doi:10.3969/j.issn.1673-4785.2010.02.005

命名实体的网络话题 K-means 动态检测方法

刘素芹[1],柴 松[1,2]
(1.中国石油大学 计算机与通信工程学院,山东 青岛 266555; 2.山东省军区 自动化工作站,山东 济南 250013)

摘 要:针对传统的网络话题检测方法在文本特征表示方面的不足及 *K*-means 聚类算法面临的问题,提出了一种基于命名实体的网络话题 *K*-means 动态检测方法.该方法对传统话题检测的特征表示方法进行了改进,用命名实体和文本特征词相结合表示文本特征,用命名实体对文本表示的贡献大小表示命名实体的权重;另外,利用自适应技术对 *K*-means 聚类算法中的 K 值进行自收敛,对 *K*-means 聚类算法进行了优化,利用 *K* 值的动态选取来实现网络话题的动态检测.实验结果表明,该方法较好地区分了相似话题,有效提高了话题检测的性能.

关键词:命名实体;网络话题;动态检测;K-means 聚类;自相似度;话题向量
中图分类号:TP18 **文献标识码**:A **文章编号**:1673-4785(2010)02-0122-05

K-means dynamic web topic detection method based on named entities

LIU Su-qin[1], CHAI Song[1,2]
(1. College of Computer & Communication Engineering, China University of Petroleum, Qingdao 266555, China; 2. Automation Workstation, Military District, Shandong Province, Ji'nan 250013, China)

Abstract: Current text representation models are not suitable for web topic detection, and the traditional K-means clustering algorithm has some drawbacks. The authors developed a dynamic K-means detection algorithm for web topics on the basis of named entities. In the new method, the representation model of the traditional topic detection method was modified. The text was represented by a combination of named entities and text features. The weight of the named entity was described by its contribution to the representation. The number of clusters *K* in the K-means algorithm self-converged by the use of an adaptive technique. The K-means algorithm was optimized, achieving a dynamic detection of web topics by using dynamic selection of *K* values. Experimental results indicated that the new method detects and distinguishes between similar topics effectively, thus significantly improving the performance of topic detection.

Keywords: named entity; web topics; dynamic detection; K-means clustering method; self-similarity; topic vector

网络话题检测与追踪[1](topic detection and tracking, TDT)旨在开发出一种能在没有人工干预的情况下自动判断新闻数据流话题的新技术[2].话题检测主要研究将新闻报道、新闻专线等来源的数据流中的报道归入不同的话题并在必要时建立新话题.相似话题的报道中有大量的相同词汇,容易造成话题误判,传统增量聚类方法很难解决这一问题[3],Kumaran 利用命名实体来解决此问题.详细分析可以得知,利用命名实体虽然能在一定程度上区分相似话题,但新闻报道中的命名实体的数目有限,仅仅依靠命名实体而放弃描述话题内容的大量其他关键词,必然造成对话题框架概括不全面,从而影响话题检测的性能[4].

本文将文本中的命名实体及除命名实体之外的特征词进行分别提取,并赋予不同的权重,将新闻文档表示成基于命名实体及特征词的双特征向量;然后在此基础上对 K-means 聚类方法[5]进行研究,结合自相似度策略来确定 *K* 值,解决了聚类算法中 *K* 值自收敛的问题,最终实现利用 *K* 值的动态选取来实现网络话题的动态检测.试验结果表明,与传统的话题检测方法相比较,该方法能够很好地解决海量网络数据环境下相似话题难以区分的问题,有效实现对网络话题的动态检测,该话题检测方法优于传

收稿日期:2009-12-04.
通信作者:刘素芹. E-mail:liusq@upc.edu.cn.

统话题检测方法.

1　基于命名实体的话题向量构造

基于命名实体的话题向量构造主要包括对网络文本中的命名实体提取以及对命名实体赋予一定权重,并将其与关键词合并构成网络话题向量.

1.1　命名实体提取

命名实体首次作为一个专门术语出现是在消息理解会议 MUC-6 上[6]. 根据消息理解会议的定义,命名实体分为七大类:人名、地名、机构名、日期、时间、百分数和货币. 这些短语都是文本中最基本的信息元素,往往指示了文章的主要内容,在对文本的理解上,命名实体的作用较之普通文本特征词来说是非常重要的. 因此,对文本中的命名实体进行专门提取,并予以与普通特征词不同的权重,将命名实体与普通特征词结合来表示文本,可以有效提高传统特征词向量对文本向量的表示,常用的命名实体主要包括人名、地名、机构名和时间实体.

