# 一种面向网络话题发现的增量文本聚类算法 *

殷风景，肖卫东，葛 斌，李芳芳
（国防科学技术大学 C4ISR 技术国防科技重点实验室，长沙 410073）

**摘 要**：为满足网络舆情监控系统中话题发现的需要，并克服经典 single-pass 算法处理网络文本聚类中受输入顺序影响和精度较低的主要不足，提出了 ICIT 算法，继承了 single-pass 算法的简单原理，保证了网络文本聚类的实时性；通过正文分词时标注词性选择名词动词进行正文向量化、建立文本标题向量来与文本正文向量共同表征文本、采用 average-link 策略、引入“代”的概念分批进行文本的聚类，以及在每批次聚类后添加报道重新选择调整所属的步骤来提高聚类的质量。实验证明了 ICIT 算法在提高话题发现准确度上的有效性和实用性。

**关键词**：话题发现；文本聚类；增量聚类；准确度；ICIT 算法
**中图分类号**：TP391 **文献标志码**：A **文章编号**：1001-3695(2011)01-0054-04
doi:10.3969/j.issn.1001-3695.2011.01.013

# Incremental algorithm for clustering texts in internet-oriented topic detection

YIN Feng-jing, XIAO Wei-dong, GE Bin, LI Fang-fang
(*C4ISR Technology National Defense Science & Technology Key Lab, National University of Defense Technology, Changsha* 410073, *China*)

**Abstract**: To meet the needs of topic detection for monitoring the public opinion on internet, this paper proposed an incremental clustering algorithm called ICIT to improve the two main disadvantages of single-pass algorithm, that was, being easily effected by the order of inputs and low precision. ICIT inherited the simple principle from single-pass to ensure clustering internet texts in real time and overcame its shortage by selecting only nouns and verbs from content as the content's vector expression, using vector expression of title with content's vector expression to express the text better, adopting average-link comparison strategy, introducing generation to accomplish batch process and add a stage for texts to reconsideration and adjust their ascription after first clustering. Experiments approve ICIT's validity and practicability in heightening the precise of topic detection.

**Key words**: topic detection; text clustering; incremental clustering; precise; ICIT algorithm

## 0 引言

互联网已经被确立为继报纸、广播、电视之后的第四代媒体，中国互联网络信息中心（CNNIC）发布的《第 25 次中国互联网络发展状况统计报告》显示，截止 2009 年 12 月 31 日，我国网民人数达到 3.84 亿，手机上网用户达 2.33 亿。此外，数据显示，中国 56% 的网民经常在网上发表意见，84.3% 的网民认为互联网是最重要的信息渠道，48% 的网民对互联网的信任程度比电视高。网络的广泛普及、网络信息的爆炸增长和网民参与热情的空前高涨使得对网络内容的监控和管理变得十分迫切和紧要，而面对海量的互联网信息，人工监管是不可想象的，于是网络舆情监控系统应运而生。

网络舆情是通过互联网传播的公众对现实生活中某些问题所持有的较强影响力、倾向性的观点和言论，是网民关注的热点，是民众讨论的焦点，主要通过新闻评论、BBS 论坛、博客、聚合新闻（RSS）、转贴等实现并加以强化，集中反映一个时期网络舆论的中心[1]。网络舆情监控系统是针对网络舆情进行信息采集、话题发现、热度评估、跟踪预警和分析处理的信息系统，话题发现是其重要的内容。话题发现依靠聚类的方法，将报道聚合成若干簇，簇内的报道之间相似度高，簇间的报道相似度低，以此来整合网络上大量的重复信息和同一话题内的不同信息。使用 TDT 相关技术后，人们以话题为粒度发现新事件并了解事件的发展，在应对信息爆炸和新知识发现上有较重要的意义[2]。同时，网络信息的快速更新和数量巨大，要求话题发现所采取的聚类方法能够针对每次更新的新增报道进行聚类而不用将新增报道与已有报道全部重新聚类，因为那样的计算代价巨大，浪费明显，这就是增量聚类。

