CULTURAL ALGORITHM BASED ON EVOLUTION STRATEGIES AND ITS APPLICATION IN TOPIC CLUSTERING

A Dissertation Submitted

to the Graduate School of Henan Normal University

in Partial Fulfillment of the Requirements

for the Degree of Master of Engineering

By

Luo Leiming

Supervisor: Prof. Wang Xiaodong

April, 2010

摘 要

话题识别与跟踪，作为一项旨在帮助人们应对信息过载问题的研究，以新闻专线、广播、电视等新闻媒体信息流为处理对象，将语言形式的信息流分割为不同的新闻报道，监控对新话题的报道，并将涉及某个话题的报道组织起来以某种方式呈现给用户。它的研究目标主要是对网络信息流进行一定的预处理之后对报道进行切分、话题识别、话题跟踪等，在这些任务中不可避免地要用到一些数据挖掘的理论知识以及相关的算法实现，所以选取什么样的分类聚类算法，达到什么样的效果以及对结果如何评价，都是目前该领域正在研究的热点问题。

话题识别是话题识别与跟踪的一项子任务，聚类算法是话题识别的核心技术。本文针对话题识别的聚类算法做研究，用 K-means 模型作为聚类模型，并结合进化策略的文化算法作为其进化寻优机制来对算法进行设计。针对以上思路，本文主要内容如下：

(1) 对聚类算法中用到的进化算法进行详细探讨，包括进化算法的三个主要分支遗传算法、进化规划、进化策略。通过算法比较，确定进化策略作为 K-means 聚类模型下的文化算法的种群空间，并对进化策略中的重要算子进行详细研究，为聚类算法设计奠定基础。

(2) 依据文化算法的框架分别对文化算法的种群空间、信仰空间以及这两个空间中的通信协议即影响函数和接受函数进行研究，探讨各种函数的工作机制，并研究嵌入文化算法框架的进化策略种群空间。

(3) 根据话题识别的聚类算法要求，对文化算法中的种群空间和信仰空间等进行设计，提出结合 K-means 算法的混合聚类算法。选取一定的语料对话题文本进行聚类实验，对实验结果进行分析，验证了提出的算法在话题识别中应用的有效性。

关键词：进化策略，文化算法，话题识别与跟踪，进化算法，K-means

ABSTRACT

Topic detection and tracking, as one study of helping people to cope with overload information, deals with the flow of information such as newswire, radio, television and other news media, splits the language forms of information flow into different news reports, monitors the New topic stories and displays the related reports to users in a particular way. Its primary goal is to segment reports, detect and track topics after preprocessing the network information flow, which will inevitably involve some concepts of data mining and related algorithms. So it is the hot problem of this field that choosing what kind of classification and clustering algorithms, achieving what kind of effect and how to evaluate the results.

Topic Detection is a sub-task of Topic Detection and Tracking and clustering algorithm is its core technology, so this paper chooses the clustering algorithm, which is used to identify topics, as the research purposes, uses K-means model as a clustering model and gives a culture algorithm combined with evolution strategies as its evolutionary optimization mechanism to achieve the design of algorithm.

(1) The evolutionary algorithms, which are used in clustering algorithm, are discusses in detailed and this paper presents three main branches of evolutionary algorithms such as genetic algorithms、evolutionary programming and evolution strategies. Through comparing various algorithms, we adopt evolution strategies as the population spatial of the Cultural Algorithm under the K-means framework, then we give a detailed analysis on the important operator in evolution strategies, thus we have a clear understanding about the evolutionary mechanism of evolution strategies.

(2) This paper discusses the population space and the belief space of cultural algorithm respectively as well as the two space communication protocols which contains the influence function and receiving function under the cultural algorithm framework, points out the working mechanism of various functions, embeds the evolution strategies in the culture algorithm as its population space.

(3) Finally, this paper has deeply researched the application in the topic detection by using the new evolution strategies culture algorithm, designs the cultural population space and beliefs space and other related technologies according to the demand on topic detection clustering algorithm, a hybrid clustering algorithm combinating K-means algorithm is then proposed. After that, we have carried out a clustering

experiments by choosing a certain topic of the text corpus, and the results prove that the new algorithm is feasible and effective in the application of topic detection.

KEY WORDS: Evolution Strategies, Cultural Algorithm, Topic Detection and Tracking, Evolutionary Algorithm, K-means

目 录

第一章 绪 论

1.1 研究背景及意义

中国互联网络信息中心（CNNIC）在京发布了《第 25 次中国互联网络发展状况统计报告》[1]（以下简称《报告》）。《报告》数据显示，截至 2009 年 12 月，我国网民规模已达 3.84 亿，互联网普及率进一步提升，达到 28.9%，网民数量稳居全球第一。这些网民阅读新闻首选互联网，而且习惯于在网上随时发表意见。目前常用的网民意见表达载体有新闻跟帖、留言、论坛/BBS、贴吧、博客/个人空间、播客、网络调查等。网络对人们的影响力越来越大，然而网络信息如此丰富，这就必然在信息筛选时给我们带来一些问题，如何在非常繁重的公务之余读透互联网？靠自己或信息部门人工浏览和摘报网络信息，难免挂一漏万；如何在信息爆炸和信息过载的情况下有效地组织和分析信息并快捷准确地获取有用信息成为人们在新的信息科技革命时代关心的主要问题之一。

作为一项旨在帮助人们应对信息过载问题的研究，话题识别与跟踪[2]（Topic Detection and Tracking，TDT），以新闻专线、广播、电视等新闻媒体信息流为处理对象，将语言形式的信息流分割为不同的新闻报道，监控对新话题的报道，并将涉及某个话题的报道组织起来以某种方式呈现给用户。它的研究目标主要是对网络信息流进行一定的预处理之后进行切分、话题识别、话题跟踪等任务。在这些任务中关键问题之一是选取什么样的分类聚类算法。

针对话题识别任务，其关键研究的内容是话题聚类技术，其中主要聚类算法有基于划分的 K-means 聚类算法和 Single-Pass 聚类方法、基于层次的 GAC 层次凝聚法等。基于划分的聚类算法在初始值选取时比较敏感，适用于样本较集中时，即类内距离小而类间距离大的情况，这种情况下对聚类阈值的选取比较容易。而样本不集中时，由于样本的处理顺序不同，阈值选取也就会随之不相同，对结果的影响比较大。而 GAC 算法比较适合于话题的回溯识别，不适合话题的在线识别，不能够满足话题动态性的要求，应用范围受到一定的限制。本文基于进化策略文化算法研究话题聚类，提出 K-means 聚类模型下的聚类算法。该算法通过信仰空间保留较优信息并指导种群空间的搜索，有效地利用全局信息，可避免传统划分方法易陷入局部最优以及对初始阈值选取敏感的问题。

另外算法的信仰空间也可以随着样本组成的种群空间变化进行动态的更新，这对 GAC 算法不能适应动态信息也是一种弥补。因此，通过对进化策略文化算法进行深入研究可以在上述两方面对传统的算法进行改进，具有很好的理论及现实应用价值。

1.2 国内外研究现状

1.2.1 话题识别与跟踪

TDT 的概念最早是 1996 年[2]美国国防高级研究计划署（DARPA)根据自己的需求，提出的一种新技术，这种技术希望计算机能在没有人工干预的情况下自动判断新闻数据流的主题。1997 年，研究者用漏检率和误检率两个指标评价 TDT 系统的性能，开始了对这项技术进行初步研究。并且还用一种识别错误权衡图（DET，Detection Error Tradeoff plot）来直观地描绘 TDT 系统发生错误的情况。从 1998 年开始，DARPA 以及后来的美国国家标准技术研究所（MST）资助并主持了话题识别与跟踪系列测评会议。它是一种评测驱动的研究方式，具有如下特点：明确的形式化研究任务、公开的训练与测试数据、公开的评测比较。它将研究置于公共的研究平台上，使得研究之间的比较更加客观，从而让研究者认清各种技术的优劣，起到正确引导研究发展方向的目的。参加该测评的机构包括著名的公司、大学和研究所，如 IBM Watson 研究中心、GE、Dragon systems、BBN 公司以及 CMU、UMASS、Cambridge 等一流大学[3]。1998 年 TDT 在首次公开测评中，将漏检率与误检率结合起来的新指标——错误识别代价——作为主要的评测指标。这次的评测除了原有的英语外，还引入了汉语（普通话）。美国国家标准技术研究所从 1999 年开始主持 TDT 系列评测，都取得了令人振奋的效果。

TDT 研究在中国大陆的研究起步比较晚，但近几年发展很快。中科院计算所、东北大学、北京大学、哈尔滨工业大学、中山大学、复旦大学、大连理工大学、四川大学等单位都对话题识别与跟踪的关键技术进行了相关研究，取得了较好的成果。

在 TDT 研究所用到技术方面，话题识别子任务中主要用到的是聚类技术。在聚类算法的选择上国内外大同小异，M.Steinbac[4]等人对常用的层次聚类算法和 K-means 算法进行了详细地比较和阐述。Young-Woo Seo[5]提出了基于神经网络的迭代竞争算法，对比层次算法和 K-means 算法取得较好的聚类效果。B.Choudhary[6]等在特征向量的生成中考虑句子中词汇间的语义关系，用神经网络方法对特征向量进行聚类，实验证明该方法比仅考虑词频作为特征的方法有更好的表现。国内杨建武[7]介绍两种常用的聚类算法，

基于平均分组（Group Average Clustering，简记为 GAC）的层次聚类算法（GAC-based hierarchical clustering）和单路径聚类算法（Single-Pass clustering），指出基于平均分组的 GAC 算法较适宜于回溯话题识别，单路径聚类算法常用于在线话题识别。贾自艳[8]等人借鉴 Single-Pass 聚类思想，通过大量分析网络新闻特点并结合新闻要素给出了一种基于动态进化模型的事件识别和追踪算法，该算法可以自动对新闻语料进行组织生成不同话题下的新闻专题，从而能为用户提供比较个性化的服务。税仪冬[9]针对增量式聚类初始时话题模型不够充分和准确，并随处理报道数量增多，误检与漏检的累积效应被放大的问题，提出了周期分类和 Single-Pass 聚类相结合的话题识别与跟踪方法，实验表明这种方法是有效的，能够降低漏检率与误检率，减少归一化错误识别代价。赵华[10]结合 Single-Pass 聚类思想提出了基于时间信息的动态阈值模型，克服话题检测中使用静态阈值的缺点。可以看出为适应话题聚类的特点，算法还需要在初始值的选取以及对动态性要求方面作进一步研究。

1.2.2 文化算法及其在聚类分析中的应用情况

文化算法[11]自 1994 年由 Reynolds 提出以来，在国外有了比较多的应用。Reynolds 和 Chung 于 1995 年起利用文化算法求解全局优化问题。Chung 关注于解决静态无约束实值函数优化，他提出了两种知识类型：形势知识和规范知识，后来 Zannoni 和 Reynolds 于 1996 年将遗传规划嵌入文化系统框架（即 CAGP），用于控制规划进化过程（Program Evolution Process），另外国外学者还将文化算法用于图像分割[12]、语义网络[13, 14]、动态优化问题[15]、数据挖掘[16]等。

国内文化算法应用刚刚起步，其中杨海英[17]等提出一种基于文化算法的负载均衡自适应机制将文化算法应用到对服务器性能权值的进化计算中，通过评价服务器的负载状况“获得优化的性能权值”并自适应地转换到集群的分配器中，使事务在集群系统中得到合理分配。张鹤峰[18]等人一种将 Chan（一种定位基本算法）算法与文化算法相结合的算法，利用该算法解决 TDOA（一种蜂窝网定位技术）定位估计中遇到的非线性最优化问题。吴英[19]等将文化算法应用到电力系统无功优化中，与改进遗传算法（SGA）和改进粒子群算法（CPSO）比较，得到了比较理想的结果。另外张涤[20]研究了基于文化算法的聚类分析技术，其中提到了 Web 数据的聚类。孟凡荣[21]提出了基于文化算法的模糊聚类算法，通过实验证明新算法在一定程度上避免模糊 C 均值算法对初始值敏感和容易陷入局部最优解的缺陷。刘纯青[22]分析了 K-means 聚类算法所存在的不足，提出了基

于文化算法的新聚类算法，证明其全局收敛性能优于基于遗传算法的 K-means 聚类算法。

总之，对文化算法的应用研究才兴起不久，相比其他成熟的进化算法，其在聚类以及其他技术中应用还需要进一步研究，相信随着研究的深入，其应用领域将会越来越广泛，也会在话题聚类中有良好的表现。

1.3 研究内容及结构安排

1.3.1 研究内容

本文把进化策略作为文化算法种群空间的进化构架，并设计相应的信仰空间以及用于两个空间通信的影响函数和接受函数，提出一种基于进化策略的文化算法，之后把基于进化策略的文化算法用到话题识别中的文本聚类中，最后再与传统的 K-means 聚类算法进行实验对比。

1.3.2 结构安排

本文共五章，各章主要内容如下：

第一章绪论。提出问题并分析目前 Web 话题识别和文化算法的研究现状，给出本文创作的意义，在对现状进行详细分析的基础上确定出本文的研究内容。最后给出论文所作的工作、文章的组织结构。

第二章文化算法及话题识别基本原理。由于进化策略是进化算法的一种，该章节首先对进化算法进行了简要的阐述，之后论述文化算法相关知识，对其框架以及其在各个行业的应用作相关介绍，第三节是话题识别的有关知识，指出话题识别是什么课题，以及话题识别分为哪些种类，最后对话题识别中用到的方法和流程作了大致介绍。

第三章进化策略文化算法设计。由于基于进化策略研究文化算法框架下的种群空间，所以本章先对进化策略这一进化算法再做更为详细的介绍，重点对其内部的重要算子加以讨论，以对进化策略有一个更为清晰的认识，之后对文化算法中信仰空间以及影响函数和接受函数进行了设计，并将其用于无约束函数优化中。

第四章话题聚类设计与评价。将新设计出的进化策略文化算法应用到话题识别中去，首先要对话题文本数据进行预处理，形成话题文本的向量表示，以适于话题聚类算法。第二节则是根据具体的聚类算法要求设计适合于话题聚类的种群空间和信仰空间。

之后给出算法的总体设计和描述并介绍了聚类性能评价方法，最后通过实验对针对话题聚类的算法进行仿真实验并给出结果分析，以判断进化策略文化算法的聚类效果。

第五章是总结与展望部分。对论文的研究工作进行总结，对不足和有待改进的地方进行分析，并对下一步的工作进行设想。

第二章 文化算法及话题识别基本原理

2.1 进化算法

进化算法是一种模拟生物在自然界的遗传和进化过程以及机制，从而求解优化问题的自适应人工智能技术。它的基本思想来源于达尔文的生物进化学说，即生物通过遗传和突变，按照"优胜劣汰、适者生存"的规则进行生物种群的进化过程。比如美国 Michigan 大学的 J.H.Holland 教授模拟生物进化和遗传过程，首次提出遗传算法（Genetic Algorithm，GA）[23]。美国的 L.J.Fogel 在研究人工智能的过程中于上世纪 60 年代提出来了一种随机的优化方法，这种方法也借鉴了自然界生物进化的思想，即是进化规划算法。他认为智能行为必须包括预测环境的能力，以及在一定目标指导下对环境作出合理响应的能力。另一个比较著名的进化算法进化策略（Evolution Strategies，ES）是出德国的 I.Rechenberg 和 H.P.Schwefel 于 1963 年提出的。除了上述常用的进化算法外，一些新颖的优化算法也得到了迅速发展，如人工神经网络（ANN）在一定程度上模拟了人脑的组织结构；蚁群算法（ACO）受启发于自然界蚂蚁的寻径方式；模拟退火（SA）思路源于物理学中固体物质的退火过程，粒子群优化算法（PSO）[24]则是源于对鸟群和鱼群群体运动的研究。

2.1.1 生物学背景

从生命进化史我们可以看出生物进化的道路是曲折的，其进化过程表现为种种特殊的千奇百怪的情况。这期间有生物类型的增多，有某些生物物种的灭绝，甚至还有些生物性状的退化，但从总体上来看，整个生物界的演化还是一种进化过程，是一种从简单到复杂，从低级到高级的过程。

当然，总体上的这种进步过程中，是以生物种群之间的相互竞争来实现的，资源的有限性导致了这种情况的发生，一种生物为求生存，它必然要和另一种需要同样资源的生物争夺这些资源，在这种生存竞争过程中大部分时候是以消灭对方作为结果。这样，获得资源的生物得以延续，未获得资源的生物会被淘汰掉，延续下来的生物同样进行着生存竞争，并由自然环境决定其能否继续延续下去。这里生物的延续是一种生物遗传的作用，而生物能够适应环境是生物产生了有利于环境生存的变异，达尔文就是通过对生

物界存在的这种遗传和变异现象进行长期的观测并借鉴前人研究成果的基础上提出了生物进化论，即“物竞天择，适者生存”的自然选择学说。

达尔文的生物进化论指出，遗传和变异是决定生物进化的内在因素。遗传是指物种的两代之间之间在性状上存在的相似现象，而变异则是父子两代性状上又存在或多或少的不同。随着现代细胞科学以及遗传学的发展，我们知道遗传和变异的物质载体是DNA（脱氧核糖核酸），它由两条脱氧多核苷酸链反向平行盘绕所形成的双螺旋结构组成，在细胞内它组织成染色体结构，其上具有遗传效应的片段称为基因。生物性状的延续即是基因通过复制把其携带的遗传信息传递给下一代，在基因复制过程中也可能发生突变产生变异性状的个体（另外染色体层次上也可以发生变异）。正是通过这种遗传（基因的复制）和变异（基因的交叉、重组和变异），才产生了如今丰富多彩的生物世界。生物的遗传使得生物能够保留适应环境的性状而是物种能够稳定的延续，而生物的变异特性（虽然是一个很小的概率）使得生物能够适应外界变化的环境而使物种得以进化。

