基于混沌关联维特征提取的大数据聚类算法

谢 川
(空军工程大学航空航天工程学院 西安 710038)

摘 要 大数据聚类过程是一个随机的非线性处理过程,具有很高的不确定性。由于传统方法需要先验知识进行学习,不能很好地适应大数据的实时变化情况,无法有效实现大数据聚类,因此提出一种基于混沌关联特征提取的大数据聚类算法。分析了传统方法的弊端,通过重构相空间建立了一个多维的状态空间向量与混沌轨迹,使原系统中很多几何特征量保持不变,为分析原系统的混沌特征提供有效依据。将平均互信息量取第一个最小值时的横坐标所指的时间延迟作为重构相空间的最佳时间延迟,采用虚假最近邻点算法对最佳嵌入维数进行选择。将提取的关联维数这一特征量作为大数据聚类的混沌特征量,依据提取的混沌关联维特征对大数据进行聚类。仿真实验表明,所提算法能够有效提高数据的聚类效率,减少能耗,是一种有效的数据聚类方法。

关键词 混沌关联维特征,大数据,聚类
中图法分类号 TP399 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.6.046

Big Data Clustering Algorithm Based on Chaotic Correlation Dimensions Feature Extraction

XIE Chuan
(College of Aeronautics and Astronautics Engineering, Air Force Engineering University, Xi'an 710038, China)

Abstract Big data clustering process is a kind of stochastic nonlinear processing and has very high uncertainty. Because the traditional methods need prior knowledge to learn, are not good to adapt to the real-time change situation of big data and unable to effectively implement large data clustering, we put forward a kind of big data clustering method based on chaotic correlation feature extraction. We analyzed the disadvantages of the traditional methods, established a multidimensional state space vector and the chaotic trajectory by phase space reconstruction. Much of the geometry characteristic information in the original system remains same, which provides the effective basis for the analysis of chaotic characteristics of the original system. Time delay referred by the abscissa when the average mutual information obtains the first minimum is as the best time delay of reconstructing phase space, and the false nearest neighbor algorithm is used to select the best embedding dimension. The extracted correlation dimension is used as the haotic correlation characteristics of bige data clustering, and big data is clustered based on the extracted chaos correlation dimension feature. The simulation results show that the proposed algorithm can effectively improve the efficiency of the clustering of data, reduce energy consumption, and is an effective method of data clustering.

Keywords Chaos correlation dimension feature, Big data, Clustering

1 引言

随着计算机处理与存储容量的持续增长以及硬件和软件成本的逐渐降低,根据实际使用需求,收集的数据在采集过程中呈爆炸式增长[1,2]。对大数据的有效管理成为当前计算机信息化管理的真正挑战[3,4]。因此,研究大数据的聚类方法具有重要意义,已经成为相关学者研究的重点课题,受到越来越广泛的关注[5,6]。

大数据聚类就是对海量数据进行聚类的过程,常用的大数据聚类算法主要包括BP神经网络算法、云计算算法、分布式算法,相关算法的研究取得了一定的成果,其中文献[7]提出一种基于云计算的大数据聚类算法,利用云计算在数据存储、数据管理和虚拟化等方面的技术优势,构建了基于云计算的大数据管理和处理模式,但该方法存在计算过程复杂的问题;文献[8]提出一种基于LabVIEW的大数据聚类算法,应用位置标记的方法,通过循环分批读取,解决了大数据块文本数据的快速聚类难题,但该方法需要建立统一的数据模型来对其进行管理,实现过程非常复杂;文献[9]提出一种基于Hadoop云计算平台的大数据聚类算法,能够在一定程度上解决Hadoop系统架构下数据复杂结构查询的局限性问题,但没有考虑数据查询中的实时性问题。文献[10]提出一种基于神经网络的大数据聚类算法,提升了数据聚类性能,但在算法的改进过程当中并未考虑重复读取操作对大数据聚类效率的影响。文献[11]提出了一种基于分布式技术的大数据聚类算

到稿日期:2015-08-16 返修日期:2015-09-16 本文受陕西自然科学基金:无铅焊点在多场耦合作用下的失效行为及寿命预测方法(2015JM6345)资助。