CMU[2] 使用经典 single-pass 算法进行新事件的探测，该方法虽然计算简单、运算速度快，但其检测性能过分依赖新闻语料的到达顺序。雷震等人[3] 提出一种改进 K 均值算法用于热点话题发现，该算法使用密度函数法进行聚类中心的初始化，以便客观地选择初始聚类中心。该算法既可以用于在线探测，也可以用于回溯探测，且执行结果受新闻语料处理顺序的影响较小，但是当新报道到来时，该算法需要针对所有样本重新计算，无法保证话题发现的实时性。赵华等人[4] 针对新闻报道的特点，通过内容分析法将话题表示为标志中心向量和内容中心向量，解决包含相同核心内容的两个话题难以区分的问题，但是内容分析法需要额外词典的支持，增加了大量的人工维护工作。曾依灵等人[5] 以切分词为基础进行多级过滤的拼接，

**收稿日期**：2010-06-17；**修回日期**：2010-07-23 **基金项目**：国家自然科学基金资助项目（60903225）

**作者简介**：殷风景（1985-），男，硕士研究生，主要研究方向为指挥信息系统、信息资源管理（yinfengjingshiwo@163.com）；肖卫东（1968-），男，教授，博导，博士，主要研究方向为信息管理、智能决策技术；葛斌（1980-），男，博士，主要研究方向为对等网、信息集成、知识管理；李芳芳（1983-），女，博士，主要研究方向为信息集成、知识管理.

提取出能够代表网络热点话题的信息串,这种方法的问题在于切分词以及多级过滤的效率难以保证,难以满足在线、实时话题发现的应用。周亚东等人[6]在分析流量内容中热点词语与热点话题关系的基础上,采用基于高密度连接区域的密度聚类方法,形成热点词语簇,得出网络热点话题的属性描述,该方法是基于 DBSCAN 算法思想,与基于 K 均值的算法存在相同问题,即无法满足热点话题发现的实时性。王巍等人[7]提出了基于多中心模型的网络热点话题发现算法,将报道内容之间的关联关系层次化,提高了对网络话题的描述能力,有效消除了网络话题相关报道内容侧重点变化对网络话题发现准确性的影响。

话题发现与跟踪(topic detection and tracking,TDT)的评测中常用的聚类方法是 single-pass 聚类,其原理简单、计算速度快,然而该算法的缺点也很明显:受输入顺序的影响,且聚类结果精度差[8],离实际应用还有一定距离。围绕着该算法的主要不足,大量的改进算法不断出现。本文所采用的聚类算法 ICIT(incremental algorithm for clustering internet texts)是以某省网络舆情监控系统为应用背景,借鉴 single-pass 聚类方法简单高效的原理,又重点克服其缺点,经多处改进而成,兼顾了网络话题发现的实时性和准确性要求,并在应用中证实了其可用性和有效性。

## 1　话题发现

### 1.1　话题发现的流程

话题发现是为网络舆情监控服务的,它的任务是把大量讨论同一事件同一话题的新闻报道聚合在同一个簇下,以减少重复和冗余;同时,可以以话题为粒度考察其受关注的程度即热度,更好地评估某个话题的潜在影响力,为舆情的监管和预警提供依据。

按照处理过程,话题发现可以分为以下几个步骤:敏感信息采集、网络文本处理、文本内容分词、文本向量化、网络文本聚类和话题热度评估。敏感信息采集是把具有一定敏感性或潜在敏感性的互联网原始信息(如贴文或负面新闻)利用爬虫工具抓取下来;网络文本处理即把原始互联网信息进行包含剔除无关字符、提取正文和必要信息(如来源、作者、发表时间等)等操作的清洗;文本内容分词是在词典的指导下,将内容分解为一个个可以被计算机识别的汉语词语,以便后序的相应操作;文本向量化依据分词后的词频统计值将文本表示为一组关键词及以其词频为权重的空间向量;网络文本聚类是计算文本向量之间的距离或者相似程度,以确定两个文本是否同属一个话题;话题热度评估是综合考虑话题中报道的点击数、回帖数、报道频率和时间频率等参数,来评估该话题的受关注程度。其中采集和处理为聚类提供数据来源,是基础;分词和向量化是必要的转换以使后面的步骤得以进行;聚类是话题发现的核心,聚类算法的效率和精度关系到话题发现的有效性;话题热度评估不是话题发现的必然组成部分,但是从网络舆情监控的角度出发,它显然有益于话题发现的数值表达和定量把握。这个过程如图 1 所示。