2.1.2 进化算法基本原理

进化算法仿效生物的进化和遗传，将生物学中的变异进化特性用于优化计算中。不仅进化算法的指导性原则与生物学吻合，而且基本术语也相似。

进化算法是以字符串或字符段为运算基础的，字符串或字符段这相当于生物学中的染色体，其上有一系列字符组成，每个字符都有自己的含义，相当于生物学中的基因。进化算法的迭代过程，类似于生物学的逐代进化。进化算法中的选择（复制），体现生物界中“自然竞争、优胜劣汰”。进化算法的交换（重组），相当于生物界的交配，进化算法的突变，相当于生物界的变异。

进化算法中具体各种名词解释如下：

种群：进化计算在最优解的搜索过程中，一般从原问题的一组可行解出发改进到另一组较优解，再从这组较优解出发进一步改进寻优。在进化计算中，每一组解称为“种群”（Population），而每一个解称为一个“个体”（Individual），当然这在不同的进化算法选择中有具体的编码方式和表现形式。

编码：进化计算中每一个解被看成是一个生物个体，一般要求用一条染色体（Chromesome）来表示，即用一组有序排列的基因（Gene）来表示，这与在普通的搜索算法中，解的表达可以采用任意的形式不同，普通搜索算法一般不需要进行特殊的处理。这就要求当原问题的优化模型建立之后，还必须对原问题的解（即决策变量，如优

化参数等）进行编码。

遗传算子：应用下述三种操作（至少前两种）来产生新的群体：

复制：把现有的个体字符串复制到新的群体中。

变异：将现有的字符串某一位的字符随机改变。

交叉：把两个父代个体中的部分结构加以替换重组而生成新个体的操作。

进化算法的一般步骤为：

(1) 初始化：随机生成初始种群 $P(0)$，进化代数计数器 $t=0$；

(2) 评价种群 $P(t)$ 适应度；

(3) 根据一定规则选取当前种群 $P(t)$ 的一部分作为下一代的解的基础；

(4) 对（3）中选取出的解进行操作（重组和变异），生成新一代解 $P'(t)$；

(5) 评价 $P'(t)$ 的适应度；

(6) 考察是否达到终止条件，若满足终止条件输出最优个体终止循环，否则 $t=t+1$，转到第三步继续执行。

其伪代码描述如下：

```
public  EC( )
{
    t=0;
    Initialize P(t);
     CalculateFitness(P(t));

    while(t <= tmax)
    {

      P′(t)=Recombination(P(t)); //重组
      P″(t)=Mutation(P(t));        //变异
      P(t+1)=Reproduction(P′(t)∪P″(t)); //选择
      CalculateFitness(P(t));
      if(Fitness(P(t))>Fit)    //若满足终止条件，则终止
        break;
    }
}
```

2.1.3 进化算法主要分支

目前研究的进化算法主要有三种典型的算法：遗传算法、进化规划和进化策略。这

三种算法是彼此独立发展起来的，各自有不同的侧重点，不同的生物进化背景，各自强调了生物进化过程中的不同特性。

下面对这三种典型算法作一介绍：

1. 遗传算法

遗传算法是最具有代表性的进化算法，也是最早提出的一个进化算法，二十世纪六十年代密歇根大学的 J.Holland 教授在研究能学习的机器时就发现生物进化过程中蕴含的朴素的进化思想可以用到机器的优化计算上来。于是他开始对达尔文的进化论以及孟德尔的遗传定律进行深入研究，对生物在种群繁衍与进化期间染色体之间的复制、杂交、变异等机制进行抽象处理，最终提出了一种具有历史意义的进化算法。由于它源于模拟生物的遗传进化的思想，因而 Holland 教授为其命名为遗传算法[25]（Genetic Algorithm，GA），并将其发表在 1975 年他的专著《Adaptation in Natural and Artificial Systems》上。

(1) 表示法和适应值度量

根据编码方式的不同，遗传算法主要可分为二进制型、序列型和浮点型等三种，分别适用于不同类型的具体工程优化问题。标准遗传算法作用于确定长度的二进制位串上，即 $I=\{0,1\}^l$。这种表示法可以直接采用于伪布尔目标函数。为了解决函数优化的问题 f: $\prod_{i=1}^{n}[u_i,v_i]\to R(u_i<v_i)$，一般是将位串分为 n 段，每段长度为 l_x，即 $l=n\cdot l_x$，每段表示分量 $x_i\in[u_i,v_i]$ 的二进制代码。位段译码函数 Γ^l：$\{0,1\}^{l_x}\to[u_i,v_i]$ 的常见形式为：

$$\Gamma(\alpha_{i1},\cdots,\alpha_{il_x})=u_i+\frac{v_i-u_i}{2^{l_x}-1}\left(\sum_{j=1}^{l_x}\alpha_{ij}2^{j-1}\right) \tag{2-1}$$

其中 $(\alpha_{i1},\cdots,\alpha_{il_x})$ 记为个体 $\alpha=(\alpha_{11},\cdots,\alpha_{nl_x})\in I^l$ 的第 i 段。那么一个个体的译码函数为 $\Gamma=\Gamma^l\times\cdots\Gamma^n$，其相应的适应值函数可以表示为 $\Phi(\alpha)=\delta(f(\Gamma(\alpha)))$，其中 δ 为比例变换函数，它使得适应度都非负而且最优个体的适应度最大。线性比例变换、幂比例变换以及指数比例变换函数是常用的比例变换函数，另外一些方法见文献[26,27]。需要指出的是，二进制编码虽易于实现，并有类似于生物染色体的组成而使算法易于用生物遗传理论解释的特点，但也有其相应的缺点，如在求解连续优化问题时存在的 Hamming 悬崖问题，即相邻两整数的二进制编码可能有较大的 Hamming 距离，例如 7 和 8 的二进制为 0111 和 1000，则算法从 7 变到 8 需要改变所有的位，搜索效率也就随之降低了；另外求解高维问题是二进制的编码过长同样对搜索效率是一种阻碍。所以，还是需要根据

具体问题选择不同的编码方式以适应具体问题的要求。

(2) 变异算子

在标准遗传算法中，二进制编码的变异算子非常简单，它即是以较小的概率 p_m（变异概率）作用在位串上，随机地改变串上的某一位的值，即如果原来位上的数字为 1，则变为 0；若是 0，则变为 1。变异概率 p_m 的值一般取为 0.001 到 0.01 之间，它不依赖目标变量的维数和位串的总长。变异算子 $m'(p_m)$ 作用个体 $(s_1,\cdots,s_l)$ 后将其变为 $(s_1',\cdots,s_l')$，其中：

$$s_i' = \begin{cases} s_i, & \theta_i > p_m \\ 1-s_i, & \theta_i \le p_m \end{cases} \quad \forall i \in \{1,\cdots,l\} \tag{2-2}$$

这里 $\theta_i \in random[0,1]$，即 θ_i 是 0 与 1 之间随机产生的一个数，它需要对每一位都要重新产生随机数进行判断。

(3) 重组算子

在标准遗传算法中，杂交即重组算子是主要的遗传算子，它对两个不同个体上的有用段进行重新组合。经典的遗传算法是一种单点杂交，其杂交算子 $r': I^\mu \to I$ 也是作用在位串上，以概率 p_c 对两个个体进行重组，p_c 的作用范围一般是为[0.6，1.0]。如两个父辈个体 $s=(s_1, \ldots, s_l)$，$v=(v_1, \ldots, v_l)$ 被随机地从群体中选择进行杂交，杂交点为第 h 个位，则其产生的两个子代个体为：

$$s'=(s_1, \ldots, s_{h-1}, s_h, v_{h+1}, v_l) \tag{2-3}$$

$$v'=(v_1, \ldots, v_{h-1}, v_h, s_{h+1}, s_l) \tag{2-4}$$

其中杂交点 h 为 1 到 l 之间的一致随机整数。另外还有两点杂交方式，即是同时选取两个杂交点，位于两杂交点之间的位串相互交换，两侧的位串保留在原串。当然根据杂交点数的增多可以有 m 点杂交算子，其原理是一样的。另外一种需要说的杂交算子是均匀杂交算子，这种杂交算子是依概率交换两个父串的每一位。其大概原理是随机地产生一个与父串相同位数的二进制串，作为两个父串杂交模板，即该模板上位数为 0 时两父串对应位不交换，而模板为 1 时两父串对应位相互交换，所得的两个新串即为后代串。不过还没有明确的理论证明哪一种杂交算子能够更好的满足现实需要，不过人们正在试图为解决这一问题而做出深入的研究。

(4) 选择算子

标准遗传算法采用的是一种依据选择概率大小进行复制的选择算子$s:I^{\mu}\to I^{\mu}$，个体α_i被选择的概率由它在整个种群中的相对适应值给出：

$$p_S(\alpha_i)=\frac{\Phi(\alpha_j)}{\sum_{j=1}^{u}\Phi(\alpha_j)},\forall i\in\{1,\cdots,u\} \tag{2-5}$$

其中$\Phi(\alpha_i)$为群体中第i个成员的适应值，u为群体规模。依据这个概率求出的值，并在整个种群空间内进行排名，一种方法是选取排名靠前的μ个个体作为父个体产生下一代，另一种方法是把每个个体的概率分布看作是在一个圆盘上的比例分布，先随机生成一个[0，1]之间的随机数r，若$p_S(\alpha_1)+p_S(\alpha_2)+\cdots p_S(\alpha_{i-1})<r$并且同时满足$r\le p_S(\alpha_1)+p_S(\alpha_2)+\cdots p_S(\alpha_i)$，则选择个体$i$，重复$\mu$次即得到$\mu$个个体。从上面讨论发现，这里的比例选择算子不符合出现负适应值的情形或最小化任务的情况，此时要采用适应值比例变换的方法。

与传统优化算法相比较，遗传算法的主要特点有以下四个方面：

(1) 群体搜索策略。传统优化算法采用点到点的搜索方式；而遗传算法则采用群体到群体的搜索方式，因而较易于达到全局最优。

(2) 进化搜索不依赖于目标函数的梯度信息，只需给出能够评价个体相对优劣的指标（适应度）即可。因此，遗传算法的适用面更广，尤其适合于处理复杂的非线性问题，或是非确定性多项式时间复杂性类（NP-hard）问题，包括目标函数为高维、不连续、不可导或带有噪声的优化问题；而传统优化方法对此难以或着根本无法解决。

(3) 进化过程具有有向随机性，即遗传算法能够逐步寻优迭代，而与穷举法中没有方向的遍历所有点的做法存在本质区别。同时，遗传算法的搜索由于是种群集合的同时变化，与传统优化方法相比，相对耗时致使优化效率相对较低。

(4) 简单通用，鲁棒性强。

2. 进化规划

上个世纪 60 年代中期，L.J.Fogel 等人为有限状态机的演化提出了进化规划模型来求解预测问题[28]。在这个进化模型中，机器的状态基于均匀随机分布的规律来进行变异，即这些机器的状态变换表是通过在对应的离散、有界集上进行一致随机变异来修改。在研究中他们用有限字符集中的符号组成的序列描述模拟环境，于是问题变成了怎样去响应当前状态下的符号序列以获得最大的收益，这里的收益是由环境中将要出现的下一个

符号及预先定义好的效益目标确定。进化规划即是根据被正确预测的符号数来度量适应值。通过变异，父辈群体中的每个机器产生一个子代，并从这样的两代中各选出最好的一半组成新的一代。通过对正态分布变异算子的研究，D.B.Fogel 将进化规划扩展到求解实数值问题中去。

(1) 表示法和适应值度量

进化规划通过假设一个有界子空间 $\prod_i^n[u_i,v_i]\subset R^n$ 的基础上将搜索区域扩展到空间 $I=R^n$，其中 $u_i<v_i$，即个体成为目标变量向量 $\bar{\alpha}=\bar{x}\in I$。进化规划把目标函数值通过某种比例变换方式将其变换成正值，并加入某个随机改变 θ：$\Phi(\alpha)=\delta(f(x),\theta)$ 来得到适应值，其中 δ 是某种比例函数。

(2) 变异算子

标准进化规划采用的是高斯（Gaussian）变异算子，它把一个标准偏差作用与个体 x 的每个分量 x_i，这个标准偏差值取为适应值 $\Phi(\alpha)=\Phi(x)$ 的一个线性变换的平方根，即 $m'(x)=x'$ 其中

$$x_i'=x_i+\sigma_i\cdot N_i(0,1)\,, \tag{2-6}$$

$$\sigma_i=\sqrt{\beta_i\cdot\Phi(x)+\gamma_i}\,,\ \ \forall i\in\{1,\cdots,n\} \tag{2-7}$$

其中 $N_i(0,1)$ 表示对每个重新采样的指标 i，作用于该指标的具有期望值为 0 标准差为 1 的正态分布（即 Gaussian 分布）的随机变量，系数 β_i，γ_i 是待定参数。一般将 β_i 和 γ_i 置为 1 和 0。此时

$$x_i'=x_i+\sqrt{\Phi(x)}\cdot N_i(0,1) \tag{2-8}$$

这里与进化策略和遗传算法不同的地方是，进化规划中没有用到重组（交叉）算子。

(3) 选择算子

在 μ 个父辈个体每个都需要经过一次变异，产生 μ 个子代。之后进化规划利用一种随机 q 竞争选择方法从父子两代的集合中选择 μ 个个体，其中 $q\geq1$ 是选择算法的参数。具体作用过程是：对每个个体 $\alpha_k\in P(t)\bigcup P'(t)$，其中 $P'(t)$ 是变异后的群体，从 $P(t)\bigcup P'(t)$ 两代个体中随机选取 q 个个体，把它们与 α_k 都按适应值进行比较，找出其中比 α_k 差的个体数目 w_k，并把 w_k 作为个体 α_k 的得分，显然 $w_k\in\{0,\cdots,q\}$；同样的过程来计算所有 2μ

个个体的得分，并按得分 $w_i(i\in\{0,\cdots,2\mu\})$ 的高低降序对个体进行排序；选择得分 w_i 排在前 μ 的个体作为下一代群体。更形式化地可以表述为：

$$w_i=\sum_{j=1}^{q}\begin{cases}1, & if\ \Phi(\alpha_i)\le\Phi(\alpha_{h_j})\\ 0, & if\ \Phi(\alpha_i)>\Phi(\alpha_{h_j})\end{cases} \tag{2-9}$$

$h_j\in\{0,\cdots,2\mu\}$ 为一致整型随机变量，这里要求对每个比较都要重新采样。这样最好的个体将有可能被置为最大适应值，从而保证最好的个体能够生存下来。

3. 进化策略

进化策略 ES[29]是于二十世纪六十年代由德国的 H.-P.Schwefel 和 I.Rechenberg 在研究流体动力学中的弯管形态优化过程中，共同开发出的一种适合于实数变量的、模拟生物进化的一种优化算法。其优化能力主要依靠变异算子的作用，后来受遗传算法的启迪，也引入了杂交算子，不过杂交算子在进化策略只起到辅助作用。

(1) 表示法和适应值度量

在进化策略中，搜索点是 n 维向量 $x\in R^n$，个体的适应值等于其目标函数值，即 $\Phi(\alpha)=f(x)$，其中 x 是个体 a 的目标变量部分。此外每个个体可以包括至多 n 个不同的方差 $c_n=\sigma_i^2(i\in\{0,\cdots,n\})$ 和至多 $n\cdot(n-1)/2$ 个协方差 $c_{ij}(i\in\{0,\cdots,n-1\},j\in\{i+1,\cdots,n\})$，从而至多 $w=n\cdot(n+1)/2$ 个策略参数可以和目标变量组合在一起构成一个个体 $a\in I=R^{n+w}$。不过，一般只考虑方差，从而 $a\in I=R^{2n}$，有时甚至对所有目标变量只用一个共同的方差，这时 $a\in I=R^{n+1}$。

(2) 变异算子

在进化策略中，全体 $\alpha=(x,\sigma)$ 在变异算子作用下变为 $\alpha'=(x',\sigma')$，其中

$$\sigma_i'=\sigma_i\cdot\exp(\tau'\cdot N(0,1)+\tau\cdot N_i(0,1)), \tag{2-10}$$

$$x_i'=x_i+\sigma_i'\cdot N(0,1),\quad i=1,\cdots,n。 \tag{2-11}$$

其中 $N(0,1)$ 表示具有期望值为 0 标准差为 1 的正态分布随机变量，τ 和 τ' 是算子集参数，分别定义整体和个体步长。

(3) 重组算子

在进化策略中，重组算子 $\gamma':I^{\mu}\to I$ 可以按下列方式产生一个个体：

$$x_i'=\begin{cases} x_{S,i} & \text{无重组} \\ x_{S,i}\text{或}x_{T,i} & \text{直接重组} \\ x_{S,i}+\Theta(x_{T,i}-x_{S,i}) & \text{加权平均重组} \end{cases} \tag{2-12}$$

