DOI:10.19651/j.cnki.emt.1802103

基于语义关联特征的大型信息管理系统数据挖掘技术

张 稼 陆兴华
(广东工业大学华立学院 广州 511325)

摘 要:为了提高大型信息管理系统的数据检索和挖掘能力,提出了一种基于语义关联特征提取的大型信息管理系统数据挖掘技术。构建云存储模型进行大型信息管理系统中大数据分布式存储设计,结合大数据信息流的特征重组方法进行信息管理系统的优化结构重组,在重组的信息管理系统拓扑结构中提取信息管理分布数据的语义关联维特征量,以语义关联特征量为训练样本集进行信息管理系统的集成调度和数据挖掘,采用模糊 C 均值算法进行大型信息管理系统中分布数据语义关联特征的自适应融合和聚类处理,采用特征压缩器进行大型信息管理系统的存储空间降维处理,提高目标数据挖掘能力和信息管理系统的自适应调度能力。仿真结果表明,采用该方法进行大型信息管理系统数据挖掘的准确性较好,语义关联聚类性较强,提高了对信息管理系统目标数据的检索和调度能力。

关键词:语义;关联特征;信息管理系统;数据挖掘;信息检索
中图分类号:TP391;TN919 **文献标识码**:A **国家标准学科分类代码**:413.99

Data mining technology of large-scale information management system based on semantic association feature

Zhang Jia Lu Xinghua
(Huali College Guangdong University of Technology, Guangzhou 511325, China)

Abstract: In order to improve the ability of data retrieval and mining in large-scale information management system, a data mining technology of large scale information management system is proposed based on semantic association feature extraction. The cloud storage model is constructed to design the big data distributed storage in large information management system, and the optimized structure of the information management system is reorganized with the feature recombination method of big data information flow. The semantic association dimension feature quantity of the information management distribution data is extracted from the reorganized information management system topology, and the integrated scheduling and data mining of the information management system is carried out using the semantic association feature quantity as the training sample set. The fuzzy C-means algorithm is used for adaptive fusion and clustering of semantic association features of distributed data in large-scale information management system, and the feature compressor is used to reduce the dimension of storage space of large information management system. Improve the ability of target data mining and adaptive scheduling of information management system. The simulation results show that the method is accurate and semantic association clustering is strong, which improves the retrieval and scheduling ability of the target data in the information management system.

Keywords: semantics; association feature; information management system; data mining; information retrieval

0 引 言

随着大数据信息和大型云存储技术的发展,采用大型信息管理系统进行大数据存储和信息检索,能提高信息管理的效率,大型信息管理系统因此被广泛应用在图书馆信息管理、高校事务管理、能源信息管理以及电网信息管理等各个领域[1]。大型信息管理系统的构建是建立在数据库设计和信息检索算法基础上,结合信息特征提取和优化的信息调度技术,进行信息管理系统的总体设计构架和数据库开发,提高信息检索和集成信息处理能力,研究大型信息管理系统的数据挖掘技术,对提高信息管理系统的分布式检索和调度能力方面具有重要意义[2]。

在大型信息管理系统和云数据库介质中,需要通过有效的数据挖掘方法进行大型信息管理系统中分布数据信息

收稿日期:2018-09-18

访问和调度，提高数据的利用效率和资源共享程度[3]，传统方法中，对大型信息管理系统的语义特征挖掘是建立在大型信息管理系统中分布数据比特信息流的统计分析和特征提取基础上，典型的数据挖掘方法有关联规则挖掘方法、语义本体模型挖掘方法、频谱特征提取方法和模糊 K 均值挖掘方法等[4]，构造大型信息管理系统中的大数据特征分布的模糊聚类中心，结合谱特征提取和语义分析方法，进行数据挖掘，取得了一定的研究成果。其中，文献[5]提出一种基于模糊 C 均值聚类的大型信息管理系统中的目标管理数据挖掘方法，构建大数据的分类调度模型，结合自相关匹配方法进行数据挖掘，具有较高的数据挖掘精度，但该方法进行信息管理系统的数据库检索的准确性不好，数据挖掘的实时性不高。文献[6]提出一种基于关联规则信息融合的信息管理系统数据挖掘方法，结合自适应回归分析方法进行信息管理数据的关联特征提取，对提取的信息管理数据的关联规则特征量进行属性分类识别，但该方法进行信息管理系统挖掘的抗干扰能力不强，数据调度的准确度不高。

