13:42:30	事件发生时间
SERVERITY	1	风险等级
ACTION	0	是否阻断
ID	[50193]某某入侵攻击事件	事件 ID 事件名
SRC_IP	192.168.10.244	源 IP
SRC_PORT	138	源端口
DST_IP	192.168.10.255	目的 IP
DST_PORT	138	目的端口
PROTOCOL	UDP.NETBIOSDGM	协议
SRC_USER	Xiaoming	用户
*PROVINCE	北京	源 IP 地址位置
*ASSET_NAME	SMP 采集服务器	目的资产名
*ASSET_USER	张三	资产负责人
*USER_PHONE	13675171277	资产负责人联系方式

3.1 算法有效性验证

衡量改进后 K-Means 算法有效性的重要指标是观察算法的迭代次数和准确率相较于优化前有无改进。准确率 $=\frac{u}{m}\times 100\%$，代表放入聚类簇中的数据点个数占数据集总数的比例。分别使用传统的 K-Means 算法和优化后的 K-Means 算法进行实验，结果如图 5、图 6 所示。

从图 5 可以看出，由于采用 MapReduce 并行化架构，对于同一数据集，改进后的算法迭代次数少于传统 K-Means 算法，随着数据量的增大，优势更

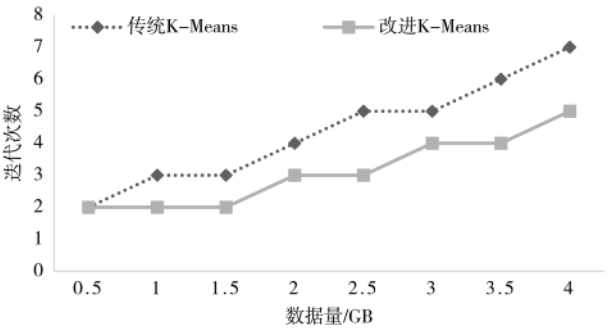

图 5 传统 K-Means 算法与改进的 K-Means 算法迭代次数对比

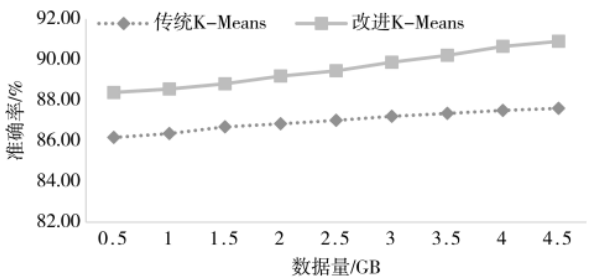

图 6 传统 K-Means 算法与改进的 K-Means 算法准确率对比

明显。

从图 6 可以看出，改进后的 K-Means 算法准确率有所提升，有更多数据点被正确划分到类簇中，聚类效果更佳。

3.2 算法时间验证

算法聚类时间是另一个衡量算法性能的重要指标，针对同一数据集，算法耗时越短代表执行效率越高。本文将并行 K-Means 优化算法和串行传统 K-Means 算法在聚类时间上进行了对比，改变数据集的大小进行实验。实验结果如图 7 所示。

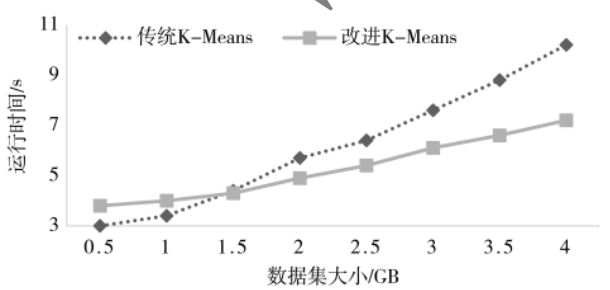

图 7 传统 K-Means 算法与改进的 K-Means 算法聚类时间对比

由图 7 可以看出，当数据量较小时，传统 K-Means 算法稍快于改进并行 K-Means 算法，随着数据量增大到 1.5 GB 时，改进算法运行时间开始比传统算法短，且随着数据集的增大，优势越来越明显。产生这种现象的原因是当数据量较小时，改进 K-Means 算法采用了 MapReduce 并行架构，较传统串行算法更耗时，随着数据量的增大，改进算法对数据片进行并行处理的优势显现出来，算法运行时间大大缩短，证明了改进 K-Means 并行算法更适用于大数据处理[10]。

4 结论

本文提出的改进 K-Means 算法克服了传统 K-Means 算法在寻找初始聚类中心的随机性，克服了孤立点对聚类效果的影响，降低了算法时间开销，提升了执行效率，并基于 MapReduce 实现了并行化，使其适应于大数据处理要求。通过实验，证明了新算法在有效性和时间复杂度方面都优于传统 K-Means 算法。基于该算法实现的日志追责系统具有可靠性高、可拓展性强、执行效率高等优点，有一定的实用价值。

参考文献

[1] 曹蓉蓉.大数据环境下网络安全态势感知研究[J].数字图书馆论坛，2014(2)：11-15.

[2] ANSARI Z，AFZAL A，SARDAR T H.Data categorization using Hadoop MapReduce-based parallel K-Means clustering[J].Journal of the Institution of Engineers(India)：Series B，2019，100(2)：95-103.

[3] 张舒婷.基于大数据分析的网络安全态势评估[J].现代电子技术，2019，42(13)：106-109.

[4] XU H，YAO N，HAN Q，et al.Parallel implementation of K-Means clustering algorithm based on MapReduce computing model of Hadoop[J].Metallurgical and Mining Industry，2015，7(4)：213-223.