谢 川(1974-),男,博士,副教授,主要研究方向为飞行数据智能处理、检测技术教学与科研。

法，利用分布式并行计算平台，对大型关系数据库平台进行优化改造，该方法具有很高的可行性，但其存在处理效率低的问题。

针对传统方法的弊端，本文提出一种基于混沌关联特征提取的大数据聚类算法，分析了大数据聚类原理，对相空间进行重构，为分析原系统的混沌特征提供有效依据。将提取的关联维数作为大数据的混沌特征量，依据该特征量，实现大数据聚类。经实验验证，所提算法能够有效提高数据的聚类效率，减少能耗。

2 大数据聚类原理

大数据聚类原理通过传播近邻信息得到最优的类代表点集合（一个类代表点对应为实际数据集中的一个数据点），通过该算法进行聚类的准确性较好，其基本思想如下：

假设一个数据集是 $\xi=\{e_1,\cdots,e_N\}$，用 $d(i,j)$ 描述数据 e_i 与数据 e_j 之间的距离，也就是 e_i 与 e_j 之间的相似度；用 K 描述一个正整数，ξ 的 K 聚类中心点问题就是从数据集 ξ 中获取 K 个具有代表性的中心点 $e_{i_1},\cdots,e_{i_K}$。上述中心点可使全部数据之间的相似度总和达到最小，也就是全部数据点 e_j 和其聚类中心代表点 e_{i_K} 之间的总和达到最小。

本节给出了一种找出 K 中心点的有效方法，同时符合上述相似度之和最小的条件。用 $\sigma(i)$ 描述和数据点 e_i 最近的中心点，则该算法的目的就是找到最合理的一组 σ，使得下述定义的 $E[\sigma]$ 达到最大化。

$$E[\sigma]=\sum_{i=1}^{N}S(e_i,e_{\sigma(i)})-\sum_{i=1}^{N}X_i[\sigma] \tag{1}$$

将欧氏距离看作是数据之间的相似度时，$S(e_i,e_j)$ 被定义成 $-d(i,j)^2$，也就是 $S(e_i,e_j)=-\|e_i-e_j\|^2$。如果 $i=j$，则将 $S(e_i,e_j)$ 称作是数据点 i 的参考度，用 $P(i)$ 或 s^* 进行描述。随着 $P(i)$ 值的逐渐增加，数据点 i 成为聚类代表点的可能性也逐渐增加。式(1)中右边第二项说明，如果将 e_i 作为数据集中某些数据点的代表中心点，则其一定是自身的中心点，也就是当 $\sigma(\sigma(i))\neq\sigma(i)$，则 $X_i[\sigma]=\infty$；否则 $X_i[\sigma]=0$。

式(1)除了可约束数据集相似度总和最小之外，也可在大数据聚类时使各数据项的平方误差 $d(i,\sigma(i))^2$ 之和失真最小。式(1)中并未直接给出需聚类成多少个簇，可通过参考度 s^* 判断一个数据点是否适合成为中心代表点，通常取 S 相似均值作为 s^* 的值。

将式(1)作为约束条件对大数据进行聚类，是通过数据点间不断传递消息获取最优结果的。分析图 1 可以看出，算法中传递的是下述两个消息。

Responsibility：$r(i,k)$ 用于描述数据点 k 被选作数据点 i 的聚类中心的程度；

Availability：$a(i,k)$ 用于描述数据点 i 选择点 k 作为其聚类中心的程度。

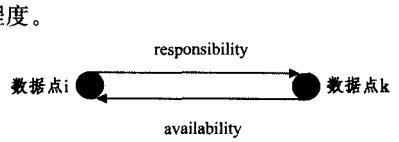

图 1 数据点之间传递消息示意图

在开始时，全部 Responsibility 与 Availability 消息均被置为 0，其值可通过下述公式不断进行迭代调整：

$$r(i,k)=S(e_i,e_k)-\max_{k',k'\neq k}\{a(i,k')+S(e_i,e_{k'})\} \tag{2}$$

$$r(k,k)=S(e_i,e_k)-\max_{k',k'\neq k}\{S(e_i,e_{k'})\} \tag{3}$$

$$a(i,k)=\min\{0,r(k,k)+\sum_{i',i'\neq i,k}\max\{0,r(i',k)\}\} \tag{4}$$

$$a(k,k)=\sum_{i',i'\neq k}\max\{0,r(i',k)\} \tag{5}$$

数据点 e_i 的代表中心点最终可被描述如下：

$$\sigma(i)=\arg\max\{r(i,k)+a(i,k),k=1,\cdots,N\} \tag{6}$$

如果代表中心点在一定次数迭代过程中并未发生改变，则算法将停止，同时将聚类结果作为最终的结果。

通过上述原理对大数据进行聚类，计算过程相对简单，而且聚类准确度较好，但其需要先验知识进行学习，不能很好地适应大数据的实时变化情况，无法有效实现大数据聚类，因此提出一种基于混沌关联特征提取的大数据聚类算法。