在真实文本中,中文句子不是以词为单位的,而是以字为单位. 为了降低中文命名实体提取的复杂度,常常把分词信息用在中文命名实体提取中,但是分词的错误在命名实体提取过程中如果无法得到纠正,会导致错误蔓延. 为此,制定了 4 种规则,以此来修正中文分词导致的命名实体提取时的错误.

第 1 种规则为合并规则,主要修正长实体在分词时被作为几个连续的实体切分错误以及本属于支配关系的连续实体切分错误;第 2 种规则为同指人名规则,旨在找到指代同一人名的词,并统一进行提取;第 3 种规则为边界修正规则,主要用于实体切分时丢失了自身一部分的错误,此类错误主要为地名提取时,经常会发现丢失后缀的现象;第 4 种规则为类型修正规则,这种规则主要用于修正命名实体提取时的类型判断错误,地名提取时此类错误经常发生.

1.2　话题向量构造

利用以上这 4 种规则对中文词语切分之后的命名实体作具体的修正,然后提取新闻文档中的命名实体,并结合特征词分别赋予不同的权重,以此来完成话题向量的构造. 具体构造步骤如下:

1)预处理. 扫描文档,对文档进行分词,进行停用词去除处理. 经过预处理所得到的词表为文档初始词表.

2)词频统计. 对所提取的每一个特征词进行词频统计,将词所对应的词频作为该词所对应的权重;并根据词频统计结果对初始词表进行排序.

3)特征提取. 对排序后的词表采取截断处理,抽取特定数目的词作为该文档的特征词,形成最优词表.

4)拆分特征词表. 对特征词表进行扫描,加入修正规则,实现对其中命名实体的有效提取,将所提取的命名实体构造成命名实体词表,余下的特征词形成关键词表.

5)构造话题向量. 将所拆分后的词表分别处理,对命名实体词表和关键词表分别赋予不同的权重,最终形成话题向量. 权重的赋予根据经验值,将命名实体词表的权重加大,经过实验验证,命名实体词表与关键词表的权重比例在 3.5∶1 时效果最佳.

2　网络话题 K-means 聚类动态检测方法

在构造的网络话题向量的基础上,检测方法采用 K-means 聚类算法,来完成网络话题的动态检测. 本部分首先分析 K-means 聚类算法及其存在的问题,然后针对 K-means 聚类中 K 值的确定问题,引入了基于自相似度的最大最小原则[7-8],利用自相似度的自收敛策略来确定 K 值的选取,解决了 K-means 聚类话题检测中预先设定话题个数的问题,实现了话题的动态检测.

2.1　K-means 聚类算法

给定 d 维数据集 $X=\{x_i \mid x_i \in \mathbf{R}_d, i=1,2,\cdots,N\}$,将其聚成 K 个类别 $\omega_1,\omega_2,\cdots,\omega_K$,质心为 $c_1,c_2,\cdots,c_K$,其中 $c_i=(1/n_i)\sum_{x\in\omega_i}x$,$n_i$ 是类 ω_i 中数据点的个数. 聚类目标函数为:$J=\sum_{i=1}^{K}\sum_{j=1}^{n_i}d_{ij}(x_j,c_i)$,其中 $d_{ij}(x_j,c_j)$ 是 x_j 与 c_i 之间的欧氏距离.

K-means 聚类步骤如下:

1)从 X 中随机选择 K 个初始参照点 $c_1,c_2,\cdots,c_K$;

2)以 $c_1,c_2,\cdots,c_K$ 为参照点,对 X 进行划分. 满足:若 $d_{ij}(x_i,c_j)=\min\limits_{m=1,2,\cdots,K}d_{im}(x_i,c_m)$,其中,$j=1,2,\cdots,K$,$i=1,2,\cdots,N$, 则将 x_i 划分到类 ω_j 中;

3)根据式 $c_i=(1/n_i)\sum_{x\in\omega_i}x$,重新计算类的质心 $c_1^*,c_2^*,\cdots,c_K^*$;

4)若对于任意 $i\in\{1,2,\cdots,K\}$,$c_i^*=c_i$ 都成立,则算法结束,当前的 $c_1^*,c_2^*,\cdots,c_K^*$ 代表最终的聚类结果;否则,令 $c_i=c_i^*$,重新执行 2).