下标 S 和 T 指从 P（t）中随机选取的两个父辈个体，$\theta\in[0,1]$ 为一致随机变量。

(4) 选择算子

在进化策略中，选择是按完全确定的方式进行。$(\mu,\lambda)-ES$ 是从几个子代个体集中选择 $\mu(1\le\mu\le\lambda)$ 个最好的个体；$(\mu+\lambda)-ES$ 是从父辈和子代个体的并集中选择 μ 个最好的个体。虽然 $(\mu+\lambda)-ES$ 保留最优的个体能保证性能单调提高，但这种策略不能处理变化的环境，因此，目前选用最多的还是 $(\mu,\lambda)-ES$，研究表明比例 $\mu/\lambda\approx 1/T$ 是最优的。

2.1.4 进化算法特点

为说明进化算法的特点，先讨论一下传统的寻优算法，传统算法主要有以下三种：

(1) 基于导数的方法。这种方法是一种解析法，即利用数学分析中求函数极值的方法，令函数的一阶导数为零，在一阶导数为零处往往是函数的极值点；另一种方法就是爬山法，它根据函数的一阶导数确定梯度的方向，并沿梯度增大的方向逐步向上迭代迁移，最终到达极值点，即峰顶。

(2) 穷举法。顾名思义，即是把所有可能的情况列出并通过比较找到最优解。求解函数极值时依次搜索 x,y 平面内所有 x_i,y_i 的 $f(x_i,y_i)$ 值，以求出函数的最大值位置。显然，这是一种很笨拙的方法，并且在多维空间下这是一种费时费力的方法，几乎不太可能实现。

(3) 随机搜索法。在 x,y 平面内随机确定搜索点 x_i,y_i，并通过计算其函数值 $f(x_i,y_i)$ 而尽力经过少量搜索找出函数的最大值点。很显然，这种方法没有严格的收敛性条件，具有很大的偶然性，在多维空间下同样比较耗时。

综合上面传统算法的说明以及对进化算法的讨论，可以看出进化算法具有与传统算法不能具有的以下优点[30]：

(1) 有指导搜索。进化算法是以适应度即目标函数为指导搜索策略，既不是盲目式的乱搜索，也不是穷举式的全面搜索。通过变量适应度的驱动使进化算法逐步逼近目标值，达到寻优效果。

(2) 自适应搜索。进化算法在搜索过程中，需要种群空间进行复制（选择）、交换（重

组)、突变等操作，并根据适应度的大小确定个体是否保留，这体现"适者生存，劣者淘汰"的自然选择规律，而不需要添加其他额外的作用，就可以使群体的品质得到不断的改进，即是一种自动适应环境的能力。

(3) 渐进式寻优。进化算法从随机产生的初始可行解开始迭代，这样一代代地反复计算，并以子代的结果优越于父代为目的，逐渐得出更优的结果而最终达到最优点，是一个逐渐寻优的过程。

(4) 隐式并行性。进化计算具有显著的隐式并行性（implicit parallelism)。进化算法虽然在每一代只对有限解个体进行操作，但处理的信息量为群体规模的高次方。

(5) 黑箱式结构。进化算法类似于黑箱结构，只要确定了种群的编码方式并初始化好种群后，根据适应度函数引导突变方向，直至搜索到最优解，至于里面是怎么一步步变异和交换重组的不需要关心，只关心它的输出的解是否达到了最优，这样便于处理因果关系不明确的问题。

(6) 应用广泛。传统的优化算法，一般是将所解决的问题用数学函数式表示，并要求该函数的一阶导数或更高阶导数存在。而进化算法则是根据特定的问题确定某种字符集描述问题，然后根据适应度的大小区分个体优劣。其中的选择、重组、突变等操作都是固定的，由计算机自动执行。它只有一些简单原则要求，在实施过程中，而不需要额外的干预。

(7) 鲁棒性比较强。即在存在噪声的情况下，对同一问题的进化算法的多次求解中得到的结果是相似的。进化算法的鲁棒性在大量的应用实例中得到了充分的验证。

2.2 文化算法

2.2.1 文化算法的提出

被称为"人类学之父"英国人泰勒（E.B.Tylor）在 1871 年发表的《原始文化》一书中最早把"文化"作为专门术语来使用。在那本书中他给文化下了定义：文化是一个复杂的总体，它包括知识、信仰、道德、艺术、法律、风俗，以及人类在整个社会里所习得的一切能力和习惯。从那以后，也有不少西方学者对文化下了定义，到目前为止形成了成百上千种关于文化的定义。

20 世纪 60 年代，人们在对控制论和系统论的研究中对"文化"一词有了一种新的思考，他们把文化看作一个能与环境进行交互的系统，并能够对系统中的人产生这样或那

样的影响，包括正面或负面的。在这个时期，伴随着这些理论的研究一门新的学科文化生态学（Cultural Ecology）出现了，这门学科重点研究的是文化的存在和发展的资源、环境、状态，及其相互影响的规律。到了20世纪70年代，人们开始重点研究一个系统中的有用信息是怎样在利用文化的过程中产生的。1973年时，文化人类学家克利福德·格尔兹（Geertz）注重从人的自然进化和文化互动的发展史来对对文化展开研究，他提出文化"不是被附加在完善的或实际上完善的动物身上，而是那个动物本身的生产过程的构成要素"，即文化是人类生产过程中用来解释他们的经验和指导他们行动的意义结构。1994年，Durham通过研究总结把文化描述为"一个通过符号编码表示众多概念的系统，而这些概念是在群体内部及不同群体之间被广泛和历史般长久传播的"[31]。

在人类社会中，文化是存在于一定文明、社会及社会群体（尤其是一个特殊的时代）中的包含了知识、习俗、信念、价值等的复杂系统，它被看作存储信息的载体，这些信息在社会群体之间和群体内部广泛地传递，并能被社会所有成员继承，从而有效地指导成员的行为解决问题。生物学中的模型和概念给解决计算问题提供了灵感。文化算法的提出是为进化计算系统中随时间增长不断累积经验的个体成分的进化建立模型。

在上一节中提到的多种进化算法只是模拟了现实世界生物进化的过程，还有一种进化算法是模拟人类社会进化的过程，并且许多情况表明，文化在社会进化中发挥着举足轻重的作用，它能使种群以超越单纯依靠基因遗传生物进化的速度进行进化和对环境地适应[32]。伦斯基认为社会进化与生物进化之间存在许多相似性，它们都是基于信息系统以编码形式的进化过程，生物进化中有种群基因的编码形式，人类社会进化则是世代相传的知识、经验等字符编码形式。但它们之间也是有差异的：在生物进化中，基因作为保存信息的载体并通过新有机体的复制来传递，它是一种无意识的进化，生物并不知道基因的突变方向是向适应环境的方向改变，或是相反的方向，并且速度比较缓慢，后天获取的技能不能通过遗传传给下一代；而在社会进化中，文化的主导作用十分明显，是一种有意识的进化方式，人类通过有意识的选取不同类型的符号进行传播，并且传播速度比基因遗传的方式更快、比基因遗传更直接、更灵活、更广泛，后天获得的知识形成文化并继续传递给下一代。在传递的过程中文化不断累积，其中不仅包括技术的积累，同样也包括知识的积累[31]。另外学者Renfrew[33]也指出，随着时间的迁移人类在进化过程中逐渐掌握了提取、编码和传播信息知识的能力，这种能力是其他物种所不具有的，是我们人类特有的能力。正是这种能力使人类出自然选择进化，发展到有意识的主动进

化，并形成、积累和传递人类特有的经验，大大加快了人类社会发展的速度。所以说人类社会的进化超越了自然选择的水平，甚至可以说人类发展到当今这个阶段以及其未来的发展文化都将起到决定性的作用。

正是受到上述思想的启示，文化算法才得以形成。最早对文化进行数学建模的是美国人 Reynolds，他通过对进化计算系统的经验积累建模研究，最早提出文化系统的演化模型[34]并于 1994 年定义了文化算法。他指出文化算法分别从微观和宏观两个不同层面分别模拟生物层面的进化和文化层面的进化，是一种多进化过程的算法并且各进化过程又相互影响相互促进。或者可以这样认为，文化算法是一个分层的双层进化系统，能够提供在两个不同进化层面（信仰空间和种群空间）上的交互和协作。两个空间有相应的通信协议即接受函数和影响函数来完成两个层面的通信，这样各个层面就形成统一的整体并能相互影响以提高算法的计算效率。

2.2.2 文化算法的框架

文化算法是一种解决复杂计算的新型全局搜索优化算法，它模拟了人类社会的演化过程。在人类社会中，文化以各种不同种类的形式作为信息的载体存在着，这些不同形式的信息通过各种方式潜在地影响社会中的群体成员，并通过人类自身的实践活动解决相应的问题。与其他进化算法有所区别，文化算法是一种基于知识的双层进化系统，其包含两个进化空间：一个是由具体个体组成的种群空间，通过进化操作和性能评价进行自身的迭代求解；另一个是由在进化过程中获取的经验和知识组成的信仰空间。图 2-1 说明了文化算法的基本框架[35]。

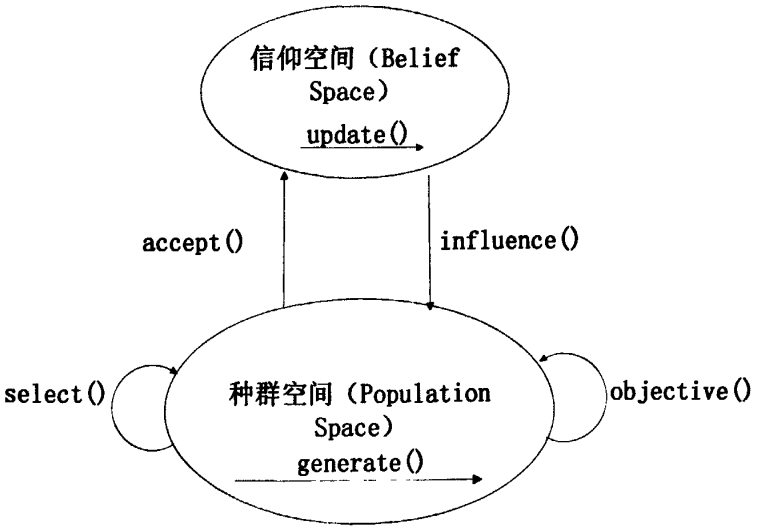

图 2-1 文化算法基本框架

如图 2-1 所示，种群空间与信仰空间分别从微观和宏观两个层面模拟了人类社会的进化过程，两个空间相对独立，都以各自的进化准则进行进化。首先，种群空间中的个体通过 objective()函数来计算其适应值，之后两个空间由相互通信的协议即接受函数 accept()和影响函数 influence()进行沟通。其中 accept()函数决定哪些个体作为影响信仰空间的个体，信仰空间通过 update()函数用从种群空间中选出的精英来更新信仰空间的知识，以形成新的知识。下一步信仰空间将影响种群空间的个体，在 influence()函数的作用下新的个体将会被生成（generate()）。这些新生成的个体和老一代个体一起通过 select()函数形成新一代的种群。这里的信息有两条反馈路径，一条是通过 accept()函数和 influence()函数，另一条是通过个体自身的经验以及 objective()函数。这两条反馈路径一起构成了种群空间和信仰空间的双重继承机制。种群空间和信仰空间相互交流并相互促进，类似于人类文化的进化机制。

文化算法的伪代码如下：

```
Begin
    t=0;
    Initialize   POPSpace(t):              //初始化种群空间
    Initialize   BLFSpace(t);              //初始化信仰空间
    repeat
            evaluate POP(t);               //评价种群个体
            update(BLFSpace (t)，accept(POPSpace (t))；  //更新信仰空间
            generate(POPSpace (t)，influence(BLFSpace (t))；//种群空间进化
            t=t+1;
            select POPSpace (t)   from POPSpace (t-1)
    until (termination condition achieved)
end
```

其中 POPSpace(t)表示第 t 代种群空间，BLFSpace(t)表示第 t 代信仰空间，算法从初始化两个空间开始，之后进入循环直到终止条件满足而结束。

文化算法框架是一种多层进化过程的计算模型，从计算模型的角度来看，任何一种符合文化算法要求的进化算法都可以嵌入文化算法框架，作为种群空间的一个进化过程。所以采用不同的进化算法作为文化算法的种群空间，就会有不同的文化算法。

根据上面的描述，总结文化算法具有以下特点：

(1) 双重进化继承：在种群空间和信仰空间分别继承父代的信息；

(2) 种群空间的进化是由信仰空间中保存的先进种群经验即知识来进行引导；

(3) 支持种群空间和信仰空间的层次结构；

(4) 两个空间分别进行自适应的进化；

(5) 不同空间的进化可以按不同的速度进行，也可以按不同的频率更新；

(6) 支持不同算法的混合问题求解；

(7) "文化"改变的不同模型可表达于一个模型之内。

2.2.3 文化算法的应用

Reynolds 和 Chung 等利用文化算法求解非线性约束优化问题[36]，他们指出在问题的约束变量变多时信仰空间将在问题求解过程中起到很重要的作用。Chung 和 Reynolds 将进化规划（Evolutionary Programming）和遗传算法数值优化系统 GENOCOP（GEnetic algorithm for Numerical Optimization for COnstrained Problems）结合起来组成文化算法的实验平台，这个平台能够组织各种不同的进化模型对数值约束问题进行求解。Chung 在对一些基准问题图形特征进行分类的研究中提出一种信仰空间中知识表示方法，他将信仰空间中的知识分成两种类型：形势知识（Situational Knowledge）和规范知识（Normative Knowledge）[37]，并将这两种知识用于指导搜索进程。

Jin 等提出了一种 n 维知识解模式，把其称为"信仰元"[38]（Belief-cell），这种模式能够提供一种精确地机制来支持非线性约束知识的获取、保存和整合。在文化算法框架中，信仰空间可以"容纳"一些模式集，这些模式集可以被用来指导进化种群的搜索，或者可以说基于模式的这些区域通过修剪不可行区域并促进可行区域的方式直接地用于指导优化搜索。另外 Jin 也将文化算法的框架用于数据挖掘[16]，他指出在多维空间优化搜索过程中，可以对有关的变量进行跟踪以保证搜索朝向最优点进行。Sternberg 将文化算法用于基于规则的专家系统中来对汽车保险公司保险索赔进行欺骗检测[39]，当专家系统中重组工程变得异常复杂时用文化算法对动态的变化做出响应，为专家系统提供一种自适应的响应能力。

Saleem 使用文化算法处理动态优化问题[40]，他指出文化算法提供一种动态环境中的推理机制，并通过研究表明文化算法在搜索多维空间中随时间变化的峰值时表现出令人振奋的结果。Ricardo 和 Carlos 将差分进化嵌入文化算法框架中作为其种群空间提出了一种文化差分进化算法[41]，并用于有约束优化问题中。

目前，国内对文化算法的研究也是方兴未艾，最早出杜琼等人在2005年在文献[42]中对文化算法作了介绍，之后本文首次提出基于进化策略的文化算法并将其作为Web话题识别的聚类算法。其中杨海英[17]等提出一种基于文化算法的负载均衡自适应机制，将文化算法应用到对服务器性能权值的进化计算中，通过对服务器的负载状况进行评价以"获得优化的性能权值"来指导资源的分配，使系统的使用趋于合理。张鹤峰[18]等人一种将Chan（一种定位基本算法）算法与文化算法相结合的算法，利用该算法解决TDOA（一种蜂窝网定位技术）定位估计中遇到的非线性最优化问题。吴英[19]等将文化算法应用到电力系统无功优化中，与改进遗传算法（SGA）和改进粒子群算法（CPSO）比较，得到了比较理想的结果。另外张涤[20]研究了基于文化算法的聚类分析技术，其中提到了Web数据的聚类。总体来看该算法的研究相对其它比较成熟的进化算法还是刚刚起步，应用的范围还比较小。

2.3 话题识别与跟踪

2.3.1 话题识别与跟踪相关定义

话题识别是TDT的一项子任务，TDT作为一项旨在帮助人们应对信息过载问题的研究，以新闻专线、广播、电视等媒体信息流为处理对象，将语言形式的信息流分割为不同的新闻报道并监控对新话题的报道，最后将涉及某个话题的报道组织起来以某种方式呈现给用户。它的研究目标是要实现按话题查找、组织和利用来自多种新闻媒体的多语言信息。这类新技术是现实中急需的，比如：自动监控各种信息源（如：广播、电视等，现在更多的是网络），从中检测出各种突发事件、新事件以及关于已知事件的新信息，它们可广泛用于网络舆情检测、信息安全、证券市场分析等领域。另外，还可以找出有关用户某一感兴趣话题的所有报道，研究这一话题的发展历程等。

为了区别于语言学上的概念，TDT对话题、报道、主题等概念进行了具体定义。

话题（Topic）：在最初的研究阶段（1999 年前），话题和事件的定义相同。一个话题指出某些原因、条件引起，发生在特定时间、地点并可能伴随某些必然结果的一个事件，比如"2009 年 12 月 7 日哥本哈根气候大会"。现在使用的话题定义相对宽泛一些，即一个话题由一个种子事件或活动以及与其直接相关的事件或活动组成。根据话题的定义，一篇报道只要论述的事件或活动与一个话题的种子事件有着直接的联系，那么这篇报道就与该话题相关。比如搜寻某次矿难事故的幸存者以及对矿难家属的慰问都被认为