3 基于混沌关联维特征提取的大数据聚类算法

3.1 混沌关联维特征提取

大数据中的混沌特征通常表现为无明显规则和次序、非同期性的复杂的折叠和扭曲，混沌特征非常复杂，需采用关联维数对其进行描述。

3.1.1 相空间的重构

数据序列很大程度上属于非线性时间序列，而非线性时间序列的关键是相空间重构，相空间重构能够使原系统中很多几何特征量保持不变，建立了原始时间序列和多维空间分析的桥梁，在多位相空间中有效提取数据的混沌关联维特征。

相空间重构方法如下：假设时间序列是 $\{x_1,x_2,\cdots,x_N\}$，则相空间重构结果可描述成：

$$X=[X_1,X_2,\cdots,X_K]$$

$$=\begin{bmatrix} x_1 & x_2 & \cdots & x_K \\ x_{1+\tau} & x_{2+\tau} & \cdots & x_{K+\tau} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1+(m-1)\tau} & x_{1+(m-1)\tau} & \cdots & x_{M+(m-1)\tau} \end{bmatrix} \tag{7}$$

其中，$K=N-(m-1)\tau$，τ 用于描述时间延迟；m 用于描述嵌入维数。若 $m\geqslant 2d+1$，则动态系统的几何结构将被完全打开，其中 d 用于描述系统混沌吸引子的维数。

嵌入维数 m 和时间延迟 τ 的选择是相空间重构的关键，选择合理的 m 和 τ 才能准确重构反应原系统特征的相空间，下面给出详细的选择方法。

针对时间延迟 τ 的选择，本文将延迟时间互信息取第一个最小值时的横坐标所指的时间延迟 τ 作为重构相空间的最佳时间延迟。在数据分布的区间内，建立数据的概率分布曲线。用 p_i 描述 $x(t)$ 出现在数据分布曲线区间 i 内的概率；用 $p_{ij}(\tau)$ 描述 $x(t)$ 出现在 i 内和出现在延迟了一定时间量 τ 后的延迟 $x(t+\tau)$ 出现在区域 j 的联合概率。则延迟时间互信息可描述成：

$$I(\tau)=-\sum_{ij}p_{ij}(\tau)\ln\frac{p_{ij}(\tau)}{p_ip_j} \tag{8}$$

如果 $I(\tau)=0$，则 $x(t+\tau)$ 将无法预测，也就是 $x(t)$ 与 $x(t+\tau)$ 相互独立，并且 $I(\tau)$ 越小，$x(t)$ 与 $x(t+\tau)$ 越相互独立，所以可将 $I(\tau)$ 达到最小时，与横坐标相应的时间延迟 τ 作为重构相空间的最佳时间延迟。

针对嵌入维数 m 的选择，本文采用虚假最近邻点算法对其进行计算。根据 Takens 定理，在 m 维相空间中形成的 m 维矢量可描述成：

$$X(n)=\{x(n),x(n+\tau),\cdots,x(n+(m-1)\tau)\} \tag{9}$$

获取相空间重构的最小嵌入维需符合式(10)所描述的条件，若符合，则将 $X_{\eta(n)}$ 称作是 X_n 的虚假最近邻点。

$$\frac{|x_{\eta(n)+m\tau}-x_{n+m\tau}|}{\|X_{\eta(n)}-X_n\|_2^{(m+1)}}\geqslant R_{tol} \tag{10}$$

其中，R_{tol} 用于描述阈值，通常 R_{tol} 取 15。这时需求出虚假最近邻点占的比例曲线，若虚假最近邻点的比例低于 5%，则认为获取的 m 就是所求的相空间重构最小嵌入维数。