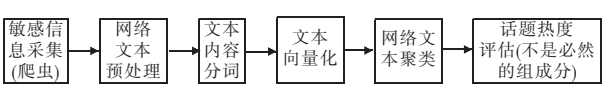


图 1　话题发现流程

### 1.2　Single-pass 聚类算法介绍

该算法依次处理输入的互联网文本(在这之前已经经过向量化处理),一次一篇,以增量的方式进行动态聚类,将文本向量与已有话题内的报道进行比较,根据相似度度量进行匹配。如果与某个话题模型匹配,则把文本归入该话题,如果该文本和所有话题模型的相似度度量均小于某一阈值,则将该文本表示为一个新的话题种子,创建新话题[9]。不同的阈值设置可以得到不同粒度大小的话题。

Single-pass 算法聚类过程如下:

a)接收一篇互联网文本向量 $d$;

b)$d$ 逐一与已有的话题中各报道进行相似度计算,并取最大者作为与该话题的相似度(single-link 策略);

c)在所有话题间选出与 $d$ 相似度最大的一个,以及此时的相似度值;

d)如果相似度大于阈值 $TC$,$d$ 所对应的互联网文本被分配给这个话题模型文本类,跳转至 f);

e)如果相似度小于阈值 $TC$,$d$ 所对应的文本不属于已有的话题,创建新话题,同时把这篇文本归属创建的新话题模型文本类;

f)本次聚类结束,等待新文本到来。

该算法聚类过程的流程如图 2 所示。

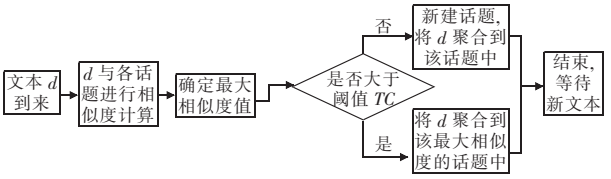


图 2　single-pass 聚类算法执行流程

## 2　ICIT 聚类算法

### 2.1　算法的主要思想

K 均值聚类方法需要先输入簇的数目,而面对互联网的实时文本流,这是根本无法实现的,并且该算法难以实现增量聚类;基于密度的算法如 DBSCAN 等,用于回溯探测还可以,用于实时的网络话题时,在时间复杂度上很难满足应用的要求。于是 single-pass 算法成了主要的选择,但要满足实际应用的提高,该算法也还有许多地方需要改进,主要为提高其聚类精度和克服受输入顺序影响两方面。

在文本向量化的过程中,通常是将文本的正文内容分词后进行词频和反文档频率统计,再依据 TFIDF 公式计算出各个特征词的权重,用文本特征词的权重组合为一个空间向量来数值化文本,进行相似计算。TFIDF 公式如下:

$$W_i=\frac{TF_i(t,d)\log\left(\frac{N}{DF(t)}+0.01\right)}{\sqrt{\sum_k TF_i^2(t,d)\ \log^2\left(\frac{N}{DF(t)}+0.01\right)}}$$

其中:$W_i$ 表示第 $i$ 个特征词的权重;$TF_i(t,d)$ 表示词 $t$ 在文档 $d$ 中的出现频率;$N$ 表示文档总数;$DF(t)$ 表示包含词 $t$ 的文档数目。

然而在实际应用中,笔者发现这样做有两个很大的缺点:a)大量文本正文提取出的大量关键词构成了海量的特征词集合,使得文本向量化时得出的向量维度很高(通常在几万维以上),这不仅造成了很大的计算开销,影响了算法的效率,更因

文本向量矩阵的稀疏造成各文本间相似度区分不明显,不利于相似度的比较和话题的区分;b)对于互联网文本(如新闻)而言,主要信息都会在标题中体现,换言之,标题中词语对文本的区分度要远大于正文中词语的区分度。

ICIT 算法的改进为:对正文的分词添加词性标注,词频统计仅针对具有实际意义的名词和动词,这既大大减少了特征词的数目,又最大限度地保留了关键信息;对标题也进行相应的分词和统计,并单独构成一个标题向量。当两篇文本需要进行相似度计算时,分别计算它们的标题向量和正文向量的相似度,然后予以加权求和作为这两篇文本最终的相似度。ICIT 算法采用夹角余弦公式计算两向量的相似度,为标题相似度和正文相似度赋予的权重分别是 0.7 和 0.3。夹角余弦公式为

$$\mathrm{sim}(D_i,D_j)=\frac{\sum\limits_{k\in N,i\neq j}D_{ik}D_{jk}}{\sqrt{\left(\sum\limits_{k\in N}D_{ik}^2\right)\left(\sum\limits_{k\in N}D_{jk}^2\right)}}$$