是对该矿难事故直接相关的话题。

报道（Story）：指一个与话题紧密相关的、包含两个或多个独立陈述某个事件的子句的新闻片断。

主题（Subject）：它的含义更广些。话题与某个具体事件相关，而主题可以涵盖多个类似的具体事件或者根本不涉及任何具体事件。如"矿难事故"是一个主题，而"2009年2月22日山西焦煤集团屯兰矿特别重大瓦斯爆炸事故"则是一个话题。再比如，"自然灾害"是一个主题，而属于此主题类别的文本未必有与之直接相关的事件发生，如讲述自然灾害预防知识的文章。

2.3.2 话题识别与跟踪任务分类

TDT研究是一项综合的任务，需要比较多的自然语言处理理论以及相应的数据挖掘算法，根据不同的应用以及相应技术的完善，TDT评测会议把TDT分为五项任务[44]，这五项任务对新闻报道的处理各有侧重。下面分别对这五项任务的研究内容进行一下简单的描述：

(1) 报道切分（Story Segmentation）：即是从一个信息源获得的报道流（主要是流媒体数据：如电视，广播等）中切分出具有完整结构和统一主题的报道。如从一段包括娱乐、时事、体育新闻的广播中切分出属于不同领域的报道，该任务的主要切分对象是音频信号或誊录音频信息获得的文本语料，由于这类数据内部不同报道之间通常没有明确标记，需要通过报道切分来确定报道边界标记的位置。

(2) 话题关联识别（Story Link Detection）：主要任务是判断两篇新闻报道是否属于同一个话题，并对属于同一个话题的多个报道对按时间顺序处理。

(3) 首次报道识别（First Story Detection）：即是对报道流中的每个报道，判断其是否描述了一个新话题，即是否为该报道流中对一个话题的首个描述。首次报道识别是话题发现的第一步工作，被认为是TDT中最有难度的任务，因为其要用到其他相关的子任务作为其技术支持。

(4) 话题跟踪（Topic Tracking）：对于已经确定了的话题，对其后续报道进行过滤追踪研究，以判断舆论走势。该任务与信息过滤比较相似，区别在于话题追踪的需求描述和测试对象的时间效应比较强，随时间动态演化。

(5) 话题识别（Topic Detection）：也叫话题发现或话题检测，主要任务是检测和组织系统预先未知的话题，对报道流建立一个报道簇，簇内所有报道是对同一话题的描述。

它的特点是因为系统内欠缺相关话题的先验知识，实际是由首次报道识别和话题追踪两方面技术共同实现。该任务与文本聚类比较类似，但话题发现是实时聚类，新闻报道按时间顺序流入系统。

上述任务不是相互孤立的，从对他们的描述中可以看出其中一些任务是另一些任务的基础，如报道切分和关联识别是基础；而关联识别则是其中的核心技术，因为判断两个话题之间的关联程度都要用到这种技术；话题识别和话题跟踪则是研究的目的与重点。

单就针对话题识别方面，具体可以分为以下几个方面[45]：

(1) 在线话题识别（Online Topic Detection，OTD）：它的主要任务是识别新话题并收集后续相关报道。

(2) 新事件识别（New Event Detection，NED）：正如 TDT 研究体系中所提到的，首次报道识别（FSD）任务忽视了话题出现的不连贯性，从而使检测到的新话题经常是某些已知话题在不同时期出现的相关事件。因此，新事件识别（NED）逐渐成为辅助话题识别（TD）的重要组成部分。NED 与 FSD 任务很相似，唯一的区别在于前者提交的最新事件可能相关于历史上的某一话题；后者必须输出话题最早的相关报道。

(3) 事件回溯识别（Retrospective News Event Detection，RED）：RED 的主要任务是回顾过去所有发生过的新闻报道，并从中检测出未被识别到的相关新闻事件。对于 RED 研究方向的理解必须涉及到事件与话题的定义。事件是发生在特定时间和地点的事情，而话题则不仅包含作为种子的事件或活动，同时也包含与其直接相关的事件与活动。因此，RED 的任务实际上是辅助话题识别系统回顾整个新闻语料，从中检测相关于某一话题但以前却未被识别到的一类新闻事件。

(4) 层次话题识别（Hierarchical Topic Detection，HTD）：它是 TDT2004 定义的一项新的话题识别任务。HTD 是面向话题识别中两种不恰当的假设提出的，其中一个假设是所有报道与相关话题的近似程度都在一个层次上，而另一个假设是每篇报道只可能相关于一个话题。实际上，报道的主题与话题的相关程度往往分布于不同层次，比如“山西煤矿整合”和“人民币跨境贸易结算”两篇报道，虽然它们都相关于同一话题“2009 中国十大经济事件”，但是主题侧重点的差异造成它们与话题的对应程度处于不同层次。此外这两篇报道都可以分别划分到“矿业改革”类和“人民币国际化”类的话题模型当中，因此报道不总是仅仅相关于一个话题，往往不同话题的相关报道存在交集。

2.3.3 话题识别方法流程

Web 话题识别与跟踪主要方法和流程为：

(1) 网络新闻语料获取：首先确定将要研究的领域确定话题语料的采集源，如各大新闻门户网站或专业性的军事娱乐类网站等，之后选择设计一个网络爬虫用于定时采集指定时间、指定站点内的全部新闻网页，然后对下载下来的网页进行文本抽取，即是去除网页文档中的各种 HTML 标签以及无用的文字后，得到新闻标题、新闻正文、URL、来源网站、新闻时间等，以文本的方式进行存储。

(2) 语料预处理：对获取到的文本语料进行预处理，由于语料最终要在计算机中进行计算，所以要对文档进行形式化处理。需要确定选用一种文档表示模型（通常是向量空间模型 VSM），通过一定的特征项选择规则（如：TF/IDF）对文档中的特征项进行提取和权值表示，最终形成一个话题语料的向量表示。

(3) 聚类算法实现：在选取聚类算法之前，首先要确定一种相似度计算方法，如基于欧式距离的相似度计算，或基于向量夹角余弦的相似度计算方法，通常选用向量夹角余弦作为相似度（如下式所示）计算方法。最后再根据实际需要选取一种或几种聚类算法，对文档向量化后的数据进行聚类。

$$D(d_i,d_j)=\cos(d_i,d_j)=\frac{\overrightarrow{d_i},\overrightarrow{d_j}}{|\overrightarrow{d_i}|\times|\overrightarrow{d_j}|}=\frac{\sum_{k=1}^{n}w_{i,k}\times w_{j,k}}{\sqrt{\sum_{k=1}^{n}w_{i,k}^2}\times\sqrt{\sum_{k=1}^{n}w_{j,k}^2}} \tag{2-13}$$

其中 d_i,d_j 分别表示两篇文档的向量空间表示，$w_{i,k}$ 表示第 i 篇文档第 k 个向量的权值。

文档的聚类要根据不同的应用对聚类质量和效率的特定要求，来选用合适的聚类算法。一般 Web 话题识别中用到的主要聚类算法文档聚类算法可分为两类[46]：

(1) 层级式聚类算法：即聚类结果可以表示成树状结构，叶子节点表示初始的类，非叶子节点表示不同层次的聚类结果，根据聚类方向的不同又可以分为凝聚式（自底向上）和分裂式（自顶向下）两种。以凝聚式为例其典型算法 AHC（层次聚合聚类）算法的过程是首先假设所有文档自成一类，然后将最相似的两类合并，并继续这一过程，直到最后将所有文档合并为一类，形成一棵聚类树。

(2) 划分式聚类算法：划分式聚类不同于层级式聚类，其类别结构一般没有清晰的层次关系，算法通过不断的迭代来完成样本数据的最优分配，但其本质是一种贪心算法，

容易陷入一种局部最优解。这类算法的典型代表是 K-means 聚类算法、KNN 聚类算法等。K-means 聚类算法需要事先给定 K 值和初始划分，但其聚类效果就直接受到 K 值和初始划分的影响；KNN 即 K 最近邻算法，它的基本思想是如果一个文档与它最相近的 K 个文档都属于某一个类，那么该文档也属于该类。由于面向互联网新闻报道流的话题识别任务处理的数据量很大，因此 TDT 系统中经常使用的是 1NN 聚类算法，即 Single-pass 聚类算法，其实质是求与当前文档距离最近的一个类，在 Single-pass 聚类中一般用类的中心代表该类，而类的中心被定义为该类中所有文档的平均向量。

2.4 本章小结

本章在三个方面对本论文的基础知识作了论述，分别是进化算法、文化算法和话题识别技术。因为进化策略是进化算法的一种，先对进化算法进行了介绍，以其生物进化的过程为背景指出进化算法提出的进化依据，而后对其基本原理进行了描述，并对进化算法的三个主要分支遗传算法、进化规划、进化策略分别作了阐述，最后对进化算法的特点作了总结；第二节是文化算法的相关知识，在文化算法的提出、框架以及其在各方面优化问题中的应用以及文化算法的特点做了论述；第三节则是通过话题识别与跟踪中相关概念以及其任务分类和方法流程的阐述对话题识别与跟踪作了介绍。

第三章 进化策略文化算法设计

3.1 进化策略的重要算子

在进化策略文化算法研究中，种群空间的设计是首要研究内容，而进化策略作为该算法种群空间的进化过程，其中用到的算子（如变异算子和选择算子）的选取将直接影响算法的进化效果，本节将论述进化策略的重要算子，讨论各算子的选取对算法进化过程的影响。

3.1.1 变异算子

在 ES 中，变异算子通常作为基本的遗传操作算子，它是遗传变化的主要根源。变异算子的设计一般与问题有关，它的质量体现在两个方面：一个是产生随机数花费的代价，另一方面是突变产生的后代在一可行域上的分布是否合理。通用的变异算子设计方法学现在还没有被明确的建立起来，Beyer 在文献[47]中对几种比较成功的变异算子进行分析，从理论角度提出了在设计变异算子时一般要遵循的三条基本原则，即是可达性（Reachabihyt）、无偏性（Unbiasedness）、可伸缩性（Scalability）三原则。

可达性（Reachability）。所谓可达性指对于任一给定的父个体状态(y_p,s_p)，经过有限的变异或进化之后能达到任意的其他状态$(\tilde{y},\tilde{s})$，此处，把个体对应的基因看作是个体空间的状态。这一条件也是证明算法全局收敛性的必要条件。

无偏性（Unbiasedness）。这一要求的思想其实是源于达尔文进化论。我们可以看到在进化过程中，选择和变异是两个不同有时候甚至是有点矛盾的作用：选择通过对最佳适应值信息的选取以指导搜索过程进入可能存在更优解的搜索区域，而变异则不使用任何适应值信息，只能使用源于父种群的搜索空间信息用于对搜索空间的探索。因此在进化策略中，任何被选择作为父代的个体都是没有倾向性的，任何突变算子都不会引入偏差。对于给定的父个体要做到尽量无偏，这是设计变异算子的原则。这一思想自然会让我们想到“最大熵原理”，应用这一原理，可以指导进化策略在无约束的实数搜索空间R^n使用正态分布，而在整数搜索空间Z^n可以是几何分布。

可伸缩性（Scalability）。可伸缩性指变异步长或者平均变异长度都是可以调节的，

以适用于适应值函数的特性，这样对自适应要求的目的是为了确保进化策略系统（即是结合目标函数的进化策略算法）的“进化性”。这里用一种非正式化的方式把“进化性”直观地描述为这样一种思想，它使得进化代数如同是在适应度空间建立的一条“平滑的”随机进化路线，指导变异的生成。由于进化空间的性质由目标函数和变量算子决定，因此适应度空间的“平滑性”（并不一定是目标函数的）可以被认为是有效的进化最优化的前提条件。

另外需要在此说明的是，这三条原则仅仅是一般性的指导原则，当然在现实需求上根据所处理的问题不同，它们的重要性也会有所差别，与这一原则相冲突的变异算子在特定的问题中不一定失败。

变异算子根据实际应用不同可以分为实数搜索空间，二进制搜索空间和组合搜索空间的变异算子。这里将重点讨论实数搜索空间的变异算子。对于实数搜索空间 R^n，CES（Classical ES）的变异算子可表示为 $x'=x+z$。$z=(z_1,z_2,\cdots,z_n)$ 为 n 维随机生成的变量。CES 使用的随机变量满足 Gaussian 分布，目标变量的各个分量（基因）在同一个时间进行变异，变异的步长（标准差）σ 既可以作为个体本身的基因与个体目标变量基因一起进化，也可以作为外在参数形式以某种控制方式如 1/5 成功规则来决定，以随进化过程的深入而动态的改变变异步长。当标准差 σ 作为外在参数时，则所有分量的变异步长都相同。这种方式常称为自适应型步长（Adpative σ）控制策略，使用 Gaussian 分布的随机变量时：

$$z=\sigma(N_1(0,1),N_2(0,1),\cdots,N_n(0,1)) \tag{3-1}$$

$z_i=N_i(0,1)$, $i=1,2,\cdots,n$ 是互相独立的随机变量。$N(0,\sigma)$ 的概率密度函数（probability density function，pdf）为：

$$P_N(z)=\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2}\frac{z^2}{\sigma^2}) \tag{3-2}$$

如图 3-1 所示，变异步长 σ（标准差）控制了变异增量的分散程度。

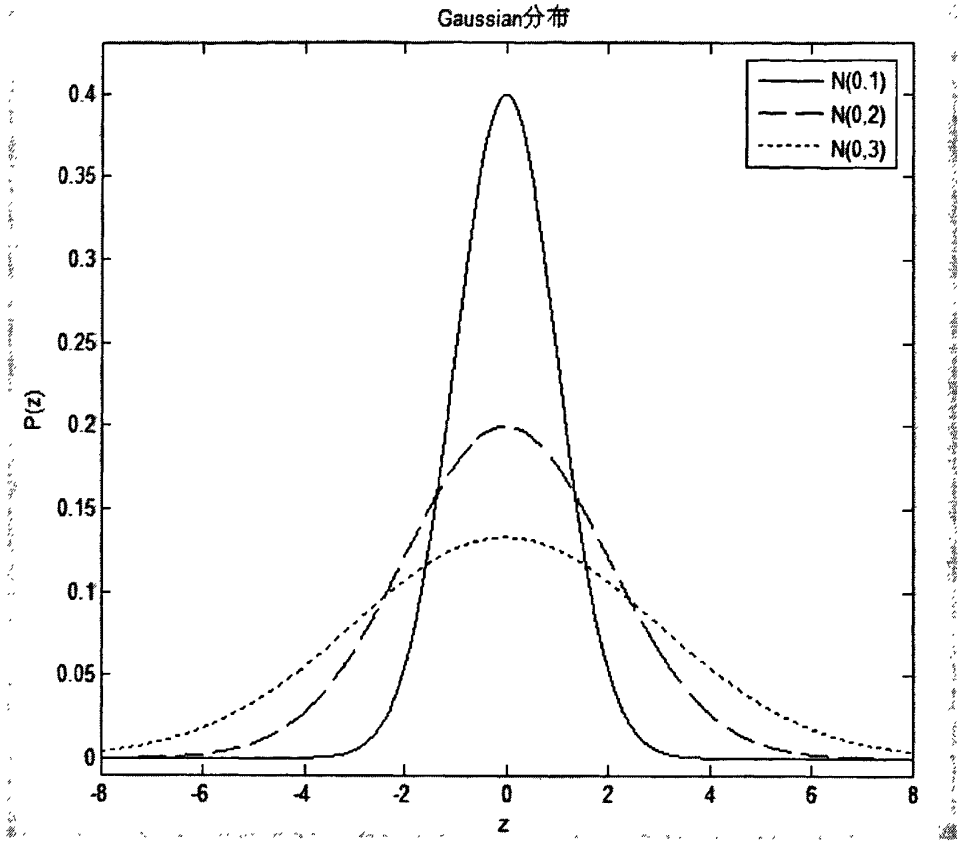

图 3-1 不同方差的正态分布图

如上图所示 Gaussian 分布的图像曲线比较集中在均值附近，这样的话在变异时就不利于跳出局部极值点。Yao.Xin[48]等根据 Cauchy 分布具有较强的分散性的特点提出了 Cauchy 变异，如图 3-2 所示，其概率密度函数具有较大的拖尾，那么在变异步长的随机取值时就更加离散，这样有利于提高算法的全局搜索能力。使用 Cauchy 分布时：

$$z=\sigma(C_1(0,1),C_2(0,1),\cdots,C_n(0,1)) \tag{3-3}$$

$z_i=C_i(0,\sigma)$ 是相互独立的 Cauchy 分布随机变量，$C(0,\sigma)$ 的概率密度函数为：

$$P_C(z)=\frac{1}{\pi}\frac{\sigma}{\sigma^2+z^2} \tag{3-4}$$

图 3-3 给出了 Cauchy 概率分布与 Gaussian 概率分布的比较图，从图中可以看出，Gaussian 正态分布比较集中于均值附近，而 Cauchy 分布相比来说就具有较强的发散性。