为了防止 4)中出现无限循环的情况,通常设置一个固定的阈值 th,当对于所有的 c_i,都有 $|c_i^*-c_i|<\text{th}$ 时,算法结束.

利用 K-means 聚类实现话题检测需要解决以下问题:

1)聚类类别数 K 的确定. 事先不知道所检测话题的个数,所以需要确定 K 的值.

2)初始质心的选择. K-means 是以质心为参照点进行聚类的,质心的选取决定最终所聚话题的核心内容,因此,如何选取聚类初始质心在动态话题检测中尤为重要.

2.2 自相似度收敛策略

假设文本中一个事件为一个文本单元,任意 2 个文本单元间的距离用余弦相似度来计算. 比如文本单元 i 的特征向量为 $\boldsymbol{x}_i=(a_1,a_2,\cdots,a_n)$,$j$ 的特征向量为 $\boldsymbol{x}_j=(b_1,b_2,\cdots,b_n)$,则 2 个文本单元的相似度:

$$\mathrm{sim}(\boldsymbol{x}_i,\boldsymbol{x}_j)=(\sum_{m=1}^{n}a_m b_m)/[\mathrm{sqrt}(\sum_{m=1}^{n}a_m^2)\times \mathrm{sqrt}(\sum_{m=1}^{n}b_m^2)]. \tag{1}$$

式中:a_m、$b_m(1\leqslant m\leqslant n)$ 为文本单元的第 m 个特征对应的权值,n 为 2 个文本单元的特征并集总数.

设聚类样本集合为:$X=\{x_1,x_2,\cdots,x_N\}$,N 为样本个数. 计算所有样本间的相似度 $\mathrm{sim}_{ij}=\mathrm{sim}(\boldsymbol{x}_i,\boldsymbol{x}_j)$,当 $i=j$,则 $\mathrm{sim}_{ij}=1$.

定义全局平均相似度:

$$\bar{s}=\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\mathrm{sim}_{ij}/N(N-1)/2. \tag{2}$$

定义最大最小平均相似度:

$$\bar{s}_{\min\max}=\{\max(\mathrm{sim}_{ij})+\min(\mathrm{sim}_{ij})\}/2. \tag{3}$$

式中:$i=1,2,\cdots,N-1$,$j=i+1,i+2,\cdots,N$.

定义平均自相似度,类别 ω_k 的平均自相似度:

$$\bar{s}_{kk}=\frac{s_{kk}}{n_k(n_k-1)/2},\ n_k=|\ \omega_k\ |. \tag{4}$$

式中:

$$s_{kk}=\sum_{i\in\omega_k}\sum_{\substack{j\in\omega_k\\ j\neq i}}\mathrm{sim}_{ij}. \tag{5}$$

i 与 j 同为类 ω_k 中的样本,因此称 s_{kk} 为类 ω_k 自相似度. 如果 $i\in\omega_{k_1}$ 且 $j\in\omega_{k_2}$,则 $s_{k_1k_2}=\sum_{i\in\omega_{k_1}}\sum_{j\in\omega_{k_2}}\mathrm{sim}_{ij}$ 为类 ω_{k_1} 和 ω_{k_2} 的互相似度.

定义全局平均自相似度门限:

$$\bar{s}_{G\mathrm{th}}=\max(\bar{s},\bar{s}_{\min\max}). \tag{6}$$

假设选取了 K 个质心,各类的平均自相似度分别为:$\bar{s}_{11},\bar{s}_{22},\cdots,\bar{s}_{KK}$,增加一个质心后,各类的平均自相似度为:$\bar{s}'_{11},\bar{s}'_{22},\cdots,\bar{s}'_{KK},\bar{s}_{(K+1)(K+1)}$.

定义局部自适应的平均自相似度门限 $\bar{s}_{\mathrm{th}}$:

$$\bar{s}_{\mathrm{th}}=(\bar{s}_{11}+\bar{s}_{22}+\cdots+\bar{s}_{kk}+\bar{s}'_{11}+\bar{s}'_{22}+\cdots+\bar{s}'_{KK})/2K. \tag{7}$$

该门限值随每次聚类动态变化.