Single-pass 算法依照文本到达的先后顺序依次处理,在第一次读取文本时就可以确定它的所属类簇,这是该算法的优势,但同时也给它带来了一定的不利:文本读入的先后会对结果有较大的影响。而从理论上讲,数据源和各参数确定时,聚类结果应该是确定的,不应该因顺序不同而有所不同。ICIT 针对这一点所作的改进就是为了减少这种因顺序给聚类结果带来的扰动。ICIT 提出一个“代”的概念,指的是到达的文本不再一个个地进行聚类,而是一批批地添入到聚类过程中,这个“批”就是“代”。每一代的数目是个固定且可调节的参数(如本文 ICIT 中取为 200),开始与已有话题类比较相似度之前,先在本代成员之间进行初步的相似比较和聚类,再将这些初步类与已有话题类进行比较和聚合。

相似比较时虽然 single-link 策略足够简单,但面对实际数据时,average-link 策略的准确度更佳,且能有效减少大类的出现[10],所以 ICIT 采用 average-link 策略。

在聚类过程中新加入的文本只经过了与话题中已有文本的相似比较,而未考虑后面加入的文本对该文本当初选择正确性的冲击和影响。事实上,后续文本的加入是可能引起该文本选择变化的。所以,ICIT 给所有(当前代)文本一个重选择的机会,在当前代内成员完成聚类后加入一个比较调整的步骤,代内成员依次计算当前聚类结果下最相似的类簇是否就是自己所处的类簇,不是则调整。因为这种调整所引起的变动是连锁和动态的,所以本文给出调整终止的条件:90% 的文本已不用调整。实验证明该终止规则是有效的,结果也是比较让人满意的。

### 2.2　算法运行描述

如图 3 所示,经过了正文词性选择向量化和标题向量化的文本,以代为单位进入聚类算法模块,先在代内进行初步的自聚,再将初步类与已有话题作比较,具体比较方法为初步类依次与各话题类中的文本进行相似计算,再将与话题内所有文本相似度的均值作为与该话题的相似度(即 average-link 策略),与预设的相似阈值比较后确定是创建新话题类还是归入已有话题类,之后进入重选择调整阶段。在该步骤中,代内每一个文本重新计算与当前所有话题类的相似度(同样采用 average-link 策略),看与它相似度最大的一个话题类是否就是它所在的类簇;如果满足调整终止条件,则结束当前一轮聚类,等待下一代文本的进入,否则将文本调整到最大相似度的类簇中进行迭代计算。

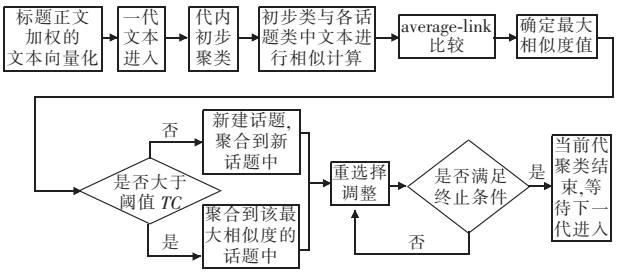


图 3　ICIT 算法执行描述

### 2.3　算法的优化分析

从本文以上内容可以看出,ICIT 算法是基于经典的 single-pass 算法的,这保证了话题发现的实时性;同时,ICIT 算法通过设“代”和代内初聚,有效地降低了文本顺序对聚类结果的影响,词性选择的正文向量化极大地降低了向量的维数和计算的复杂度,正文和标题双向量的机制提高了聚类结果的精确度,average-link 策略的采用有效提高了话题发现的准确性并减少了大类现象,算法最后的重选择也起到了提高结果精度的作用。因而,ICIT 算法具有很高的有效性和实用性。