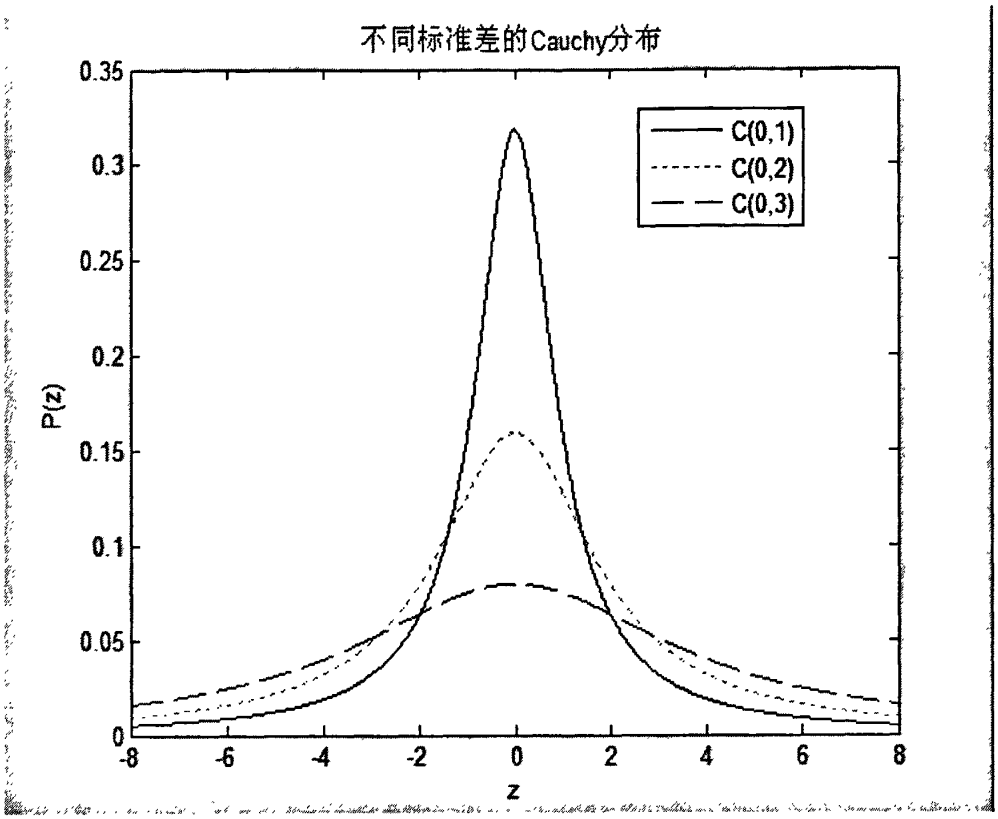

图 3-2 不同标准差的 Cauchy 概率分布

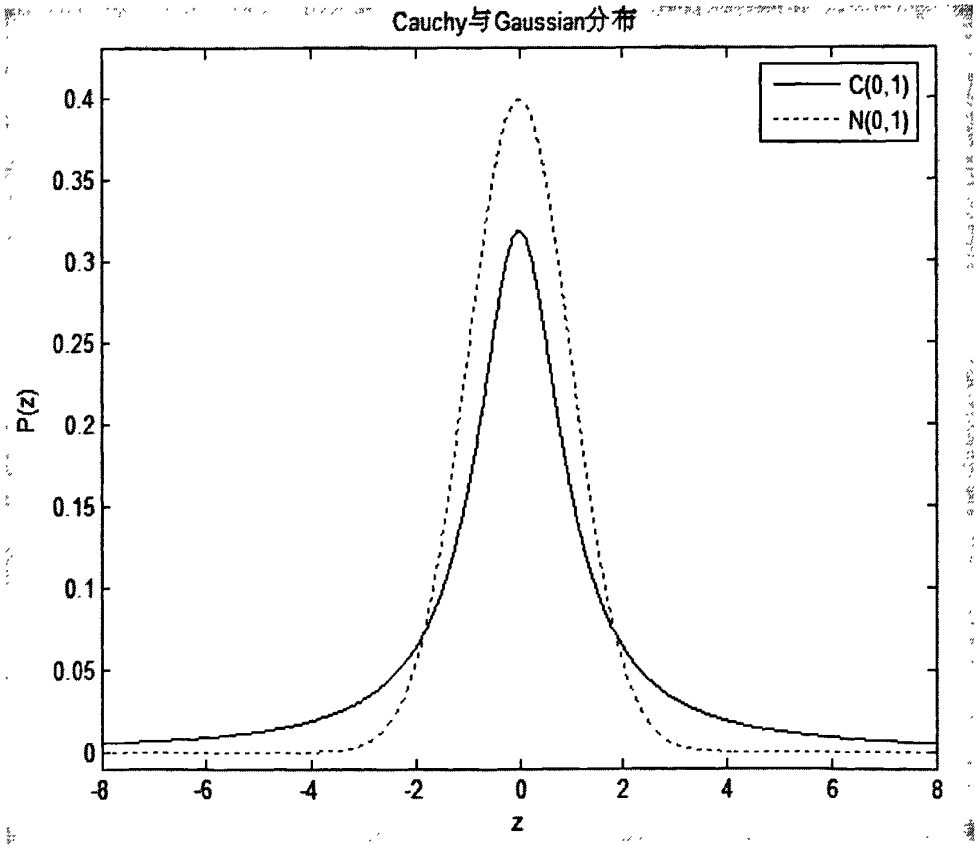

图 3-3 Cauchy 概率分布与 Gaussian 概率分布对比图

另外王云诚[49]还指出 Laplace 变异函数和指数分布函数同样可以作为进化策略的变异算子，并给出这些函数作为变异算子收敛性的证明。

林丹等人在文献[50]中应用概率论方法对各变异算子进行了详细的比较，比较了它们的局部搜索和局部逃逸性能。结果表明，Cauchy 变异算子和 Gaussian 变异算子分别具有良好的局部逃逸和局部搜索能力，而平均变异算子（Cauchy 变异算子和 Gaussian 变异算子的组合平均）在一维时同时具有良好的局部逃逸和局部搜索能力，但在高维时它的性态和 Cauchy 变异算子基本一样。由于上述变异方式都是一种全基因变异方式，即是在繁殖子代时，子代个体的每个目标变量基因都是从父个体基因变异产生，王湘中在文献[51]提出一种单基因变异的方式，并通过仿真计算试验表明对于同一变异步长，单基因变异的局部搜索能力强于全基因变异，能进化到更接近于极值点，验证了其理论分析的正确性。这些结果为设计和使用变异算子提供了指导和启发。

3.1.2 选择算子

选择机制（Selection）是进化策略的重要部分，它规定了进化的方向。进化策略中的选择机制是只有那些具有预计性质（如高适应度）的个体才有机会作为父代进行繁殖，这就如同生物的繁殖一样。不同的是在进化策略里，选择过程是确定性的。

选择策略包含两种技术：$(\mu+\lambda)-ES$，原有 μ 个个体及新产生的 λ 个新子代个体（共 $\mu+\lambda$ 个体）一起参与竞争并选取其中的 μ 个作为下一代群体的父个体，也被称为精英机制；另一种为 $(\mu,\lambda)-ES$，这种技术从新产生的 λ 个新子代个体中择优选择 μ 个个体作为下一代父代群体，这里要求 $(\mu<\lambda)$。而不考虑父代的适应度和子代相比好坏与否。显然这种选择机制是以出生过剩为基础的，又被称为遗忘机制。

在 $(\mu+\lambda)-ES$ 选择策略中，上一代的父代和子代一起加入下一代父代的选择中，这时 $(\mu=\lambda)$ 或者 $(\mu>\lambda)$ 都有可能，这种选择机制对子代数量没有限制，这样就最大程度地保留了那些具有最佳适应度的个体。但是这样也会带来一些缺点，如它可能会增加计算量而影响收敛速度。

在 $(\mu,\lambda)-ES$ 选择策略中，只有最新产生的子代才能加入选择机制中。所以在优化选择中要求 $(\mu<\lambda)$ 是必须的。从 λ 中选择出最好的 μ 个个体，作为下一代的父代，而 $(\lambda-\mu)$ 个个体被放弃，这种选择也被称为切断选择。显然它不能保障优秀父代个体被保留。

3.2 种群空间

任何基于种群进化的生物智能进化算法都可适用于文化算法的种群空间。随着智能计算领域的研究进展，遗传算法、遗传规划、进化规划、粒子群优化算法以及微分进化算法等诸多智能计算策略均被引入文化算法作为其种群空间。

本文决定应用 ES 而不是 GA 作为文化算法的种群空间主要是基于如下两点考虑：其一是 ES 是实数直接编码的，目标向量的长度仅随着待优化变量数目的增长而线性地增长。如果采用常规的 GA 二进制编码，目标向量的长度会随着个体复杂度的增加而达到让人难以忍受的程度。其二是 ES 处理约束问题比较容易，可以保证该算法能够搜索更大的参数空间。但是这并不意味着进化策略优于遗传算法或者相反，正如 NFL（No Free Lunch）定理指出的那样，无法设计出一种算法在解决所有优化问题时都优于其他算法，而只能认为该算法在此领域占有优势。比如若进化算法求解问题集 A 时的性能比模拟退火的性能好，那么必然会有模拟退火求解问题集 B 时的性能比进化算法的性能好。

3.3 信仰空间

信仰空间最早被 Chung 分为形势知识（Situation Knowledge）和规范知识（Normative Knowledge）两个部分，后来 Jin 又提出了拓扑知识（Topographical Knowledge）和约束知识（Constraint Knowledge），Saleem 又在这两人的基础上提出了历史知识（History knowledge）和领域知识（Domain Knowledge）。不同知识类型在不同方面刻画了种群中优良个体的不同特性。可根据所要解决问题的不同选择相应不同的知识类型，当然不同的知识种类在引导进化搜索的过程中所起到的作用也不尽相同。下面对其中最为基本的三种知识：形势知识、规范知识和拓扑知识作探讨。

3.3.1 形势知识

形势知识是 Chung 于 1997 年在解决静态环境实值函数优化问题时提出的[52]，用于记录进化过程中的较优个体。其结构描述为

$$S=\{s_1^t,s_2^t,\cdots,s_m^t\} \tag{3-5}$$

其中，m 为形势知识容量，即为最优个体集合。$s_i^t=\{x_i^t|f(x_i^t)\}$ 为第 t 代第 i 个较优个体，$f(x_i^t)$ 为 x_i^t 的适应值。

形势知识中记录的较优个体按照个体适应值降序排列，即满足

$f(x_{i-1}^t)>f(x_i^t), i\le m$。种群空间每代进化完成后，接受函数选取较优个体提交给信仰空间，知识更新函数从中选出最优个体，用于更新形势知识，其更新过程描述为：

$$s^{t+1}=\begin{cases}x_{best}^t & f(x_{best}^t)<f(s^t)\\ s^t & \text{其他}\end{cases} \tag{3-6}$$

其中，x_{best}^t是种群空间第 t 代最优个体。可见，形势知识是进化过程中具有优势引导作用的个体轨线的反映。

3.3.2 规范知识

规范知识由 Chung 提出[52]，用于描述问题的可行解空间。针对具有 n 维变量的优化问题，规范知识结构描述为

$$N=<N_1,N_2,\cdots,N_n> \tag{3-7}$$

n 为变量数，$N_j=<I_j,L_j,U_j>$表示每个变量的取值区间信息，每个取值区间 $I=[l,u]=\{x|l\le x\le u,x\in R\}$表示变量$x$定义域边界的值，上下界$u$和$l$由给定的值域初始化，$L_j$表示参数$j$区间下限$l_j$对应的目标函数的适应值，$U_j$是参数区间上限$u_j$对应的目标函数的适应值。

规范知识更新体现为可行搜索空间的变化。随着进化深入，搜索范围应在优势区域集中涵盖。因此，当存在较优个体超出当前搜索范围时，更新规范知识的更新规则如式（3-8）~（3-11）所示。

$$l_i^{t+1}=\begin{cases}x_{j,i} & if\ x_{j,i}\le l_i^t\ or\ f(\vec{x}_j)<L_i^t\\ l_i^t & \text{其他}\end{cases} \tag{3-8}$$

$$L_i^{t+1}=\begin{cases}f(\vec{x}_j) & if\ x_{j,i}\le l_i^t\ or\ f(\vec{x}_j)<L_i^t\\ L_i^t & \text{其他}\end{cases} \tag{3-9}$$

$$u_i^{t+1}=\begin{cases}x_{k,i} & if\ x_{k,i}\ge u_i^t\ or\ f(\vec{x}_k)<U_i^t\\ u_i^t & \text{其他}\end{cases} \tag{3-10}$$

$$U_i^{t+1}=\begin{cases}f(\vec{x}_k) & if\ x_{k,i}\ge u_i^t\ or\ f(\vec{x}_j)<U_i^t\\ U_i^t & \text{其他}\end{cases} \tag{3-11}$$

这里，第j个个体影响参数i的下边界，第k个个体影响参数i的区间上边界，l_i^t表示参数i第t次迭代时的下边界，L_i^t表示l_i^t对应目标函数的适应值。u_i^t表示参数i在第t代时的上边界，U_i^t表示u_i^t对应目标函数的适应值。

3.3.3 拓扑知识

拓扑知识也称作地势或地形知识，最早由Jin Xidong提出的信仰元概念进化而来，即将知识进行划分为以元为单位的等级[53]。拓扑知识以规范知识为基础，将可行搜索空间均分成许多称为单元（cells）的小区域。各单元与可行搜索空间具有相同变量维数，即在各维上同时进行变量划分。以2维变量优化问题为例，若各维变量可行搜索空间为[-10，10]，每一维划分成4个单元数目，则其拓扑知识如图3-4所示。图中，各单元依次编号，各单元状态通过单元属性值加以描述。

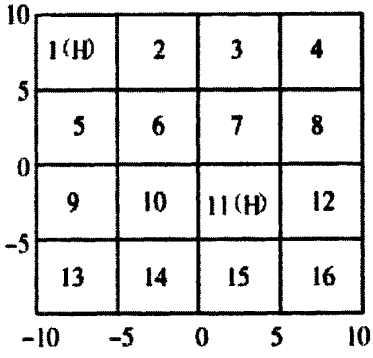

图3-4 拓扑知识（二维搜索空间）

单元属性值C根据单元所在区域的个体适应度平均值占种群平均适应度的比例来分配，所占比例越高等级越高。属性值一般分为4个等级，即：$C\in${高（HI），中（ME），低（LO），未知（OT）}。其中，该单元所在搜索区域没有被当前种群覆盖称为“未知”。上述属性值的占优关系为：HI>ME>OT>LO。文献[54]针对约束优化问题进一步将拓扑知识与约束条件相结合，通过判断单元可行性将其划分为：可行域（F）、半可行域（S）、不可行域（N）。其中，半可行域是指约束条件未被完全覆盖的单元，通常处于约束范围的边界区域，如图3-5所示。拓扑知识引导种群在可行和半可行单元进行搜索，并且避免对不可行单元的搜索。拓扑知识更新与规范知识有关，一般包含2种情况：（1）规范知识更新后，需重新对拓扑知识进行划分；（2）规范知识如果未更新，则根据单元中个体数目的多少和单元中个体的平均适应值更新拓扑知识。

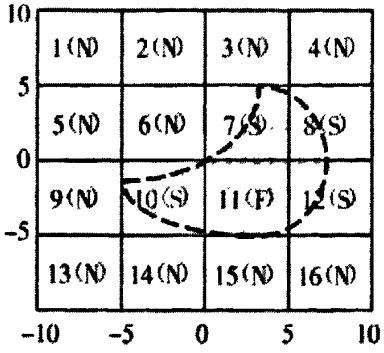

图 3-5 约束优化问题的拓扑知识（二维搜索空间）

拓扑知识反映了搜索空间中个体适应度分布状况。它在进化过程中，引导处于低属性单元的个体向较高属性单元移动。同时拓扑知识也反映了搜索空间中个体适应度分布状况，属性高的单元个体适应度也高。

另外，还有领域知识和历史知识等信仰空间的知识，其中领域知识用于引导种群沿预测的优势方向进化。该类知识在动态优化问题中可以有效捕捉环境的动态变化信息，提高进化效率。历史知识相当于进化的时序表。当搜索陷入局部最优或环境发生变化时，它引导搜索返回之前记录的可行解空间，重新开始进化。

3.4 影响函数和接受函数构造

3.4.1 影响函数

影响函数的主要作用是信仰空间中的各类知识通过该函数引导种群空间的种群进化。不同进化阶段，信仰空间中的各类知识所起作用不同；当然选用的知识类型不同，对种群进化的引导作用也不同。在进化初期，种群探索整个搜索空间，为了尽快向优势区域移动，此时规范知识及拓扑知识处于控制地位，一方面限定探索区域，另一方面引导搜索趋于具有高属性值的单元。在进化中期，更新形势知识和规范知识，进一步缩小搜索范围，并对拓扑知识进行更近一步的划分，引导种群在更小粒度上进行搜索。进化后期，搜索集中在某一局部区域，容易陷入局部最优而导致早熟收敛，这时开始引入历史知识以使种群跳出局部较优点。在整个的进化过程中，领域知识一直为种群进化提供进化方向的指引。从知识对进化过程的引导作用可以看出，影响函数的关键问题在于各类知识何时作用于种群，及其所引导的种群在整个种群空间中所占的比例。