针对上述问题，本文提出一种基于语义关联特征提取的大型信息管理系统数据挖掘技术。首先构建云存储模型进行大型信息管理系统中大数据分布式存储设计，在重组的信息管理系统拓扑结构中提取信息管理分布数据的语义关联维特征量。然后采用模糊 C 均值算法进行大型信息管理系统中分布数据语义关联特征的自适应融合和聚类处理，采用特征压缩器进行大型信息管理系统的存储空间降维处理，提高目标数据挖掘能力和信息管理系统的自适应调度能力。最后进行仿真实验分析，展示了本文方法在提高大型信息管理系统的目标数据挖掘和检索能力方面的优越性能。

1　大型信息管理系统的数据存储结构及特征分析

1.1　数据存储结构模型

为了实现对大型信息管理系统中大数据分布式存储和优化挖掘，首先构建大型信息管理系统的数据存储结构模型，采用分布式数据库拓扑模型，构建大型信息管理系统的目标数据存储数据库，大型信息管理系统中的大数据的节点分布间隔为 d，采用相似性结构重组[7]，得到大型信息管理系统的数据库存储基本块为 $m_{i,j}(i\leqslant n,j\leqslant k)$，采用有向图 $G=(V,E,W,C)$ 表示大型信息管理系统中的大数据分布式存储的网格结构模型，在分块组合模型中，得到大型信息管理系统中的大数据存储的属性分布用子图 $G_1=(M_1^g,M_1^g,Y_1)$ 表示，结合闭频繁项集融合方法，得到大型信息管理的目标数据语义训练集满足 t_i，令 $G_1\subseteq G_2\subseteq Y_1\Leftrightarrow Y_2$，令 $A=\{a_1,a_2,\cdots,a_n\}$ 为数据结构的分布幅值，大型信息管理系统中目标数出现的概率用 $P(i,j)$ 表示，$p(i,j)=\lim_{t\to\infty}p\{a_t=i,b_t=j\}$，得到大数据采样的间隔单元 $i\in[0,n],j\in[0,n]$。采用自相关特征匹配方法，数据挖掘的时间窗口表示为：

$$Y_N=X_n+\eta \tag{1}$$

其中 Y_N 表示数据关联规则特征分量，而 η 表示为互信息量，将数据存储元 $\alpha(i,j)$ 输入到大型信息管理系统的存储链路层中，得到大型信息管理系统中的大数据时间序列为 x_n，信息调度的稳态特征量为 d_n，得到大型信息管理系统分布式存储的传递函数为：

$$\alpha(i,j)=\begin{cases}0,i=0orj=0\\1,n-j<i,i\geqslant j\\1,n-i<j,j\geqslant i\\1-{}_{n-j}C_i/{}_nC_i,n-j\geqslant i,i\geqslant j\\1-{}_{n-i}C_j/{}_nC_j,n-i\geqslant j,j\geqslant i\end{cases} \tag{2}$$

采用 Parallel Sets 本体结构模型进行大型信息管理系统中的大数据分区调度[8]，大型信息管理系统的关联规则分布模型如图 1 所示。

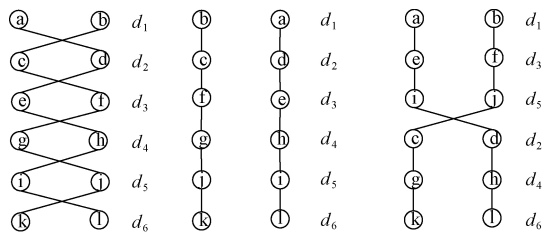

图 1　大型信息管理系统的关联规则分布模型

结合自适应分布数据重构方法进行型信息管理系统数据挖掘和语义相似性特征提取[9]，从而提高型信息管理系统的数据挖掘和信息调度性能。

1.2　语义关联维特征分析

在构建云存储模型进行大型信息管理系统中大数据分布式存储设计的基础上，结合大数据信息流的特征重组方法进行信息管理系统的优化结构重组，在重组的信息管理系统拓扑结构中提取信息管理分布数据的语义关联维特征量[10]，关联规则信息融合的判决统计量为：

$$Q_s=\frac{\langle(x_n-\bar{x})(x_{n-d}-\bar{x})(x_{n-D}-\bar{x})\rangle}{\langle(x_n-\bar{x})^3\rangle} \tag{3}$$