## 3　实验及结果分析

### 3.1　评价指标

通常,人们使用错检率和漏检率以及耗费函数来评价话题发现的质量。假设与某话题相关的互联网文本数是 $A$,不相关的互联网文本数是 $B$;其中 $A$ 里面有 $A_1$ 篇被聚到该话题类中,$B$ 中有 $B_1$ 篇被聚到该话题类。那么,错检率就是 $F=B_1/B$,漏检率就是 $E=(A-A_1)/A$。耗费函数综合考量了漏检和错检的代价,公式为

$$C=C_EP_EP+C_FP_F(1-P)$$

其中:$C_E$、$C_F$ 是漏检和错检的代价,$P_E$、$P_F$ 是漏检和错检的概率,$P$ 是文本属于某话题类的先验概率。根据 TDT 评测的标准,设定 $C_E=1$,$C_F=0.1$,$P=0.02$。

### 3.2　实验环境与实验数据

实验使用 AMD OPTERON 2G 曙光服务器进行实验,操作系统是 Linux 企业版。采用中国科学院 ICTCLAS 分词系统,测试数据是利用网络爬虫工具采集的互联网新闻报道,再人工挑选其中 10 个话题的新闻,分别是南平惨案、西南干旱、波兰总统坠机、索马里海盗、泰国时局、富士康跳楼、天安舰事件、房产新政、世博会、赵作海案。为维持一定的数据量,为每一个话题选择 100 篇相关的新闻报道,共 1 000 篇。

### 3.3　结果分析

实验中设定 ICIT 算法和经典 single-pass 算法的阈值为 0.25。话题发现的结果和有关统计如表 1、2 所示。

实验采用漏检代价 $C_E=1$,错检代价 $C_F=0.1$,先验概率 $P=0.02$,得出 single-pass 算法的漏检率、错检率和耗费函数分别为 0.217、0.004 62、0.004 792 76,ICIT 算法的漏检率、错检率和耗费函数为 0.119、0.000 66、0.002 444 68。可见,ICIT 算法中的几项改进对话题发现起到了很好的作用,提高了聚类的质量。聚类结果中出现的“其他”是没有聚合到前面十个话题中的报道的集合,这种现象是由 VSM 模型和基于欧式距离相似计算的固有弊端造成的,难以根除。

表 1　single-pass 算法运行结果统计

| 初始话题类别(10) | 南平惨案 | 西南干旱 | 波兰总统坠机 | 索马里海盗 | 泰国时局 | 富士康跳楼 | 天安舰事件 | 房产新政 | 世博会 | 赵作海案 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 话题报道数目 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | |
| single-pass 聚类后话题类别 | 南平惨案 | 西南干旱 | 波兰总统坠机 | 索马里海盗 | 泰国时局 | 富士康跳楼 | 天安舰事件 | 房产新政 | 世博会 | 赵作海案 | 其他 |
| single-pass 聚类后话题报道数 | 78 | 83 | 76 | 79 | 82 | 85 | 87 | 81 | 90 | 84 | 175 |
| $A_1$ | 75 | 79 | 70 | 75 | 78 | 80 | 82 | 75 | 89 | 80 | |
| $B_1$ | 3 | 4 | 6 | 4 | 4 | 5 | 5 | 6 | 1 | 4 | |
| $A-A_1$ | 25 | 21 | 30 | 25 | 22 | 20 | 18 | 25 | 11 | 20 | |
| $B$ | 900 | 900 | 900 | 900 | 900 | 900 | 900 | 900 | 900 | 900 | |
| 单个话题漏检率 $E=A-A_1/A$ | 0.25 | 0.21 | 0.30 | 0.25 | 0.22 | 0.20 | 0.18 | 0.25 | 0.11 | 0.20 | |
| 单个话题错检率 $F=B_1/B$ | 0.0033 | 0.0043 | 0.0066 | 0.0044 | 0.0044 | 0.0055 | 0.0055 | 0.0066 | 0.0011 | 0.0044 | |
| single-pass 平均漏检率 | 0.217 | | | | | | | | | | |
| single-pass 平均错检率 | 0.00462 | | | | | | | | | | |
| single-pass 耗费函数 | 0.00479276 | | | | | | | | | | |