如果出现

$$\frac{(\bar{s}'_{11}+\bar{s}'_{22}+\cdots+\bar{s}'_{KK})}{(|\ \bar{s}'_{11}-\bar{s}_{\mathrm{th}}\ |+|\ \bar{s}'_{22}-\bar{s}_{\mathrm{th}}\ |+\cdots+|\ \bar{s}'_{KK}-\bar{s}_{\mathrm{th}}\ |)}\geqslant \frac{(\bar{s}_{11}+\bar{s}_{22}+\cdots+\bar{s}_{KK})}{(|\ \bar{s}_{11}-\bar{s}_{\mathrm{th}}\ |+|\ \bar{s}_{22}-\bar{s}_{\mathrm{th}}\ |+\cdots+|\ \bar{s}_{KK}-\bar{s}_{\mathrm{th}}\ |)}$$

且

$$\bar{s}_{(K+1)(K+1)}\geqslant\bar{s}_{G\mathrm{th}}, \tag{8}$$

则继续选取下一个质心聚类.

3 基于命名实体的话题动态检测方法

基于命名实体的网络话题动态检测的主要思想是在文档特征提取上面进行突破,将文本中的命名实体和关键词进行分别处理,予以不同的权重,然后将二者结合构造话题向量,从话题的向量表示上加大了命名实体对文档表示的力度,丰富了词对文档表示的内容. 然后在 K-means 聚类方法中加入了基于最大最小的自相似度收敛策略,实现了 K-means 聚类方法中的 K 值的自动选取,从而实现了基于命名实体的话题动态检测. 新方法流程图如下:

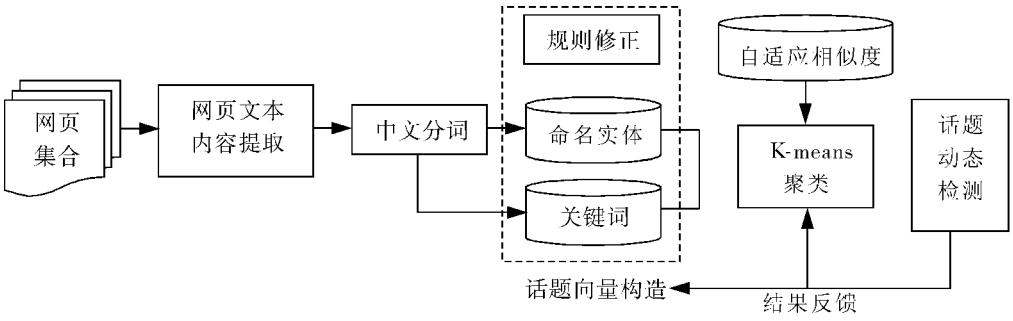


图 1 基于命名实体的网络话题动态检测流程图

Fig. 1 Flow chart of Web topic dynamic detection based on named entities

基于命名实体的话题动态检测的具体步骤如下:

1)网页文本内容提取,主要完成对半结构化的网页数据进行结构化处理,从新闻网页中抽取文本内容,去除包括网页链接、广告、版权信息等网页噪声,完成文本内容提取. 本文采用实验室已有的基于网页树的文本内容提取方法进行.

2)中文分词,主要完成对中文网页文本内容进行词语的自动切分. 本模块采用中科院开源 ICTCLAS 进行处理.

3)命名实体提取,主要在中文分词的基础上完成文本中命名实体的提取. 在本部分处理中采用本文第 1.1 节的规则修正策略,修正中文分词对命名实体的切分错误,从而实现文本中命名实体的完整而准确的提取.

4)话题向量构造. 中文分词之后对停用词进行有效去除,对其中的常用字等多频字也进行去除,并对所得词按照词频进行排序,截取预先设定好的词的个数,形成初始词表;从中提取的命名实体及其词频信息形成命名实体词表,余下的词形成关键词词表,并按照本文第 1.2 节的策略予以加权处理,最终构造话题向量.