式中：x_n 表示大型信息管理系统中的大数据采样序列；d 表示数据采样时间间隔，$D=2d$；$\bar{x}$ 表示均值；$\langle x(n)\rangle$ 表示评价指标权重，为：

$$\langle x(n)\rangle=1/N\sum_{n=1}^{N}x(n) \tag{4}$$

结合二元语义权重分析方法，得到二元语义决策矩阵为：

$$\boldsymbol{R}=\begin{bmatrix}r(V_1,V_1) & \cdots & r(V_1,V_{k-1})\\ \vdots & \vdots & \vdots\\ r(V_{k-1},V_1) & \cdots & r(V_{k-1},V_{k-1})\end{bmatrix} \tag{5}$$

构造出评价决策矩阵，结合自适应信息融合方法，得到大型信息管理系统数据挖掘的判决准则为：

$$p(Q_s)=\frac{1}{\sqrt{2\pi}\sigma_s}\exp\left[-\frac{(Q_s-\langle Q_s\rangle)^2}{2{\sigma_s}^2}\right] \tag{6}$$

$$\int_{-\infty}^{\infty}p(Q_s)\mathrm{d}Q_s=1 \tag{7}$$

依据各评价对象与最优方案构建本体聚类中心,计算 Q_0 与 $\langle Q_s\rangle$ 的绝对误差,得到大型信息管理系统数据挖掘的判决阈值为 Q_c,当满足:

$$p(|Q_0-\langle Q_s\rangle|>Q_c)\leqslant 0.05 \tag{8}$$

在置信度为 95%的置信度水平下,得到语义关联特征挖掘的准确分布概率满足:

$$0.025=\int_{-\infty}^{z_2}p(Q_s)\mathrm{d}Q_s=1-\int_{-\infty}^{z_1}p(Q_s)\mathrm{d}Q_s \tag{9}$$

式中:$z_2=-z_1$,当 $S\geqslant 2.00$,表示信息管理系统的目标数据挖掘在语义本体模型中以 95%概率聚敛,得到语义关联特征提取结果为:

$$\begin{cases}\max U=u_1+u_2+\cdots+u_n\\ u_i=p_i\\ \sum\limits_i^n p_i=1,0<p_i<1\\ \dfrac{p_1/(1-p_1)}{w_1}=\dfrac{p_i/(1-p_i)}{w_i}=\cdots=\dfrac{p_n/(1-p_n)}{w_n}=\dfrac{1}{K}\end{cases} \tag{10}$$

其中,K 表示本体映射过程中概念之间语义相似度,根据上述特征提取结果,结合自适应回归分析方法进行信息管理数据关联规则特征挖掘。

2　数据挖掘算法优化

2.1　语义关联性特征提取

以语义关联特征量为训练样本集进行信息管理系统的集成调度和数据挖掘,采用模糊 C 均值算法进行大型信息管理系统中分布数据语义关联特征的自适应融合和聚类处理[11],分布数据分布样本属性集 $i\in S_s$,其等价的相似传递性映射关系满足:

$$\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{\alpha}=\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_jQ_{ij}\geqslant 0 \tag{11}$$

建立本体之间的语义映射关系,对于含有 n 个样本的目标数据,其中样本 $x_i,i=1,2,\cdots,n$ 的关联特征分布向量为:

$$s(t)=\sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty}a_{mn}g_{mn}(t)+n(t) \tag{12}$$

式中:a_{mn} 称为待大型信息管理系统中分布数据潜在有用信息的包络幅值;$g_{mn}(t)$ 为数据统计平均值;$n(t)$ 为干扰项,对于本体模型中的每个概念也可称为类。从语义上来说,当 $\boldsymbol{\Phi}:M\to R^{2d+1}$,表示 $\boldsymbol{\Phi}(z)=(h(z),h(\varphi_1(z)),\cdots,h(\varphi_{2d}(z)))^{\mathrm{T}}$ 存在一个嵌入式的高维相空间,得到极为相似的两组本体片段数据样本序列 $\{x(t_0+i\Delta t)\}$,$i=0,1,\cdots,N-1$,由此得到信息管理系统的本体之间语义映射关系:

$$x_n=a_0+\sum_{i=1}^{M_{AR}}a_ix_{n-i}+\sum_{j=0}^{M_{MA}}b_j\eta_{n-j} \tag{13}$$