表 2　ICIT 算法执行结果统计

| 初始话题类别(10) | 南平惨案 | 西南干旱 | 波兰总统坠机 | 索马里海盗 | 泰国时局 | 富士康跳楼 | 天安舰事件 | 房产新政 | 世博会 | 赵作海案 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 话题报道数目 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | |
| ICIT 聚类后话题类别 | 南平惨案 | 西南干旱 | 波兰总统坠机 | 索马里海盗 | 泰国时局 | 富士康跳楼 | 天安舰事件 | 房产新政 | 世博会 | 赵作海案 | 其他 |
| ICIT 聚类后话题报道数 | 85 | 91 | 78 | 95 | 89 | 92 | 88 | 76 | 99 | 94 | 113 |
| $A_1$ | 84 | 90 | 76 | 95 | 88 | 92 | 88 | 75 | 99 | 94 | |
| $B_1$ | 1 | 1 | 2 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | |
| $A-A_1$ | 16 | 10 | 24 | 5 | 12 | 8 | 12 | 25 | 1 | 6 | |
| $B$ | 900 | 900 | 900 | 900 | 900 | 900 | 900 | 900 | 900 | 900 | |
| 单个话题漏检率 $E=A-A_1/A$ | 0.16 | 0.11 | 0.24 | 0.05 | 0.12 | 0.08 | 0.12 | 0.25 | 0.01 | 0.06 | |
| 单个话题错检率 $F=B_1/B$ | 0.0011 | 0.0011 | 0.0022 | 0 | 0.0011 | 0 | 0 | 0.0011 | 0 | 0 | |
| ICIT 平均漏检率 | 0.119 | | | | | | | | | | |
| ICIT 平均错检率 | 0.00066 | | | | | | | | | | |
| ICIT 耗费函数 | 0.00244468 | | | | | | | | | | |

## 4　结束语

本文面向网络话题发现的具体应用,基于经典 single-pass 算法,并针对其主要不足进行改进,提出了适用于网络文本增量聚类的 ICIT 算法,后面的实验证明所提算法是有效的,提高了聚类和话题发现的质量,具有很高的使用价值。值得一提的是,ICIT 算法中对正文选择词性进行向量化的方法客观上降低了算法的运算复杂度,使得算法的执行性能明显优于经典 single-pass 算法。

**参考文献:**


[1] 徐晓日. 网络舆情事件的应急处理研究[J]. 华北电力大学学报:社会科学版,2007(1):89-93.

[2] 张晓艳,王挺. 话题发现与追踪技术研究[J]. 计算机科学与探索,2009,3(4):347-357.

[3] 雷震,吴玲达,雷蕾,等. 初始化类中心的增量 K 均值法及其在新闻事件探测中的应用[J]. 情报学报,2006,25(3):289-295.

[4] 赵华,赵铁军,张妹,等. 基于内容分析的话题检测研究[J]. 哈尔滨工业大学学报,2006,38(10):1740-1743.

[5] 曾依灵,许洪波. 网络热点信息发现研究[J]. 通信学报,2007,28(12):141-146.

[6] 周亚东,孙钦东,管晓宏,等. 流量内容词语相关度的网络热点话题提取[J]. 西安交通大学学报,2007,41(10):1142-1145,1150.

[7] 王巍,杨武,齐海凤. 基于多中心模型的网络热点话题发现算法[J]. 南京理工大学学报:自然科学版,2009,33(4):422-431.

[8] SEO Y W,SYCARA K. Text clustering for topic detection[D]. [S. l.]:Carnegie Mellon University,2004.

[9] CALBONELL A J,DODDINGTON J,YAMRON G,*et al*. Topic detection and tracking pilot study final report[C]//Proc of DARPA Broadcast News Transcription and Understanding Workshop. San Francisco:Morgan Kaufmann Publishers,1998:194-218.

[10] ALLAN J,HARDING S,FISHER D,*et al*. Taking topic detection from evaluation to practice[C]//Proc of the 38th Hawaii International Conference on System Sciences. Hawaii,MA: Kluwer,2005:199-204.

[11] SEO Y W,SYCARA K. Text clustering for topic detection[D]. [S. l.]:Carnegie Mellon University,2004.