3.1.2 混沌关联维特征提取

本文将提取的关联维数作为大数据聚类的混沌特征量。在相空间重构的基础上，使一维时间序列在多维空间中得以扩展，从而提取混沌关联维特征。依据上节分析的过程，即可获取重构后的时间序列：

$$X_i=(x_i,x_{i+\tau},\cdots,x_{i+(m-1)\tau})^{\mathrm{T}} \tag{11}$$

通过上述过程重构的 m 维相空间中，相点 x_j 到除 x_i 本身外 x_i 的距离不超过 r 的点数可描述成：

$$Q=\sum_{j\neq i}H(r-\|x_i-x_j\|) \tag{12}$$

其中，$H(\cdot)$用于描述 Heavside 函数。

这里给出关联函数概念，将所有可能在距离上比给定的距离 r 小的点相对于占有总的点对数的比例称为关联函数，公式描述如下：

$$C_N(r)=\frac{2}{Q(Q-1)}\sum_{i=1}^{N}\sum_{j=i+1}^{N}H(r-\|x_i-x_j\|) \tag{13}$$

式中分子为 2 是为了排除重复计数，用范数描述两相点之间的距离，即可获取两相点之间的距离，也就是两个矢量的最大分差量：

$$\|x_i-x_j\|=\max_{1\leqslant k\leqslant m}|x_{i-(k-1)\tau}-x_{j-(k-1)\tau}| \tag{14}$$

针对距离不超过 r 的向量，可将其称作是具有关联性的矢量。假设一维实测序列数据是 n 个，则相空间重构中的向量点数量为 $N=m-(m-1)\tau$。求出相点中存在关联的相点对数占所有可能的 $N(N-1)/2$ 种配对的比例，将其称作关联维数，公式描述如下：

$$C_m(r)=\frac{2}{N(N-1)}\sum_{i=1}^{N}\sum_{j=i+1}^{N}H(r-\|x_i-x_j\|) \tag{15}$$

则上述获取的关联维数即为大数据聚类的混沌特征量，依据该关联维数实现大数据的聚类。

3.2 大数据聚类实现

所谓聚类分析就是将不同的样本划分成几类，同时使一个聚合类的样本比不同聚合类的样本更加相近。本文依据提取的混沌关联维特征对大数据进行聚类分析，详细实现过程如下。

(1)输入样本和参数。

输入 n 个数据样本$\{x_1,x_2,\cdots,x_n\}$，依据混沌关联维特征，从上述样本中选择 N 个聚类中心，用$\{z_1,z_2,\cdots,z_n\}$进行描述。

(2)将 n 个样本依据下述原则按照顺序划分至最近的聚类 ω_j：

$$\|x-z_j\|=\min(\|x-z_j\|) \tag{16}$$

其中，$\|x-z_j\|$ 用于描述 x 和 z_j 之间的距离。同时假设 ω_j 中存在 N_i 个样本。

(3)通过下式求出聚类中心值：

$$z_j=\frac{1}{N_j}\sum_{x\in\omega_j}xC_m(r) \tag{17}$$

(4)如果迭代次数是奇数，则直接进行步骤(6)；否则继续进行下一步骤。

(5)分裂。

假设 $L=\max(\|x-z_i\|),x\in\omega_i$，用 d_1 描述分裂距离。如果 $L>d_1$，则将 ω_i 分裂成两类，这时聚类中心可描述成：

$$\begin{cases}z_{i1}=z_i+\lambda L\\ z_{i2}=z_i-\lambda L\end{cases} \tag{18}$$

其中，λ 用于描述一个大于 0 的常数。如果 $L<d_1$，同时上一次未进行合并操作，则进行步骤(7)。

(6)合并。

假设 $l=\|z_I-z_J\|=\min\|z_i-z_j\|$，用 d_2 描述合并距离。如果 $l<d_2$，则将 ω_I、ω_J 合并成一类，合并中心可描述成：