式中:a_0 表示输入大型信息管理系统中分布数据信息流向量;a_i 是相空间嵌入维数;M_{AR} 是语义信息相关阶数;η_{n-j} 是数据采样时间间隔。在重组的信息管理系统拓扑结构中计算信息管理大数据样本的语义关联维特征量[12],结合数据聚类算法进行信息调度和关联规则挖掘。

2.2　信息管理系统目标数据挖掘

考虑两组结构相似的本体片段,构造大型信息管理系统中分布数据语义关联性特征融合的目标函数:

$$X_p(u)=\begin{cases}p\sqrt{\dfrac{1-\mathrm{j}\cot\alpha}{2\pi}}\,e^{\mathrm{j}\frac{u^2}{2}\cot\alpha}\displaystyle\int_{-\infty}^{+\infty}x(t)e^{\mathrm{j}\frac{t^2}{2}\cot\alpha-\mathrm{j}u\csc\alpha}\,\mathrm{d}t, & \alpha\neq n\pi\\ x(u), & \alpha=2n\pi\\ x(-u), & \alpha=(2n\pm1)\pi\end{cases} \tag{14}$$

式中:$x(t)$ 为海量大型信息管理系统中分布数据信息中包含挖掘目标属性集的本体片段;p 为测度距离,用欧氏距离表示为式(15)。

$$p=\|x_k-V_i\|^2 \tag{15}$$

根据语义映射 SM-Context,输出的信息管理数据目标样本集满足:

$$\sum_{i=1}^{c}\mu_{ik}=1,k=1,2,\cdots,n \tag{16}$$

结合概念的上下文特征寻优方法进行信息管理系统的目标数据调度,得到输出最优解:

$$F_j=\sum_{k=1}^{n}X_{kj},Q_j=\sum_{k=1}^{n}(X_{kj})^2 \tag{17}$$

明确概念所包含的语义成分,由此提取到大型信息管理系统中分布数据的语义关联维特征量[13],得到特征提取的迭代式:

$$x_i^{(k+1)}=(1-\omega)x_i^{(k)}+i(b_i-\sum_{j=1}^{i-1}a_{ij}x_j^{(k+1)}-\sum_{j=i+1}^{n}a_{ij}x_j^{(k)})\quad i=1,2,\cdots,n;k=1,2,\cdots,n \tag{18}$$

以语义关联特征量为训练样本集进行信息管理系统的集成调度和数据挖掘,以交集的逻辑形式进行数据分类挖掘,得到数据挖掘输出结果为:

$$a_{ii}x_i^{(k+1)}=(1-\omega)a_{ii}x_i^{(k)}+\omega(b_i-\sum_{j=1}^{i-1}a_{ij}x_j^{(k+1)}-\sum_{j=i+1}^{n}a_{ij}x_j^{(k)}) \tag{19}$$

式中:ω 为语义映射的初始权重;a_{ij} 为析取关系的联合特征量;$x_j^{(k)}$ 表示一个概念的语义与它周围的元素的差异值。对语义关联特征量输入到模糊 C 均值分类器中[14],采用特征压缩器进行大型信息管理系统的存储空间降维处理[15],提高目标数据挖掘能力和信息管理系统的自适应调度能力。

3　仿真实验与结果分析

为了测试本文方法在实现大型信息管理系统中分布数据优化挖掘和信息检索的应用性能,进行仿真实验,实验在

MATLAB 仿真软件中进行，实验的硬件环境配置为 Windows 10 系统的 PC，4 GHz 双核 Core i5 处理器，大型信息管理系统中分布数据库为 Micro-Clusters 2017 云存储数据库，信息管理系统中的分布数据包括 32 个概念集和 80 个属性集，含有 100 个语义信息实例集，语义映射的初始位置(0，0.25)，关联系数为 0.23，信息管理系统数据检索的惯性权重加速因子为 0.56，样本数据的初始采样频率为 $f_1=1.46$ Hz，终止采样频率 $f_2=2.12$ Hz，仿真时长为 200 ms，信息管理系统的管理数据包含 32 个概念、62 个属性和 100 个实例，语义关系所对应的映射如表 1 所示。