(上接第 43 页)几类算法;在对函数 $f_3$ 的优化中 MAGA 和 StGA 花费的平均函数评估次数最少,EGEA 与 HTGA 搜索结果的精度最高;在对函数 $f_4$ 和 $f_5$ 的优化中 MAGA 表现最佳,EGEA 优化结果的精度一般但花费的函数评估次数要远少于 OGA/Q、StGA 和 OEA。综合来说,基于大配子自适应参数筛选策略使 EGEA 成为了一种具有一定竞争力的无约束函数优化算法。

表 2　EGEA 与五种算法对 $f_1$ ~$f_6$ 优化结果的对比

| function | | $f_1$ | $f_2$ | $f_3$ | $f_4$ | $f_5$ | $f_6$ |
|---|---|---|---|---|---|---|---|
| EGEA | MNFE | 28786 | 11114 | 21409 | 25858 | 20391 | 9009 |
| | MFV | -12569.2367 | 0 | 0 | 2.5e-4 | 6.8e-2 | 0 |
| | (Std) | (1.5e-1) | (0) | (0) | (8.7e-3) | (3.9e-4) | (0) |
| OGA/Q | MNFE | 302166 | 224710 | 112421 | 134000 | 112559 | 112612 |
| | MFV | -12569.4537 | 0 | 4.44e-16 | 0 | 0 | 0 |
| | (Std) | (6.447e-4) | (0) | (3.9e-16) | (0) | (0) | (0) |
| MAGA | MNFE | 10862 | 11427 | 9656 | 9777 | 9502 | 9591 |
| | MFV | -12569.4866 | 0 | 4.44e-16 | 0 | 0 | 0 |
| | (Std) | (7.121e-12) | (0) | (0) | (0) | (0) | (0) |
| HTGA | MNFE | 163468 | 16267 | 16632 | 20999 | 20844 | 14285 |
| | MFV | -12569.46 | 0 | 0 | 0 | 0 | 0 |
| | (Std) | (0) | (0) | (0) | (0) | (0) | (0) |
| StGA | MNFE | 1500 | 28500 | 10000 | 52500 | 30000 | 17600 |
| | MFV | -12569.5 | 4.42e-13 | 3.52e-8 | 2.44e-17 | 2.45e-15 | 2.03e-7 |
| | (Std) | (0) | (1.1e-13) | (3.5e-9) | (4.5e-17) | (5.2e-16) | (2.95e-8) |
| OEA | MNFE | 300019 | 300019 | 300018 | 300020 | 300017 | 300014 |
| | MFV | -12569.4866 | 5.430e-17 | 5.336e-14 | 1.317e-2 | 2.481e-30 | 2.068e-13 |
| | (Std) | (5.555e-12) | (1.683e-16) | (2.954e-13) | (1.561e-2) | (1.128e-29) | (1.440e-12) |

## 3　结束语

大配子机制可有效提高族群聚类的效率以及族群进化的有效性,因此筛选出结构合理的大配子成为了提高 EGEA 性能的关键。本文提出了两类三种大配子筛选策略,通过仿真实验发现,自适应参数法能够有效提高族群组织对无约束函数的优化效率,使 EGEA 成为了一种有竞争力的优化算法。

**参考文献:**


[1] CHEN H,CUI D W,LI X,*et al*. The harmonious evolution of ethnic group algorithm[C]//Proc of the 3rd International Conference on Natural Computation. Haikou:IEEE Computer Society,2007:380-384.

[2] 陈皓,崔杜武. 基于择偶的族群繁殖机制[J]. 计算机工程,2009,35(18):7-8.

[3] 陈皓,崔杜武,严太山,等. 基于竞争指数的模拟退火排序选择算子[J]. 电子学报,2009,37(3):586-591.

[4] 陈皓,崔杜武,李雪,等. 交叉规模的优化与交叉算子搜索性能的改进[J]. 软件学报,2009,20(4):890-901.

[5] ZHONG W C,LIU J,XUE M Z,*et al*. A multiagent genetic algorithm for global numerical optimization[J]. IEEE Trans on Systems,Man,and Cybernetics,Part B:Cybernetics,2004,34(2):1128-1141.

[6] LEUNG Y W,WANG Y P. An orthogonal genetic algorithm with quantization for global numerical optimization[J]. IEEE Trans on Evolutionary Computation,2001,5(1):41-53.

[7] TSAI J T,LIU T K,CHOU J H. Hybrid Taguchi-genetic algorithm for global numerical optimization[J]. IEEE Trans on Evolutionary Computation,2004,8(4):365-377.

[8] TU Z G,LU Y. A robust stochastic genetic algorithm(StGA) for global numerical optimization[J]. IEEE Trans on Evolutionary Computation,2004,8(5):456-470.

[9] LIU J,ZHONG W C,JIAO L C. An organizational evolutionary algorithm for numerical optimization[J]. IEEE Trans on Systems,Man,and Cybernetics,Part B,2007,37(4):1052-1064.