依据信仰空间的知识构成，影响函数分为两类[55]：（1）单类知识的影响函数，即信仰空间只含有一类知识，知识通过影响进化算子引导整个种群进化。（2）m 类知识的影

响函数。这时信仰空间存在多类知识，但各类知识作用时机和影响种群比例不是一次性决定的，一般采用随机型和轮盘赌型两种方法来确定用哪一种知识。随机性影响函数即是每次根据随机产生的随机数而选用一种信仰空间的知识作为影响种群空间的知识。这种方式每次只选用一类知识作用于种群，且作用比例为整个种群，并未充分发挥多类知识的综合引导作用；轮盘赌型影响函数是借鉴遗传算法中的轮盘赌选择机制，确定何类知识引导种群及其作用比例。假设第i类知识在赌盘上所占的区域大小为：

$$\gamma_i = \frac{\omega_i}{\sum_{i=1}^{m} \omega_i} \tag{3-12}$$

其中，ω_i是第 i 类知识的表现值。在初始化阶段$\omega_i = 1/m$，即各类知识对种群具有相同的影响程度。随着进化深入，各类知识对种群的影响程度动态变化，描述为

$$\omega_i = \frac{\sum_{j=1}^{k} f(x_j)}{k} \tag{3-13}$$

其中，k是第i类知识影响的个体数。ω_i越大，该类知识对种群的影响程度越大，其在随后进化中所影响种群比例也越大。可见，轮盘赌型影响函数每次综合使用多类知识作用于种群，且能根据各类知识对种群的引导能力来动态修正其影响个体数目。

以求解非线性无约束优化问题为例，其影响函数主要有如下四种：

CA-Versionl：仅使用规范知识调整变量变化步长。具体定义如下：

$$x'_{j,i} = x_{j,i} + size(I_i) \cdot N(0,1) \tag{3-14}$$

其中，$N(0,1)$为服从标准正态分布的随机数，$size(I_i)$为信仰空间中变量可调整区间的长度。

CA-Version2：仅使用形势知识调整变量前进方向。具体定义如下：

$$x'_{j,i} = \begin{cases} x_{j,i} + \left|\sigma_{j,i} \cdot N(0,1)\right| & if\ x_{j,i} < s_i \\ x_{j,i} - \left|\sigma_{j,i} \cdot N(0,1)\right| & if\ x_{j,i} > s_i \\ x_{j,i} + \sigma_{j,i} \cdot N(0,1) & \text{其他} \end{cases} \tag{3-15}$$

这里$\sigma_{j,i}$是指第j个个体中第i个变量的个体级变化步长，s_i是在信仰空间中变量i的最佳值。

CA-Version3：使用规范知识调整变量变化步长，形势知识调整其前进方向。具体定义如下：

$$x'_{j,i}=\begin{cases}x_{j,i}+\left|size(I_i)\cdot N(0,1)\right| & if\ x_{j,i}<s_i\\ x_{j,i}-\left|size(I_i)\cdot N(0,1)\right| & if\ x_{j,i}>s_i\\ x_{j,i}+size(I_i)\cdot N(0,1) & 其他\end{cases} \tag{3-16}$$

CA-Version4：仅使用规范知识调整变量变化步长及前进方向。这样是为了父代处在较好区间时变异仅做微小摄动，而在其它情况时则使父代的变异尽可能朝向信仰空间中规范知识所引导的区间。具体定义如下：

$$x'_{j,i}=\begin{cases}x_{j,i}+\left|size(I_i)\cdot N(0,1)\right| & if\ x_{j,i}<l_i\\ x_{j,i}-\left|size(I_i)\cdot N(0,1)\right| & if\ x_{j,i}>u_i\\ x_{j,i}+\beta\cdot size(I_i)\cdot N(0,1) & 其他\end{cases} \tag{3-17}$$

此时，l_i 和 u_i 分别是当前信仰空间中参数 i 的下限和参数 j 的上限。

3.4.2 接受函数

接受函数从种群空间选取较优个体，提交给信仰空间用于知识更新。其研究核心在于选取较优个体数目。目前，已有的接受函数有三种：固定比例接受函数、动态接受函数和模糊接受函数。

(1) 固定比例接受函数

顾名思义，固定比例该接受函数在整个进化过程中，以一个固定比例（p%）提取种群空间中的较优个体，即

$$accept()=p\% \tag{3-18}$$

对种群规模为 N 的种群选取种群中前 $p\%\cdot N$ 个较优个体。选取比例一般在 20%左右，即 $p\%=20\%$。对于执行接受函数的间隔代数则一般是预先设定。前面的章节已经讨论过，进化前期种群多样性较好，为避免早熟收敛和误导并尽快找到优势区域，不宜选取过多个体进入信仰空间。伴随着进化一步步深入，较优个体隐含着更多的有价值信息，这时应该提交更多个体给信仰空间。而到了进化后期，算法的逐渐收敛使得优势个体及其携带的隐含有效信息之间相似性越来越大，为保持知识的多样性并避免冗余信息对进化迭代的消耗，应减少接受个体数目。因此，固定比例接受函数不能满足上面的要求而需要用动态比例接受函数来调节。

(2) 动态接受函数

该接受函数引入进化代数作为动态调节因子，调节接受比例，即

$$accept() = p\% + \frac{p\%}{t} \tag{3-19}$$

其中，p%为一个预先设定的固定比例；t为进化代数。由上式可知较优个体的选取比例随着进化代数的增加而逐渐减小，变化范围为$[p\%, 2p\%]$。该动态接受函数计算简单，但进化代数不能直接反映当前进化状况。为此，Chung 将模糊逻辑引入接受函数[56]。

(3) 模糊接受函数

该接受函数引入一个新的概念"个体成功率"，所谓个体成功率 ß，是指子代个体中优于父代个体的比例，即

$$\beta = a/N \tag{3-20}$$

其中，a为子代个体优于父代个体的数目。它和进化代数结合起来一起作为接受比例的影响因素，从而能更加全面地反映当前进化状况。模糊接受函数采用模糊推理策略，以个体成功率和进化代数作为输入，较优个体接受比例作为输出。其模糊规则表，如表 3-1 所示。

表 3-1 模糊接受函数规则表

进化代数	个体成功率		
	LO	ME	HI
I	ME	ME	HI
M	LO	ME	ME
F	LO	LO	ME

可见，模糊接受函数考虑了进化代数和个体成功率两种因素，能根据当前进化状态确定较合理的个体接受比例，但计算实施起来比较复杂，同时凭经验确定的模糊推理中的隶属度函数，容易对算法性能造成较大影响。总之，上述三类接受函数各有特点。在实际应用中，应根据具体优化问题进行选择。

3.5 进化策略文化算法描述与实验

3.5.1 算法描述

本节设计一种基于进化策略的文化算法，并用于求解非线性无约束优化问题，其算

法描述如下：

(1) 初始化种群空间。随机选取 p 个候选解组成初始种群空间，每一个个体为一个 N 维实数向量。

(2) 应用适应度函数即目标函数对种群空间中的个体进行评价，求出个体的适应值。

(3) 根据给定的取值范围和初始种群空间中的候选解，按照 3.3 节给出的信仰空间形势知识和规范知识结构，生成初始信仰空间。

(4) 根据信仰空间的知识，利用 3.4 节的影响函数 influence()，对种群空间中的 p 个父个体进行变异，生成 p 个相应子个体。

(5) 评价 p 个子个体的适应度，这里同样是依据适应度函数。

(6) 对于由父子两代个体共同组成的规模为 2p 的种群空间中的每个个体，随机地从该种群空间中选取 c 个个体与之进行比较。记录每个个体与 c 个竞争者中的获胜次数(即该个体比其竞争对象的适应度高)。

(7) 选择具有最多胜利次数的前 p 个个体作为下一代的父个体。

(8) 用 3.4.2 的动态接受函数 accept()更新信仰空间，信仰空间的更新规则按照式（3-6）和式（3-8～3-11）来更新。

(9) 如果不满足终止条件，则重复（4），反之，则结束。

3.5.2 仿真实验

该仿真实验设计的目的是利用下面具有代表性的无约束非线性函数 $f_1 \sim f_4$ 来对前面所提的四种不同影响函数对应版本的文化算法进行寻优性能测试。本文所用到的非线性无约束优化函数如下：

$$f_1(x)=\sum_{i=1}^{n}x_i^2, x_i \in [-5.12, 5.12] \tag{3-21}$$

$$f_2(x)=100(x_1^2-x_2)^2+(1-x_1)^2, x_i \in [-2.048, 2.048] \tag{3-22}$$

$$f_3(x)=-\cos(x_1)\cos(x_2)e^{-(x_1-\pi)^2-(x_2-\pi)^2}, x_i \in [-100, 100] \tag{3-23}$$

$$f_4(x)=\sum_{i=1}^{n}ix_i^2, x_i \in [-1.28, 1.28] \tag{3-24}$$

为实现对比效果，上述四种算法与自适应进化规划算法[57]（Epsa）一起进行比较实验。表 3-2 描述了五种算法对上述测试函数进行仿真实验的参数配置，其中第一列为各

个函数，第二列为函数变量的维数，第三列为理论最优值或当前已知最优值(用“*”表示$f(x^*)$)，第四列为最大进化代数，第五列初始种群规模以及第六列竞争规模。每个实验都使用初始种群相同，分别对每个测试函数独立运行 20 次。

表 3-2 函数参数配置表

函数	n	$f(x^*)$	最大进化代数	种群规模	竞争规模
f_1	30	0	400	30	30
f_2	2	0	400	30	30
f_3	2	-1	180	10	10
f_4	30	0	500	30	30

五种算法对各个函数的测试结果列入表 3-3 中。表中统计了各种算法的成功率(与已知最优值的误差小于 1.0e-6 即为成功)和测试函数 20 次运行后的平均值。

表 3-3 寻优结果对照表

函数	EPsa 成功率 最优值	CA-Version1 成功率 最优值	CA-Version2 成功率 最优值	CA-Version3 成功率 最优值	CA-Version4 成功率 最优值
f_1	5% 2.1821e+00	15% 1.251564e+02	40% 3.454125e-02	25% 0.794939e+02	100% 2.547513e-35
f_2	35% 4.873022e+00	100% 2.406967e-09	100% 1.171077e-06	85% 5.028988e-31	100% 1.407213e-05
f_3	80% -8.000000e-01	0% -1.052054e-01	100% -1.000000e+00	100% -1.000000e+00	100% -1.000000e+00
f_4	0% 2.142903e-01	0% 9.495947e+01	90% 5.633432e-07	0% 1.653847e+01	100% 1.704331e-70

从结果中可以看出，新提出的算法在大部分情况下还是有效可行的，并且可以看出用规范知识调整变量变化步长及前进方向的影响函数能起到更好的效果，另外使用形势知识调整变量前进方向也有较好的表现，其他两种方法则表现一般。

3.6 本章小结

本章首先对进化策略的重要算子进行了详细的分析和探讨，指出进化策略中变异算

子设计的三条原则，以及常用的变异算子及各种变异算子在不同情况下的比较；而后又介绍了选择算子的一些情况。从第二节开始便是文化算法中种群空间和信仰空间的设计，以及影响函数和接受函数在其中的表现形式。根据以上各种函数的介绍在第五节提出一种基于进化策略的文化算法，并运用于非线性函数的优化测试中，取得了令人满意的结果。

第四章 话题聚类设计与评价

4.1 话题向量空间

Web 话题识别中的聚类技术是在话题文本的形式化表示基础上进行的，因此需要先对 Web 话题文本进行形式化处理，以形成话题向量空间模型。形式化处理过程主要包括三个步骤：网页信息采集、文本预处理以及特征选择。

4.1.1 网页信息采集

网页信息采集的目的即是从网上把感兴趣的新闻或评论下载到本地，这其中首先要确定来源信息的网站，即有一个 URL 处理器负责对待采集网站的 URL 进行管理，去除重复的 URL 并简单判断网页是否重复，把重复的网页也去除掉。由网页采集器通过各种协议来完成数据的采集，对于采集到的页面，通过网页去重检测后，需要由 URL 提取器分析其中的链接，并对链接进行必要的转换以获取真实的 URL。对下载下来的网页还需要由标签信息提取器对网页内容信息进行分析，提取页面的 Meta 信息、作者信息、页面的标题、页面的摘要等。其系统框图如下图所示。

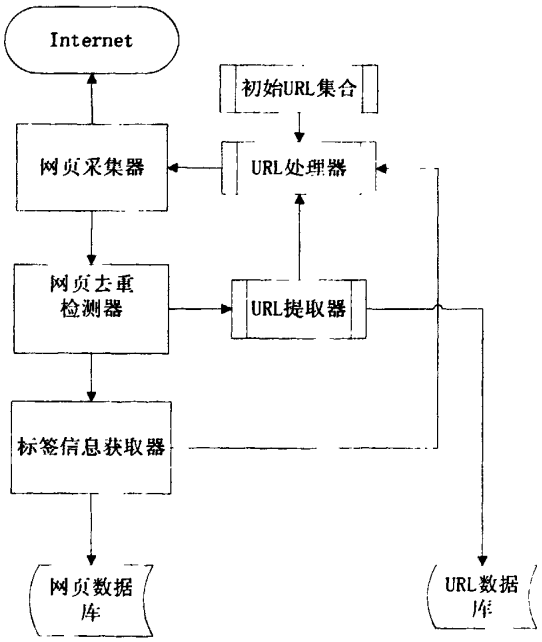

图 4-1 Web 信息采集系统框图

4.1.2 文本预处理

这个阶段对下载下来的 Web 网页进行文本处理，包括分词以及停用词的处理。所谓分词是针对中文文档而言的，因为我们知道英文文章中的词与词之间有空格作为划分，单词之间是相互独立的，而中文文章中词与词之间是紧密相连的，它们之间除了标点符号之外不存在明显的划分，因此必须对中文文本进行分词。

对文本文档进行分词过后需要对文档中的停用词进行处理，所谓停用词（Stopwords）即是指一些太常用以至没有任何检索价值或对文章标识意义不大的单词，如英文中常见的“a”、“the”、“and”、“of”等，中文中的“的”、“啊”、“阿”、“哎”等。这些词没有多大意义，并且在各类文本中出现的频率都很高，不能体现文本所表示的内容，如果在聚类过程中的特征选择或者计算相似度时考虑到这些词，那么文本之间的相似性不能表现出内容的相似性，而且这种相似性也没有什么意义，所以必须对停用词进行处理。

另外还可以根据词汇的频数把一些词频过低或过高的词汇给剔除，以尽可能的降低用于特征表示的词汇数量，方便表示和计算。

4.1.3 特征抽取

文本经过预处理后词汇数量依然是很大的，这样表示成文本向量空间的维数也就相当大，高的可以达到上千维以上，需要对文本向量进行降维处理。这样做主要基于以下两点考虑：第一，向量空间的压缩可以提高程序的效率，并提高运行速度；第二，不同词汇对文本特征向量的表示价值是不同的，一些常用的、各个类别中出现频率都比较高的词汇对文本特征表示的贡献比较小，而在某特定领域中出现频率多却在其他领域中很少出现的词汇对文本特征表示的贡献就比较大，为了提高特征表示的精度，对于每一类，应去除那些表现力不强的词汇，并尽可能保留那些表现力强的特征，以筛选出针对该类的特征项集合。

目前筛选特征项的算法比较多，但是由于大部分的筛选特征项的算法如互信息（Mutual Information）、信息增益（Information Gain）、期望交叉熵（Expected Cross Entropy）等，都是基于这样的一个前提：已知文本的分类信息。但话题识别中用到的是聚类技术，一般不会事先知道文本的分类信息，因此本文决定采用经典的词频逆文档频率(Term Frequency-Inverse Document Frequency，TF-IDF)[58]作为文档向量表示的权值计算方法，TF-IDF 的具体公式如下：

$$w_i = TF(t_i) \times IDF(t_i) = TF(t_i) \times \log(\frac{N}{DF(t_i)} + 1) \quad (4\text{-}1)$$

其中 w_i 表示某个特征项 t_i 的权值，$TF(t_i)$ 为特征项 t_i 的词频，$IDF(t_i)$ 表示特征项 t_i 的逆文档频率，其计算公式为 $IDF(t_i) = \log(\frac{N}{DF(t_i)} + 1)$，N 是文档集中文档的总数，$DF(t_i)$ 为含有特征项 t_i 的文档个数。

通过以上三方面的文本处理，并对计算出的特征项的权值根据一定的规则进行筛选，最终形成一个文档向量空间模型（Vector Space Model，VSM），其结构为

$$D = D(w_1, w_2, \cdots, w_n) \quad (4\text{-}2)$$

其中 w_i 即为式（4-1）计算出的结果，n 为特征项的总数。

两文档向量之间的相似度计算采用余弦相似度计算，公式如下：

$$CosSim(D_i, D_j) = \frac{D_i \bullet D_j}{\|D_i\| \times \|D_j\|} = \frac{\sum_{k=1}^{n} d_{ki} \times d_{kj}}{\sqrt{\sum_{k=1}^{n} d_{ki}^2} \times \sqrt{\sum_{k=1}^{n} d_{kj}^2}} \quad (4\text{-}3)$$

D_i, D_j 分别为第 i, j 篇文档的向量化表示，d_{ki}, d_{kj} 分别为第 i, j 篇文档向量的第 k 个分量。

4.2 话题种群与信仰空间的生成

4.2.1 种群空间编码

在话题文本向量空间确定之后，则需要设计适合该话题聚类的种群空间以及相应的信仰空间。由于种群空间采用进化策略作为其进化方式，那么其种群编码方式采用的也就是进化策略的种群编码方式。在进化聚类算法中，一般有两种类型染色体编码方式：一种是基于聚类中心的实数编码，每条染色体由 k 个聚类中心组成 $C = \{C_1, C_2, \cdots, C_k\}$，对于 n 维的样本向量 $C_i = (C_{i1}, C_{i2}, \cdots, C_{in})$，其染色体为长度是 $k*n$ 的浮点码串，即把 k 个聚类中心排成一列 $(C_{11}, \cdots, C_{1n}, C_{21}, \cdots, C_{2n}, \cdots, C_{k1}, \cdots, C_{kn})$；另一种是基于聚类划分的整数编码，即每个染色体如 $(k_1, \cdots, k_m)$，其中 $k_i \in \{1, \cdots, k\}$，表示总共 m 个样本，第 i 个分量为 k_i 表示该样本属于第 k_i 类。