表 1　语义关系所对应的映射

语义关系	特征映射
$A \xrightarrow{\subseteq} B$	$w(A)\subseteq w(B)$
$A \xrightarrow{\supseteq} B$	$w(A)\supseteq w(B)$
$A \xrightarrow{\equiv} B$	$w(A)\equiv w(B)$

根据表 1 的语义本体关联映射，给出参考本体与前 10 个本体之间的语义映射结果如表 2 所示。

表 2　本体之间的语义映射结果

映射	＃102	＃103	＃104	＃201
等价映射	0	43	54	39
泛化映射	0	60	80	65
具体化映射	0	60	80	60

根据上述仿真环境和参数设定，进行信息管理系统的数据挖掘仿真分析，得到在大型信息管理系统中采集的原始数据样本的时域和频域波形如图 2 所示。

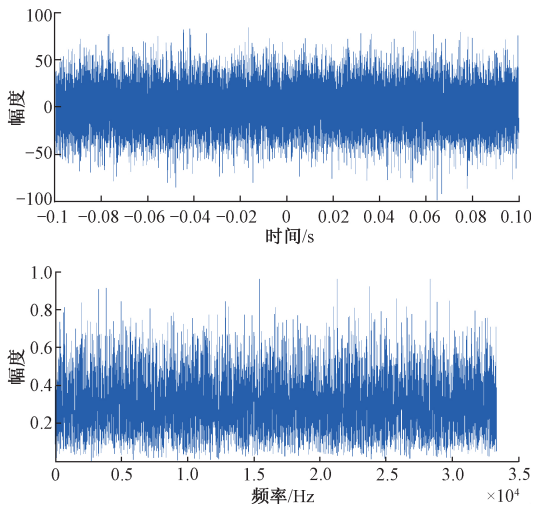

图 2　大型信息管理系统中数据采样的时域波形和频域波形

以图 2 的样本数据为测试对象集，进行语义关联特征提取和大数据挖掘，得到特征提取结果如图 3 所示。

分析图 3 可知，采用本文方法进行大型信息管理系统的数据挖掘，语义特征提取的抗干扰性较强，数据聚敛性较好。为了对比不同挖掘算法的性能，采用本文方法和传统方法在同一条件下进行大型信息管理系统的目标数据挖掘和检索，测试数据挖掘的查准率对比结果如表 3 所示，数据挖掘准确性对比结果如图 4 所示。

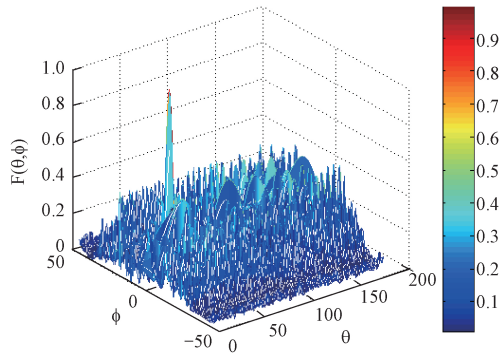

图 3　特征提取结果

表 3　查准率对比

迭代数	本文方法	粒子群优化算法	谱分析法
100	0.876	0.782	0.821
200	0.943	0.813	0.902
300	0.997	0.879	0.934
400	1	0.912	0.978

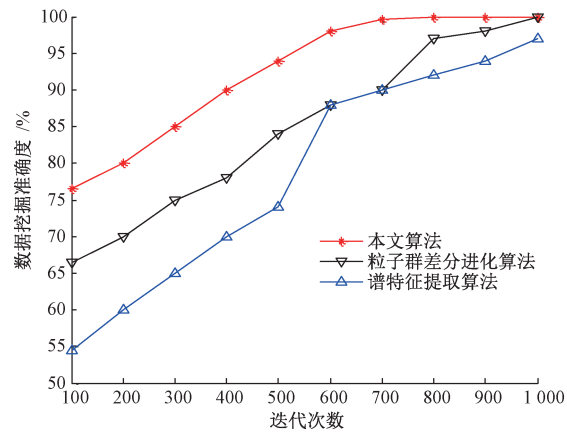

图 4　挖掘准确性能对比

分析表 3 和图 4 可知，随着迭代数的增大，查准率增大，采用本文方法的查准率比传统方法提升 12.46%，计算开销降低 23.76%，本文方法进行大型信息管理系统的存储数据挖掘和检索的准确性更高，性能优越于传统方法。但本文设计的数据挖掘方法进行信息管理系统大数据的批处理集成挖掘过程中的抗干扰能力不强，数据调度的准确度不高，下一步需要在这方面进行改进。