5)自相似度 K-means 聚类. 首先选择初始聚类质心:将上述所构造的话题向量进行相似度计算,将计算结果进行排序,选择相似度最小的 2 个样本 s_1 与 s_2 作为初始质心,其他样本再按照与初始质心的相似度分别划分到 s_1 与 s_2 所在的类别中,然后分别计算 $\bar{s}_{11}$ 与 $\bar{s}_{22}$,此时聚类类别数 $K=2$. 如果出现具有同样最小相似度的其他样本对,则同样重新计算出以此对样本为质心后各类的平均自相似度 $\bar{s}'_{11}$、$\bar{s}'_{22}$. 具体方法二者选其一:

①计算不同样本对作质心后,各类之间的互相似度 $s_{k_1k_2}$ 和 $s'_{k_1k_2}$. 比较 $s_{k_1k_2}$ 与 $s'_{k_1k_2}$,取值小的那对样本作为初始质心.

②不计算互相似度,仅利用平均自相似度进行初始质心的选取.

比较 $(\bar{s}_{11}+\bar{s}_{22})/(|\bar{s}_{11}-\bar{s}_{Gth}|+|\bar{s}_{22}-\bar{s}_{Gth}|)$ 与 $(\bar{s}'_{11}+\bar{s}'_{22})/(|\bar{s}'_{11}-\bar{s}_{Gth}|+|\bar{s}'_{22}-\bar{s}_{Gth}|)$,选取比值较大的那对样本作为初始质心;如果比值相等,则选择 $|\bar{s}_{11}-\bar{s}_{22}|$ 与 $|\bar{s}'_{11}-\bar{s}'_{22}|$ 较小的那对样本.

这种样本对选择策略有 2 个好处:既可以尽量保证聚类后各类的平均自相似度不能太小,同时也避免了选取的样本聚类后 2 个类的平均自相似度相差太大的情况.

同样的方法选取下一个质心. 聚类后计算每个类别的平均自相似度,直到不满足式(8),停止聚类,并确定类别数 K.

6)结果反馈. 自相似度 K-menas 聚类的结果作为话题检测的输出,根据输出结果人工来调整话题向量构造及自相似度 K-menas 聚类算法中的参数,具体调整包括话题向量构造中的命名实体与关键词权重的比例大小及 K-means 聚类中参数的选择等.

4　实验结果及分析

实验数据来自 TDT2005 标准语料中的中文语料,包括27 142篇新闻报道. 评测标准同样采用 TDT 标准评测,主要包括漏报率 P_{miss} 和错报率 P_{FA} 以及将漏报率与错报率合并成一个检测开销 C_{Det} 及其规范式 $Norm(C_{Det})$[9]. 其计算公式为

$$C_{Det}=C_{miss}P_{miss}P_{target}+C_{FA}P_{FA}P_{non\text{-}target}. \quad (9)$$

式中:P_{miss} 为系统的漏报率,P_{FA} 为系统的错报率,C_{miss}、P_{target}、C_{FA}、$P_{non\text{-}target}$ 为事先设定值,具体如下:$C_{miss}=1.0$,$P_{target}=0.02$,$C_{FA}=0.1$;$P_{non\text{-}target}=1-P_{target}=0.98$.

为了使得到的性能指标落在更有意义的范围内,因此将 C_{Det} 规范化得到 $Norm(C_{Det})$:

$$Norm(C_{Det})=C_{Det}/\min(C_{miss}P_{target},C_{FA}P_{non\text{-}target}). \quad (10)$$

可以看出,$Norm(C_{Det})$ 值越小系统的性能越好. 本文实验结果均采用 Micro Average[10] 计算各项指标.

实验将本文话题检测结果与传统基于增量聚类的话题检测方法[3](增量聚类法)、基于命名实体话题检测方法[4](命名实体法)、基于 K-means 聚类话题检测方法(K-means 聚类法)话题检测结果相比较,对比结果如表 1 和图 2 所示.