由于聚类问题的样本数目一般都远大于其聚类数目，采用聚类划分的整数编码方式

染色体长度太大，而采用聚类中心的编码方式更为有效简洁。所以进化策略中采用聚类中心的实数编码方式。

聚类准则函数如（4-4）式所示：

$$\min \quad J=\sum_{i=1}^{k}J_i=\sum_{i=1}^{k}\sum_{j=1}^{n}\left\|D_j-C_i\right\|^2 \tag{4-4}$$

$$C_i=\frac{1}{n_i}\sum_{l=1}^{n_i}X_i \tag{4-5}$$

其中D_j为第j个文档的向量表示，C_i为第i个聚类中心，k为聚类数目。C_i由公式（4-5）确定n_i为第i个类中所含的样本数。由于种群空间的适应度一般是求最大化问题，则在进化策略中可以用聚类准则函数的倒数即

$$f=1/J \tag{4-6}$$

作为种群空间的适应度函数对各代的适应度进行量化计算。

4.2.2 信仰空间构成及更新

信仰空间结构采用三类知识构成，定义为$<Norm,Situ,Topo>$，$Norm$是定义域子空间区域的 n 个参数的规范化知识，其结构和更新方式如 3.3.2 节所示，$Situ$为信仰空间中的形势知识，其结构和更新方式如 3.3.1 节所示。$Topo$是搜索空间里的拓扑知识，它把搜索空间根据适应值情况分成四个区域，每个区域有一个状态称为区域状态（Region-Status），其取值范围为{高（HI），中（ME），低（LO），未知（OT）}。区域状态是基于本区域的平均适应值AVR_i与当代形势知识存储的个体的平均适应值AVR来确定，定义如式(4-7)所示：

$$Region-Status=\begin{cases}HI & AVR_i>0.7AVR\\ ME & 0.4AVR<AVR_i\le 0.7AVR\\ LO & 0<AVR_i\le 0.4AVR\\ OT & 0\end{cases} \tag{4-7}$$

其中$AVR_i=\sum_{j=1}^{k}f_j(x_i)\Big/k$为第$i$个区域的平均适应度值，$f_j(x_i)$为第$i$个区域第$j$个个体的适应度值，$k$为处于第$i$个区域内的个体数目。$AVR=\sum_{j=1}^{A}f_w(x)\Big/A$为当前代数较优搜索范围内种群的平均适应值，$f_w(x)$为当代形势知识存储的第$w$个个体的适应度值，$A$

为规范化知识的规模。平均适应值越高代表着其区域状态等级愈高，这就表明拓扑知识能够将种群向适应度较高即包含具有较好聚类效果个体的区域移动。

对于拓扑知识的更新，通过轮盘式选择法来实现个体替代。如果区域状态为 HI 或 ME 的区域平均适应度值满足 $AVR_i \in [L_i^t, U_i^t]$ 时，随机在该区域内产生一个个体替换区域状态为 LO 区域内的最差个体。如果 $AVR_i \notin [L_i^t, U_i^t]$，定义区域选择概率 $\rho_k = \frac{AVR_k}{AVR}$，将区域状态为 HI 或 ME 的区域选择概率 ρ_k 与 (0, 1) 之间的一个随机数 η 比较，如果 $\rho_k > \eta$，随机选取规范化知识中存储的任一个体替换该区域内的最差个体。

4.3 话题聚类设计与描述

综上所述，在形成文本向量空间的基础上，以及对种群空间的编码讨论前提下提出基于进化策略文化算法的 K-means 混合聚类算法（K-Means Evolution Strategies based Cultural Algorithm 简称 KESCA），本算法仍采用 K-means 模型为聚类模型，只是在文化算法种群空间采用进化策略的进化过程。

混合算法以文化算法为框架，种群空间为进化策略的种群空间，编码方式采用 4.2 节的种群空间编码方式，信仰空间为 4.2 节讨论的信仰空间结构，接受函数用动态接受函数（式 3-19），影响函数为 3.4 节论述的影响函数。算法的具体步骤如下：

(1) 初始化种群空间：从所有向量样本中随机选取 k 个作为聚类中心进行编码，产生种群空间中的一个个体。同样的步骤重复 p 次，即可生成初始种群规模为 p 的种群空间；

(2) 对种群空间中的每个个体进行适应度评价，适应度函数为式(4-6)；

(3) 根据样本集和初始种群空间中的候选解，按照信仰空间结构，生成初始信仰空间；

(4) 根据 3.4.1 节影响函数 influence()版本 CA-Version1（或版本 CA-Version2），对种群空间中的每个父个体进行变异，对应生成 p 个子个体；

(5) 对于出父子两代个体共同组成的规模为 2p 的种群空间中每个个体，随机地从中选取 c 个个体，把它们与种群空间中的每个个体 α_k 都按适应值进行比较，找出其中比种群空间差的个体数目 w_k，并把 w_k 作为个体 α_k 的获胜次数；

(6) 选择前 p 个具有胜利次数最多的个体作为下一代的父个体；

(7) 执行接受操作，使用 3.4.2 节中动态接受函数即式（3-18）选择影响信仰空间的个体。

(8) 对信仰空间进行操作，根据接受函数调整信仰空间，更新形势知识、规范化知识和地形知识。

(9) 如果不满足终止条件，则重复步骤(4)；反之，则结束。

4.4 聚类性能评价

NIST 为 TDT 建立了完整的评测体系，但是由于各个研究方面针对不同的问题，而且历届评测语料之间标注方案的差异，TDT 不同任务之间的评测方法、参数和步骤不尽相同。总体而言，一般包括五个指标对 TDT 的评测结果进行比较，这五个指标有：漏检率（Miss rate，简写为 Miss）、误检率（false alarm rate，简写为 FA）、查准率（precision）、查全率（recall）和 F 值（F-measure）等。其中话题 i（在这里表示某一个聚类 i）的漏检率和误检率[7]定义为：

$$Miss_i = \frac{\text{未检测到的与话题i相关的报道数}}{\text{与话题i相关的报道总数}} \tag{4-8}$$

$$FA_i = \frac{\text{检测到的与话题i不相关的报道数}}{\text{与话题i相关的报道总数}} \tag{4-9}$$

则整个话题聚类结果的平均漏检率和误检率为：

$$P_{Miss} = \sum_i Miss / t_n \tag{4-10}$$

$$P_{FA} = \sum_i FA / t_n \tag{4-11}$$

其中 t_n 为话题聚类个数，在本文也即是总聚类数目。

查准率和查全率[59]以及 F 值[4]一般作为数据挖掘中聚类算法研究的评测标准，它的定义组合了信息检索中查准率(precision)与查全率(recall)的思想，将每个聚类结果看作是查询的结果，这样对于最终的某一个聚类类别 r 和原来的预定类别 i，

$$recall(i,r) = n(i,r)/n_i \tag{4-12}$$

$$precision(i,r) = n(i,r)/n_r \tag{4-13}$$

这里 $n(i,r)$ 是聚类 r 中包含于预定义类别 i 中的文档的个数，n_r 是聚类形成的类别中文档个数，n_i 是预定义类别中的文档个数。则聚类 r 和预定义类别 i 之间的 $f(i,r)$ 值计算如下：

$$f(i,r)=\frac{2\cdot recall(i,r)\cdot precision(i,r)}{recall(i,r)+precision(i,r)} \tag{4-14}$$

一次聚类结果最终的评价函数为

$$F=\sum_i \frac{n_i}{n}\max\{f(i,r)\} \tag{4-15}$$

这里n是所有测试文档的个数。值得指出的是，通过以上这两种方法获得的聚类评价只是对数据集作一次划分的评价。为了客观评价聚类算法的性能，有必要进行多次聚类获得其评价结果，并用其均值来评价算法。即：

$$F_c=\frac{1}{m}F_i \tag{4-16}$$

m表示进行了m次评价，F_i为第i次求出的F值。

上述几个概念并不是毫无关系的，假如某次聚类的话题 i 的结果如表 4-1 所示：

表 4-1 话题聚类结果

	话题相关（人工）	话题无关（人工）
话题相关（机器）	a	b
话题无关（机器）	c	d

其中漏检率$Miss_i=c/(a+c)$，误检率$FA_i=b/(b+d)$，查全率$recall_i=a/(a+c)$，查准率$precision_i=a/(a+b)$，显然$Miss_i+recall_i=1$。

4.5 实验与结果分析

实验初始语料集选用搜狐搜狗实验室提供的 2008 版搜狐新闻数据(SogouCS)[60]，本文使用了其中的样例数据，数据为去除 HTML 格式后的 Web 新闻文本。我们选取了其中的财经、健康、汽车、军事、体育、社会五类 Web 新闻文档，在每一类文档中随机选取 200 篇，共 1200 篇作为总的语料集，人工标注话题 15 个。

对选取的 Web 新闻文本用中科院计算所的 ICTCLAS 分词工具进行分词，并统计各个词的词频，根据词频以及确定的阈值筛选出特征词，共计 132 个特征词。并计算其 TF-IDF 值作为其权重，形成文本的向量表示。

分别用新设计的基于进化策略文化算法的 K-means 聚类算法以及传统的 K-means 聚类算法对语料集进行聚类实验，并根据所得结果整理数据如下：

表 4-2 算法 recall、precision 和 F-measure 结果

	recall		precision		F-measure	
	K-means	KESCA	K-means	KESCA	K-means	KESCA
财经	0.7356	0.7485	0.7023	0.7106	0.7186	0.7291
健康	0.6534	0.6612	0.6418	0.6502	0.6475	0.6557
汽车	0.7612	0.7835	0.7326	0.7542	0.7466	0.7686
军事	0.7015	0.7234	0.6879	0.7068	0.6946	0.7150
体育	0.6852	0.7112	0.6546	0.6903	0.6696	0.7006
社会	0.5324	0.5213	0.5004	0.5201	0.5159	0.5207

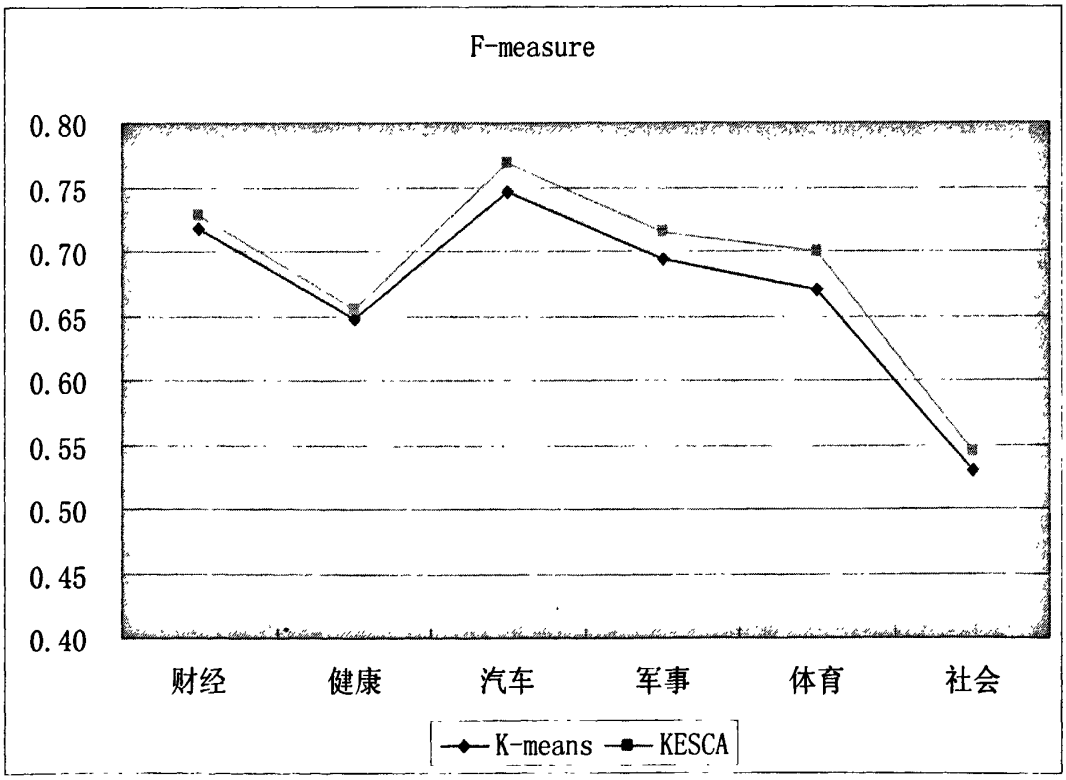

图 4-2 不同类别的 F-measure

算法误检率和漏检率如下表所示：

表 4-3 算法误检率和漏检率

	Miss		FA(false alarm)	
	K-means	KESCA	K-means	KESCA
财经	0.2644	0.2515	0.1054	0.0987
健康	0.3466	0.3388	0.1826	0.1245
汽车	0.2388	0.2165	0.0912	0.0862
军事	0.2985	0.2766	0.2815	0.2613
体育	0.3148	0.2888	0.2013	0.1465
社会	0.4376	0.4279	0.4513	0.4032

两种算法在各类别中的漏检率和误检率对比效果如下：

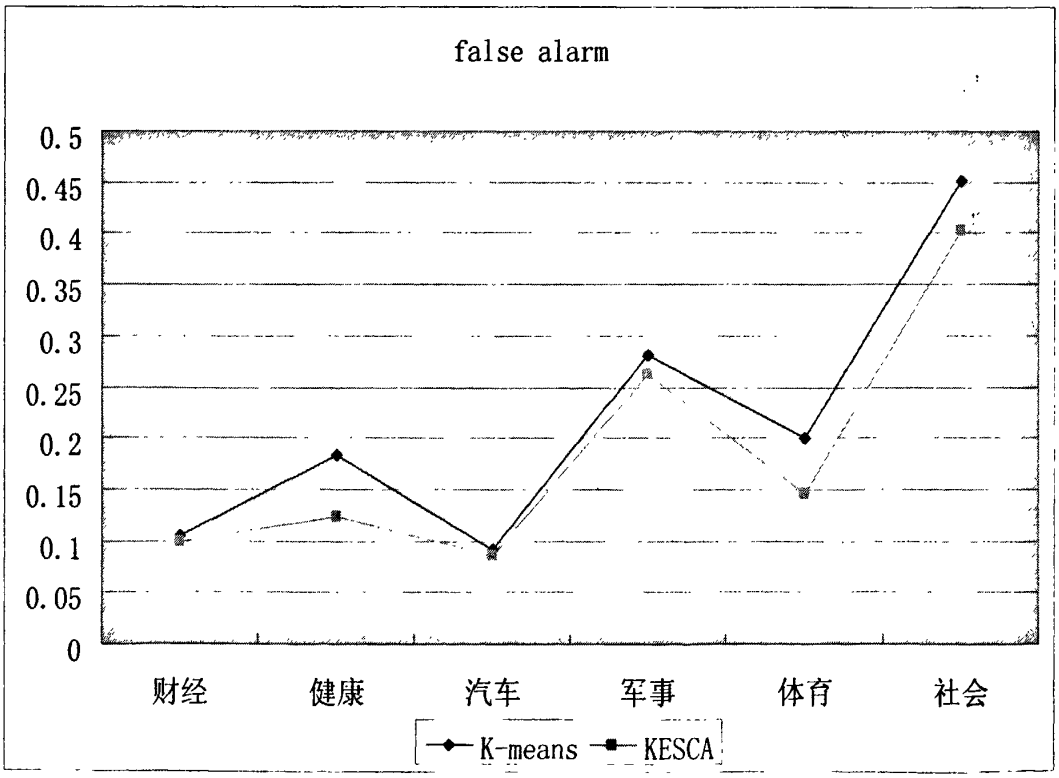

图 4-3 不同类别的误检率比较

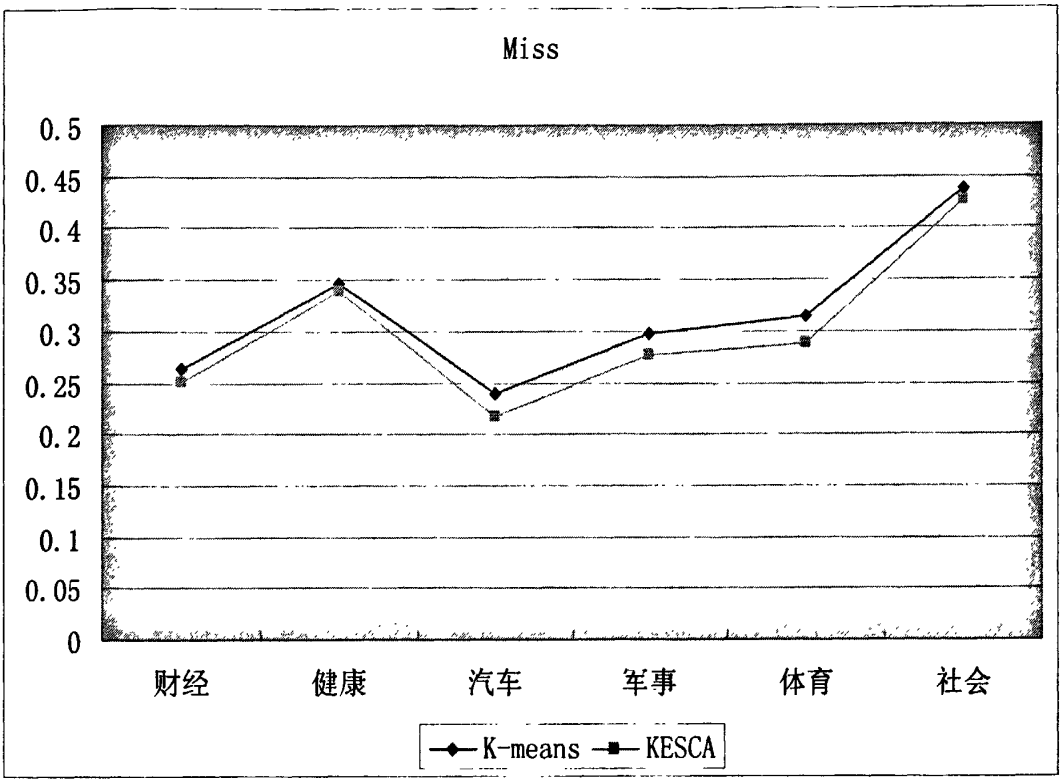

图 4-4 不同类别的漏检率比较

通过上述实验可以看出，新提出的基于进化策略文化算法的 K-means 聚类算法总体来看具有比较可行的效果，但从图 4-2 看出，汽车类的聚类结果好于其他类的结果，而社会类的结果则较差，这说明聚类结果跟具体聚类对象的特征选择有比较大的关系，寻找其中的原因应该是在特征提取时在汽车军事等领域有比较容易区分的特征词，而对于社会领域其特征词的提取就不太容易。