[5] YANG J，LI X P.MapReduce based method for big data semantic clustering[C].2013 IEEE International Conference on Systems，Man，and Cybernetics(SMC 2013)，2013：2814-2819.

[6] 毛远军.基于大数据的网络安全态势感知技术分析[J].科学与信息化，2019(5)：54，59.

[7] SUN Z.A parallel clustering method study based on MapReduce[C].Proceedings of the 1st International Workshop on Cloud Computing and Information Security(CCIS 2013)，2013，52：416-419.

[8] 董超，刘雷.基于安全态势感知在网络攻击防御中的应用[J].网络安全技术与应用，2019(8)：22-23.

(下转第 51 页)

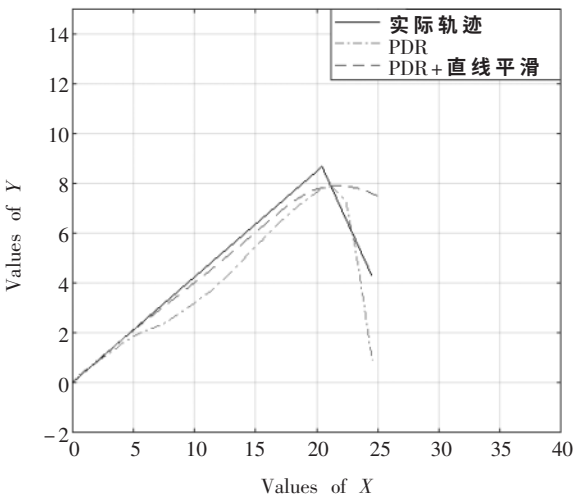

图 1 直线为主的轨迹

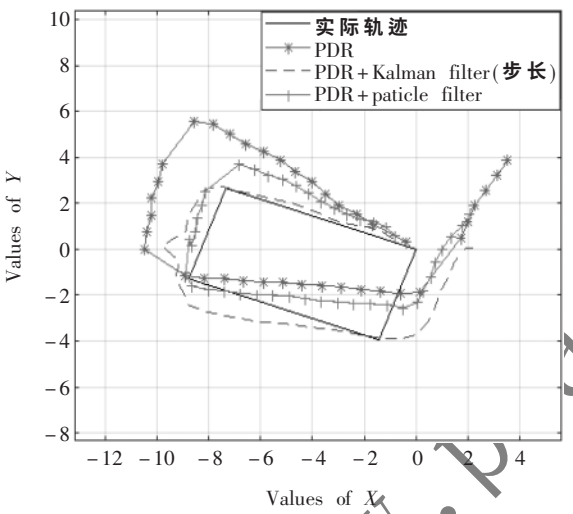

图 2 曲线为主的轨迹

轨迹，其中PDR算法的估算轨迹误差为6.20 m，采用卡尔曼滤波平滑处理步长数据后的误差降为1.67 m，在卡尔曼滤波的基础上采用粒子滤波优化结果的轨迹误差降为1.08 m。本文的误差采用公式(17)计算所得：

$$e=\sqrt{(x-x_0)(y-y_0)} \tag{17}$$

其中，(x,y)表示定位坐标，(x_0,y_0)表示实际坐标。

4 结论

本文提出的针对两种场景（直线为主的轨迹、曲线为主的轨迹）下的优化方案能够提高PDR算法的精度，具有一定的使用价值。但是PDR算法所得结果都是相对的，即需要一个初始的参考位置，在实际应用中需要加以解决。如与基于Wi-Fi或者蓝牙的指纹法相结合，指纹法能够提供初始的参考位置或者校准PDR算法的累积误差。除此外，还可以与其他技术相结合进行融合定位，如地图匹配、滤波融合等，这也是之后的一些可供参考的研究方向。

参考文献

[1] 李晓阳.WiFi技术及其应用与发展[J].信息技术，2012(2)：62.

[2] 苑宝玉.超声波室内定位系统[D].长春：长春理工大学，2010.

[3] 杨洲，汪云甲，陈国良，等.超宽带室内高精度定位技术研究[J].导航定位学报，2014，2(4)：30-35.

[4] 周梓鑫.RFID技术的发展和起源[J].黑龙江科技信息科技论坛，2013(18)：84.

[5] 汪苑，林锦国.几种常用室内定位技术的探讨[J].中国仪器仪表，2011(2)：54-58.

[6] KAPPI J，SYRJARINNE J，SAARINEN J.MEMS-IMU based pedestrian navigator for handheld devices[J].Ion Gps，2001.

(收稿日期：2020-03-10)

作者简介：

刘玲玉(1992-)，女，硕士研究生，主要研究方向：单片机应用、信号的检测与处理等。

刘狄松(1992-)，男，硕士研究生，主要研究方向：信号的检测与处理、传感器的研究与应用等。

常铁原(1964-)，男，副教授，主要研究方向：信号的检测与处理等。

(上接第40页)

[9] YANG Y，LONG X，JIANG B.K-means method for grouping in hybrid MapReduce cluster[J].Journal of Computers，2013，8(10)：2648-2655.

[10] 王进，姜新超，孙佳伟.数据仓库在网络安全态势感知中的设计与实现[J].网络安全技术与应用，2019(4)：54-56.

(收稿日期：2020-03-21)

作者简介：

江佳希(1996-)，女，硕士研究生，主要研究方向：网络安全、数据挖掘。

谢颖华(1972-)，女，硕士，副教授，主要研究方向：大数据、数据挖掘等。

版权声明