4 结　论

大型信息管理系统的构建是建立在数据库设计和信息检索算法基础上,结合信息特征提取和优化的信息调度技术,进行信息管理系统的总体设计构架和数据库开发,提高信息检索和集成信息处理能力,本文提出一种基于语义关联特征提取的大型信息管理系统数据挖掘技术。构建云存储模型进行大型信息管理系统中大数据分布式存储设计,在重组的信息管理系统拓扑结构中提取信息管理分布数据的语义关联维特征量,采用模糊C均值算法进行大型信息管理系统中分布数据语义关联特征的自适应融合和聚类处理,采用特征压缩器进行大型信息管理系统的存储空间降维处理,提高目标数据挖掘能力和信息管理系统的自适应调度能力。研究得知,本文方法进行大型信息管理系统数据挖掘的准确性较好,抗干扰能力较强。

参考文献

[1] 余晓东,雷英杰,岳韶华,等.基于粒子群优化的直觉模糊核聚类算法研究[J].通信学报,2015,36(5):78-84.

[2] 张博,郝杰,马刚,等. 混合概率典型相关性分析[J]. 计算机研究与发展,2015,52(7):1463-1476.

[3] 孙超,杨春曦,范莎,等. 能量高效的无线传感器网络分布式分簇一致性滤波算法[J]. 信息与控制,2015,44(3):379-384.

[4] 文天柱,许爱强,程恭. 基于改进ENN2聚类算法的多故障诊断方法[J]. 控制与决策,2015,30(6):1021-1026.

[5] KUMAR A, POOJA R, SINGH G K. Design and performance of closed form method for cosine modulated filter bank using different windows functions [J]. International Journal of Speech Technology, 2014, 17(4): 427-441.

[6] RAJAPAKSHA N, MADANAYAKE A, BRUTON L T. 2D space-time wave-digital multi-fan filter banks for signals consistingof multiple plane waves [J]. Multidimensional Systems and Signal Processing, 2014,25(1):17-39.

[7] 李智翔,李赟,贺亮. 采用新邻居模型的多目标分解进化算法[J].计算机工程与应用,2018,54(14):1-6.

[8] 范晓波,李兴明. 基于线性松弛方法的网络故障链路诊断[J]. 计算机应用,2018,38(7):2005-2008.

[9] 沈学利,覃淑娟. 基于SMOTE和深度信念网络的异常检测[J]. 计算机应用,2018,38(7):1941-1945.

[10] 张永,李卓然,刘小丹. 基于主动学习SMOTE的非均衡数据分类[J]. 计算机应用与软件,2012,29(3):91-93,162.

[11] 谷琼,袁磊,熊启军,等. 基于非均衡数据集的代价敏感学习算法比较研究[J]. 微电子学与计算机,2011,28(8):146-149.

[12] 李正欣,赵林度. 基于SMOTEBoost的非均衡数据集SVM分类器[J]. 系统工程,2008,26(5):116-119.

[13] 毛文涛,田杨阳,王金婉,等. 面向贯序不均衡分类的粒度极限学习机[J]. 控制与决策,2016,31(12):2147-2154.

[14] 庞俊,于戈,许嘉,等.基于MapReduce框架的海量数据相似性连接研究进展[J]. 计算机科学,2015,42(1):1-5.

[15] 袁泉,郭江帆. 新型含噪数据流集成分类的算法[J]. 计算机应用,2018,38(6):1591-1595.

作者简介

张稼,硕士、讲师,主要研究方向为信息管理与信息系统、电子商务、数据分析与预测。

E-mail:5914933@qq.com

陆兴华(通信作者),硕士、副教授,主要研究方向为嵌入式技术、无人机飞行稳定性控制方法、机器人运动控制方法。

E-mail:xhlu@gdut.edu.cn