$$z_{IJ}=\frac{1}{N_I+N_J}[N_Iz_I+N_Jz_J] \tag{19}$$

如果 $l>d_2$，同时上一次未分类，则进行步骤(7)，否则进行步骤(4)。

(7)结束迭代。

本文将具有相同混沌关联特征的数据通过上述聚类分析过程划分成一类，从而实现大数据的有效聚类。

4 仿真实验分析

为了验证本文提出的基于混沌关联特征提取的大数据聚类算法的有效性，需要进行相关的实验分析。将传统神经网络算法作为对比，主要比较两种算法的能量消耗及处理时间。本文通过模拟数据对算法进行验证，全部实验程序用 C++ 编写，在 Ubuntu 12.04 操作系统上运行，实验硬件平台为 Lenovo M4390 (i3-2100CPU,4 UB 内存,2TB 磁盘)，处理器 Intel(R)Core(TM)2 Duo CPU 2.94GHz,内存:8.00GB。实验选择了仿真参数设计中，大数据分组的产生的时间间隔为 0.1s,从仿真时间 300 s 时开始产生数据。实验中，数据量从 100 MB 到 1 GB,以 100 MB 为单位，数据呈非线性增长，对大数据进行离散调度和区间边界逼近，大数据特征采集的时间间隔为 0.1 s,参数配置如表 1 所列。

表 1 参数配置

参数	值(Mpbs)
数据数量	1000
大数据分布特征数量	5
每个数据访问系统载荷	16
数据复杂度大小(GB)	2
数据执行时间延迟(ms)	2400
最大队列大小	2200

分别采用本文算法和传统算法对不同数据量进行聚类，统计两种算法的聚类效率，结果如表 2 所列。

表 2 两种算法聚类效率比较结果

数据量(GB)	本文算法所需时间(s)	传统算法所需时间(s)
20	103	139
200	1924	5988
400	4432	12771
800	8342	29147
1024	10149	35823

分析表 1 可以看出，随着数据量的逐渐增加，本文算法和传统算法对数据聚类的时间均逐渐增加，但本文算法所需的处理时间一直明显低于传统算法，说明本文算法有很高的数据聚类效率，验证了本文算法的有效性。

为了进一步验证本文算法的有效性，本文对两种算法处理相同数据量所消耗的能量进行对比，结果如图 2 所示。

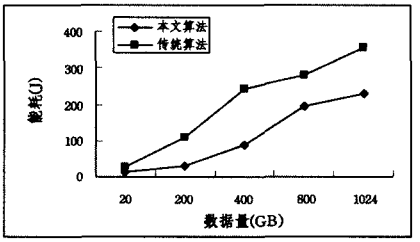

图 2　两种算法能耗比较结果

分析图 2 可以看出，当处理相同的数据量时，本文算法所需的能耗明显低于传统算法，这是因为本文算法易于实现，对数据的聚类效率高，所以消耗的能耗较少，验证了本文算法的有效性。

分析图 3 可以看出，当处理相同的数据量时，本文算法所需的时间明显低于传统算法，这是因为本文算法易于实现，对数据的聚类效率好，所以聚类时间短，进一步验证了本文算法的有效性。

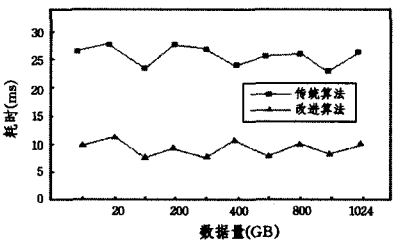

图 3　两种算法耗时比较结果

分析图 4 可以看出，随着数据量的逐渐增加，本文算法的关联维数趋于稳定，而传统算法的关联维数变化波动较大，说明本文算法有很高的数据聚类效率，验证了本文算法的有效性。

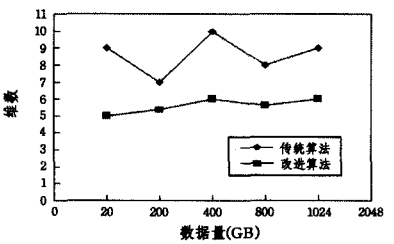

图 4　两种算法关联维数比较结果

结束语　本文提出一种基于混沌关联特征提取的大数据聚类算法，分析了传统方法的弊端，通过相空间重构建立一个多维的状态空间向量与混沌轨迹，使原系统中很多几何特征量保持不变，为分析原系统的混沌特征提供有效依据。将平均互信息量取第一个最小值时的横坐标所指的时间延迟作为重构相空间的最佳时间延迟，采用虚假最近邻点算法对最佳嵌入维数进行选择。将提取的关联维数这一特征量作为大数据聚类的混沌特征量，依据提取的混沌关联维特征对大数据进行聚类。仿真实验表明，所提算法能够有效提高数据的聚类效率，减少能耗，是一种有效的数据聚类方法。