4.6 本章小结

本章针对新提出的进化策略文化算法在话题识别中的应用作了探讨，根据话题识别的聚类算法要求设计了文化算法中的种群空间和信仰空间等相关技术，通过选取一定的语料对话题文本的聚类做实验，对结果进行了评价并证明新算法在话题识别中的应用是可行有效的。

第五章 总结与展望

5.1 论文工作总结

话题识别旨在从海量的新闻报道中识别出未知的话题以及与已知话题相关的报道。绝大多数识别算法都是基于聚类算法进行的，通过采用向量空间模型描述新闻报道和话题文本，并确定一种相似度计算方法及聚类准则函数，按照某种聚类策略对其进行聚类。

话题识别中用到的聚类算法是话题识别的核心技术，本文即以用于话题识别的聚类算法研究为目的，主要做了如下方面的工作：

对进化算法进行详细的讨论，分别阐述了进化算法的三个主要分支遗传算法、进化规划、进化策略，通过各算法的对比确定选用进化策略作为文化算法的种群空间，并对进化策略中的重要算子进行详细的分析，对进化策略的进化机制有一个清晰的掌握。

依据文化算法的框架分别对文化算法的种群空间、信仰空间以及这两个空间中的通信协议即影响函数和接受函数进行了深入的讨论，指出各种函数的工作机制，之后把进化策略嵌入文化算法作为其种群空间，设计基于进化策略的文化算法。

针对新提出的进化策略文化算法在话题识别中的应用作探讨，根据话题识别的聚类算法要求设计文化算法中的种群空间和信仰空间等相关技术，通过选取一定的语料对话题文本的聚类做实验，对结果进行了评价并证明新算法在话题识别应用中的可行有效性。

5.2 不足之处及进一步工作设想

论文研究和撰写过程中，本人阅读调研了大量的国内外文献，对本文的基础理论有了较为详细的了解，提出了一种切实可行的聚类算法，但还存在很多有待改进的地方，其不足和改进之处如下：

(1) 在第三章对新设计的基于进化策略的文化算法用于求解非线性无约束优化问题时，得到规范知识和形势知识能够更好地引导种群向更优方向的结论还显得有些不够权威，应根据各具体问题再做大量的实验，如更多的测试函数等对所设计的算法进行验证，才有可能得到更有说服力的结论。

(2) 话题识别用到的聚类技术分为两种，一种是硬聚类技术，即一个样本只能属于

一个类别，另一种是软聚类技术，即一个样本可以属于不同类别。本文的话题识别用的是硬聚类技术，而现实问题存在一个事件可能同时属于某几个不同的话题，属于软聚类的问题。因此应该对聚类算法做更好的研究，如考虑使用模糊聚类方法并设计出一种新的聚类算法，以更好地满足现实问题的要求。

(3) 针对聚类算法中 K 值的确定问题，也是需要研究的一个方面，本文是在事先知道分类情况的前提下给出 K 值，这在一定程度上不能够更好满足现实中的应用，因此需要对聚类时的 K 值确定有更深入的研究，以提出一个能自适应确定 K 值的算法来满足现实的需要。

(4) 在话题文本的形式化表示时，文本词汇特征词的权值计算采用的是 TF-IDF 权重计算方法，但话题识别要求时序性比较强，这种计算法在动态时序性要求方面不够完美，可以采用更好的权重计算方法。

综上所述，由于本人学识积累有限和知识水平的制约，以及时间的限制，论文中难免会出现不够完善的地方，甚至有可能还存在一些错误，恳请各位读者批评指正。

参考文献

[1] 中国互联网络信息中心，中国互联网络发展状况统计报告[EB/OL]. http://www.cnnic.cn/html/Dir/2010/01/15/5767.htm. 2010.01.

[2] J Allan, J Carbonell, G Doddington. Topic Detection and Tracking Pilot Study: Finalreport[A] In:Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop[C], Virginia:Lansdowne, February 1998, 194-218.

[3] 骆卫华，刘群，程学旗. 话题检测与跟踪技术的发展与研究[C]. 全国第七届计算语言学联合学术会议论文集，北京：清华大学出版社，2003：560-566.

[4] M. Steinbach, G. Karypis, V. Kumar. A Comparison of Document Clustering Techniques [C]. KDD Workshop on Text Mining(2000), Boston: MA, 2000: 109-111.

[5] B. Choudhary, P. Bhattacharyya. Text clustering using semantics[C]. In: Proceedings of the 11th International World Wide Web Conference, 2002.

[6] Young-Woo Seo and Katia Sycara. Text Clustering for Topic Detection[R]. Pittsburgh: Robotics Institute, Carnegie Mellon University, 2004.

[7] 杨建武. 文本挖掘技术[EB/OL]. http://www.icst.pku.edu.cn/course/TextMining/07-08Spring/index.html. 2008-6-12.

[8] 贾自艳，何清，张海俊等. 一种基于动态进化模型的事件探测和追踪算法[J]. 计算机究与发展. 2004, 41(7): 4721-9721.

[9] 税仪冬，瞿有利，黄厚宽. 周期分类和 Single-Pass 聚类相结合的话题识别与跟踪方法[J]. 北京交通大学学报. 2009, 33(5): 86-89.

[10] 赵华，赵铁军，赵霞. 时间信息在话题检测中的应用研究[J]. 计算机科学. 2008, 35(1): 221-223.

[11] R.G. Reynolds. An introduction to cultural algorithms, Proceedings of the Third Annual Conference on Evolutionary Programming[C], February 24-26, 1994, San Diego, California, 1994:131-13.

[12] R.G. Reynolds. Learning the Parameters for a Gradient-based Approach to Image Segmentation Using Cultural Algorithms[C]. Proceedings International Symposium on Intelligence inNeural and Biological Systems. Herndon, Virginia:[s.n.], 1995:240-247.

[13] Rychtyckyj N. Using Cultural Algorithms to Re-engineer Semantic Networks[D]. Detroit, Michigan : Wayne State University, 2001.

[14] Rychtyckyj N, Reynolds R G. Using Cultural Algorithms to Improve Performance in Semantic Networks[C]/ / Proceedin gs of 1999 IEEE Congress on Evolutionary Computation. Washington D.C.:[s.n.], 1999:1651-1656.

[15] Saleem S M. Knowledge-Based Solution to Dynamic Optimization Problems using Cultural Algorithms[D]. Detroit, Michigan : Wayne State Universit y, 2001.

[16] JIN X, R.G. Reynolds. Data mining using cultural algorithms and regional Schemata[J]. In:Proc. of the 14th IEEE Intl. Conf.on Tools with Artificial Intelligence, 2002:33-40.

[17] 杨海英，黄皓，窦全胜. 基于文化算法的负载均衡自适应机制[J]. 计算机工程与应用，2005，21：146-149.

[18] 张鹤峰，杨莘元，刘婷. 基于文化算法的蜂窝网定位技术研究[J]. 应用科技，2008，35（11）：39-42.

[19] 吴英，金从友，马利山等. 基于文化算法的电力系统无功优化研究[J]. 现代电力，2008，25（3）：36-41.

[20] 张涤. 基于文化算法的聚类分析研究[D]. 成都：西南交通大学，2008.

[21] 孟凡荣，郭晶，周勇. 基于文化算法的模糊聚类分析[J]. 微电子学与计算机，2009，26（10）：1-4.

[22] 刘纯青，杨莘元，张颖. 基于文化算法的聚类分析[J]. 计算机应用，2006，26（12）：2953-5960.

[23] John H. Holland, Adaption in natural and artificial systems[M], Englewood Cliffs, The University of Michigan Press, 1975, 66–72.

[24] Kennedy.J, Ebethart R.C. Particle swarm optimization[C]. Proceeding of IEEE International Conference on Neural Networks, 1995, 1942-1948.

[25] Hopfield J.J, Tank D.W. "Neuarl" computation of decisions in optimization Problems[M] Biologieal Cybemtics, Springer 1985, 52:141-152.

[26] Goldberg D E. Genetic Algorithms in Seareh, Optimization and Machine Learning[M]. Reading, MA: Addison-Wesley Professional, 1989.

[27] Whitley D. The Genitor algorithm and selection pressure: why rank-based allocation of repreduetive trials is best[C] In:Proceedings of the Third International Conference on Genetic Algorithms and Their Appliections. Schaffer J D ed. 1989.

[28] Fogel L J.Owens A J, Walsh M J. Artificial Intelligence through Simulated Evolution[M]. NewYork: John wiley, 1966.

[29] Back T., Schwefel H-P. An overview of evolutionary algorithms for parameter optimization[J]. Evolutionary computation, 1993.1:1-24.

[30] 陈妍. 进化策略算法与应用的研究[D]. 长春：吉林大学，2008.

[31] Durham W. Co-evolution:genes, culture and human diversity[M]. Stanford, CA:StanfordUniversity Press, 1994:100-123.

[32] 王义祥. 发展社会学[M]. 上海：华东师范大学出版社，2004：35-50.

[33] Renfrew A C.. Dynamic modeling in archaeology:what, when, and where?Dynamical Modeling and

the Study of Chang in Archaeology[M]. In: S.E. vander Leeuw, ed. Edinburgh University Press, 1994

[34] D.B.Fogel:An Introduction to Simulated Evolutionary Optimization[J], IEEE Transactions on Neural Networks, 1994, Vol.5, No.1, 3-14.

[35] R.G. Reynolds. On modeling the evolution of hunter-gatherer decision-making systems[J], Geographical Analysis, 1978, 10(1):31-46.

[36] Chung C, Reynolds R. G. A testbed for solving optimization problems using cultural algorithms[C]. In Lawrence J. Fogel, Peter J. Angeline and Thomas BacK editors, Evolutionary Programming V: Proceedings of the Fifth Annual Conference on Evolutionary Programming, MIT Press, Cambridge, Massachusetts, 1996

[37] CHUNG C.. Knowledge-based approaches to self-adaptation in cultural algorithms[D] Wayne State University, Detroit, Michigan, May 1997.

[38] JIN X, R.G. Reynolds. Using knowledge-based evolutionary computation to solve nonlinear constraint optimization problems:a cultural algorithm approach[C]. In 1999 Congress on Evolutionary Computation, Washington, D C, IEEE Service Center, 1999:1672-1678.

[39] Sternberg M, R.G. Reynolds. Using cultural algorithms to support reengineering of rule-based expert systems in Dnamic performance environments[J]. A case study in fraud detection. IEEE Transactions on Evolutionary Computation, 1997, 1 (4) :225-243.

[40] S.M.Saleem, R.G. Reynolds. Cultural Algorithms in Dynamic Environments [C], Congress on CA, 2000:1513-1521.

[41] Ricardo Landa Becerra, Carlos A. Coello Coello. Culturizing differential evolution for constrained optimization[C]. In Proceedings of the Fifth Mexican International Conference. Computer Science, 2004:304-311.

[42] 杜琼，周一届. 新的进化算法——文化算法[J]. 计算机科学，2005，32（9）：142-144.

[43] Durham W.Co-Evolution: Genes, Culture and Human Diversity[M]. Stanford:Stanford University Press, 1994.

[44] Hong Yu, Zhang Yu, Liu Ting, et al. Topic detection and tracking review[J]. Journal of Chinese Information Processing, 2007, 21 (6): 71-87.

[45] 洪宇，张宇，刘挺等. 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)，71-85.

[46] 刘星星. 热点事件发现及事件内容特征自动抽取研究[D]. 武汉：华中师范大学. 2009.

[47] Hans-Georg, Beyer and Hans-Paul Schwefel, Evolution Strategies:A comprehensive introduction[J], Natural Computing, 2002, 1, 18-19.

[48] X.Yao and Y.Liu, Fast evolution strategies In Evolutionary Programming VI: Proceedings of the

Sixth Annual Conference on Evolutionary Programming（EP97）Lecture Notes in Computer Science[C] vol. 1213，Springer，Berlin（1997）:151–161.

[49] 王云诚，方伟武．进化规划与进化策略的变异算子[J]，运筹学学报．2008，12（1）：83-92.

[50] 林丹，李敏强，寇纪淞．进化规划和进化策略中变异算子的若干研究[J]，天津大学学报，2000，33（5）：627-630.

[51] 王湘中，进化策略的变异算子与仿真平台研究[D]．湖南：中南大学，2005.

[52] Chung C J.Knowledge -based approaches to self-adaptation in cultural algorithms[D]. USA:Wayne State University，1997.

[53] Xidong J，R.G. Reynolds. Using knowledge-based evolutionary computation to solve nonlinear constraint optimization problems：a cultural algorithm approach[C]//IEEE Congress on Evolutionary Computation，1999：1672-1678.

[54] J Xidong，R.G. Reynolds. Using knowledge-based system with hierarchical architecture to guide the search of Evolutionary computation[C]//The 11th IEEE International Conference on Tools with Artificial Intelligence，1999：29-36.

[55] 郭一楠，王辉．文化算法研究综述[J]，计算机工程与应用，2009，49（5）：41-45.

[56] R.G. Reynolds，Chung C J. Fuzzy approaches to acquiring experimental knowledge in cultural algorithms[C]//The 9th IEEE International Conference on Tools with Artificial Intelligence，1997：260-267.

[57] N.Saravanan and D.B.Fogel. Leaming strategy Parameters in evolution programing:an empirical study[C]，In Proceedings of the 3rd Annual Conference on Evolutionary Programing，1994：269-280.

[58] 薛晓性，张永全，任晓东．基于新闻要素的新事件检测方法研究[J]，计算机应用，2008，28 卷，11 期：1-2.

[59] 刘远超，王晓龙，徐志明等．文档聚类综述[J]，中文信息学报，2005，20（3），55-62.

[60] 搜狗实验室，搜狐网络新闻数据[EB/OL]．http://www.sogou.com/labs/dl/cs.html.

致 谢

在论文即将完成之际，回首论文撰写过程中的点点滴滴，我的心情无法平静，当然这一路走来的成果，不只是我一个人努力的结果，这期间有很多老师和同学都给予我以莫大的帮助。

首先要感谢的是我的导师王晓东教授，无论在论文的选题、开题，还是在研究和撰写过程中，王老师多次询问研究进程，并为我指点迷津，帮助我开拓研究思路，他的精心点拨和热忱鼓励都对我论文的写作有很大的帮助。不光是在论文撰写的过程中，在我读研的三年来，王老师也以他严谨的治学作风、诚恳正直的处事原则深深地影响着我，这对我树立正确的治学观念和价值观念都有着不可磨灭的作用。生活上王老师同样给予我无微不至的关怀和帮助，使我终身难忘。在学位论文完成之际，谨向导师表示我最崇高的敬意和衷心的谢意。

同时我还要衷心感谢我的其他任课老师，正是他们对科学研究的严谨态度感染了我，使我能够不断钻研；感谢我们实验室所有同学对我的支持和帮助；特别感谢我的家人，他们给予我的殷切期望和鼓励，是我永远奋斗的动力。

同时感谢各位专家在百忙之中对此文的审阅和赐教！

攻读学位期间的科研成果

[1] Fan Lilin，Zhang Lei，Luo Leiming. The Application Research of the Automobile Industry Supply Chain Based on ASP Platform[C]. Liu Rongfang，Zhang Jin，Guan Changqian. Logistics: The Emerging Frontiers of Transportation and Development in China. America：American Society of Civil Engineers，2008(vol.3)：1831-1836.

独 创 性 声 明

本人郑重声明：所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写的研究成果，也不包含为获得河南师范大学或其他教育机构的学位或证书所使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

作者签名：罗磊明 日期：2010.5.27

关于论文使用授权的说明

本人完全了解河南师范大学有关保留、使用学位论文的规定，即：有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权河南师范大学可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。（保密的学位论文在解密后适用本授权书）

作者签名：罗磊明 导师签名： 日期：2010.5.27