表 1　不同方法话题检测结果对比

Table 1　Contrast of topic detection results with different methods　　%

检测项目	增量聚类法	命名实体法	K-means 聚类法	本文方法
漏报率	27.81	25.63	23.74	20.88
错报率	1.26	1.15	1.10	0.93
召回率	72.19	74.37	76.26	80.44
准确率	72.97	75.17	75.99	81.02
F1-Measure	74.33	75.17	77.56	80.73
Norm(CDet)	30.82	29.04	27.83	23.86

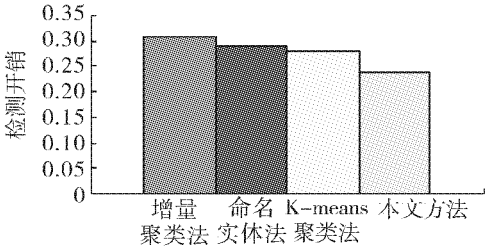


图 2　4 种方法的检测开销对比

Fig. 2　Comparison of four methods of detection of overhead

从表 1 和图 2 的检测结果对比中可以看出,基于命名实体的网络话题动态检测方法比文献[3]中

基于命名实体的话题检测方法检测准确率提高了6个百分点,召回率提高了6个百分点,主要原因在于文献[3]在话题向量表示中仅采用命名实体,忽略了命名实体之外的关键词对文本表达的作用,而新方法将命名实体和关键词结合起来构造话题向量,并且根据实际情况,针对命名实体与关键词对文本的贡献程度不同分别赋予不同的权重,使得在文档向量表示上对话题描述更充分、全面而准确,因此可以很好地区分相似话题. 新方法与基于增量聚类及K-means聚类话题检测方法相比较,无论是从检测准确率、召回率都有一定提高,并且本文在 K 值选取上实现了动态自收敛. 与传统方法相比较,本文方法无论是从检测性能,还是检测开销上都优于传统方法,是一种高性能且实用的网络话题检测方法.

5 结束语

本文在传统网络话题检测的基础上做了改进:一是在话题特征表示上,将命名实体及关键词进行分别处理,赋予不同的权重来构造话题向量,丰富了词对文档的表达,使得机器处理更能贴近人的理解;二是利用自相似度对K-means聚类中的 K 值进行自收敛,解决了K-means聚类中的问题,从而实现了利用K-means聚类对网络话题的动态检测.

参考文献:

[1] ALLAN J, CARBONELL J, DODDINGTON G. Topic detection and tracking pilot study: final report[C]// Proceeding of the DARPA Broadcast News Transcription and Understanding Workshop. San Francisco, 1998:194-218.

[2] 洪 宇, 张 宇, 刘 挺. 话题检测与跟踪的评测及研究综述[J]. 中文信息学报, 2007, 21(6):71-87.
HONG Yu, ZHANG Yu, LIU Ting. Topic detection and tracking review[J]. Journal Chinese Information Processing, 2007, 21(6):71-87.

[3] YAMRON J P, KNECHT S, Van MULBREGT P. Dragon's tracking and detection systems for the TDT2000 evaluation [C]//Proceedings of Topic Detection and Tracking Workshop. Washington, USA, 2000:75-80.

[4] KUMARAN G, ALLAN J. Text classification and named entities for new event detection[C]//Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Sheffield, 2004:297-304.

[5] YIU M C. K-means: a new generalized K-means clustering algorithm[J]. Pattern Recognition Letters, 2003(24): 2883-2893.

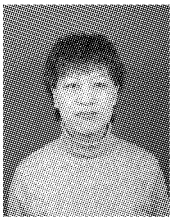
[6] SUNDHEIM B M. Named entity task definition[C]//Proc of the Sixth Message Understanding Conf. Columbia, Maryland, 1995:319-332.

[7] DING C, HE Xiaofeng. Cluster merging and splitting in hierarchical clustering algorithms[C]//Proceedings of the 2002 IEEE International Conference on Data Mining. Maebashi City, Japan, 2002: 139-146.

[8] DING C, HE X, ZHA H, et al. A min-max cut algorithm for graph partitioning and data clustering[C]//Proceedings of the IEEE Internationl Conference. San Jose, California, USA, 2001: 107-114.

[9] 骆卫华, 于满泉. 基于多策略优化的分治多层聚类算法的话题发现研究[J]. 中文信息学报, 2006, 20(1):29-36.
LUO Weihua, YU Manquan. The study of topic detection based on algorithm of division and multi-level clustering with multi-strategy[J]. Journal Chinese Information Processing, 2006, 20(1):29-36.

作者简介:



刘素芹,女,1968年生,副教授,博士,主要研究方向为计算机网络、高性能计算,近3年发表学术论文20余篇,编写教材2部.

柴 松,男,1981年生,主要研究方向为计算机网络、高性能计算及应用.