参 考 文 献

[1] Yang Ling, Zheng Si-yi. Ship radiated noise feature extraction based on chaos theory [J]. Journal of Naval Engineering University, 2014(4): 50-54(in Chinese)
杨玲，郑思仪. 基于混沌理论的舰船辐射噪声特征提取[J]. 海军工程大学学报，2014(4)：50-54

[2] Fu Qiang, Li Chen-xi, Zhang Chao-xi. Chaotic correlation dimension algorithm for G-P discussion [J]. Journal of PLA University of Science and Technology (Natural Science Edition), 2014(3): 275-282(in Chinese)
付强，李晨溪，张朝曦. 关于 G-P 算法计算混沌关联维的讨论[J]. 解放军理工大学学报(自然科学版)，2014(3)：275-282

[3] Chang Yong-zhi, Qiu Ya-ze, Zheng Zhen, et al. Based on the nonlinear correlation dimension of feature extraction of mechanical automation monitoring system [J]. Computer and Digital Engineering, 2014(12): 2311-2315(in Chinese)
常勇智，邱亚泽，郑振，等. 基于非线性关联维特征提取的机械自动化监测系统[J]. 计算机与数字工程，2014(12)：2311-2315

[4] Xiao Fei, Qi Li-lei. Big data processing technology and exploration [J]. Computer and modernization, 2013(9): 75-77(in Chinese)
肖飞，齐立磊. 大数据处理技术与探索[J]. 计算机与现代化，2013(9)：75-77

[5] Wang Bin, Wang Chao, Li Jing. Big differences between the network abnormal data feature detection algorithm simulation analysis[J]. Computer Simulation, 2013, 30(8): 277-280(in Chinese)
王斌，王超，李晶. 大差异网络异常数据特征检测算法的仿真分析[J]. 计算机仿真，2013，30(8)：277-280

[6] Sun Hai-jun. Big data processing based on cloud computing technology [J]. Journal of Information Security and Technology, 2014(11): 61-63(in Chinese)
孙海军. 基于云计算的大数据处理技术[J]. 信息安全与技术，2014(11)：61-63

[7] Han Yan, Li Xiao. Speed up big data clustering K-means algorithm improvement [J]. Computer Engineering and Design, 2015, 36(5): 1317-1320(in Chinese)
韩岩，李晓. 加速大数据聚类 K-means 算法的改进[J]. 计算机工程与设计，2015，36(5)：1317-1320

[8] Yang Zhen, Xu Min-jie, Liu Zhang-feng, et al. Big data information processing architecture and key technology research [J]. Journal of Telecom Science, 2013, 29(11): 1-5(in Chinese)
杨震，徐敏捷，刘璋峰，等. 语音大数据信息处理架构及关键技术研究[J]. 电信科学，2013，29(11)：1-5

[9] Guan Tian-yun, Hou Chun-hua. Big data technology in the application of intelligent pipe huge amounts of data analysis and mining[J]. Journal of Modern Telecommunication Technology, 2014, 42(1): 71-79(in Chinese)
管天云，侯春华. 大数据技术在智能管道海量数据分析与挖掘中的应用[J]. 现代电信科技，2014，42(1)：71-79

[10] Sun Ting, Zhang Jin-hua, Geng Guo-hua. 3d model based on local features of probability density estimation feature extraction [J]. Computer Science, 2015, 42(6): 293-29(in Chinese)
孙挺，张锦华，耿国华. 基于局部特征概率密度估计的三维模型特征提取[J]. 计算机科学，2015，42(6)：293-295

[11] Zhong Ji-yuan, Mei Kui-zhi, Wen Zhe-xi. GIST feature extraction of heterogeneous concurrent flow computing implementation [J]. Computer Engineering and Applications, 2015, 51(6): 139-144(in Chinese)
仲济源，梅魁志，温哲西. GIST 特征提取的异构并发流计算实现[J]. 计算机工程与应用，2015，51(6)：139-144

[12] Li Kang, Liu Dong. Development Research on Data-Intensive Computing Towards Massive Data Processing[J]. Journal of Sichuan Ordnance, 2015, 36(7): 93-96(in Chinese)
李亢，刘东. 面向海量数据处理的数据密集型计算发展研究[J]. 四川兵工学报，2015，36(7)：93-96