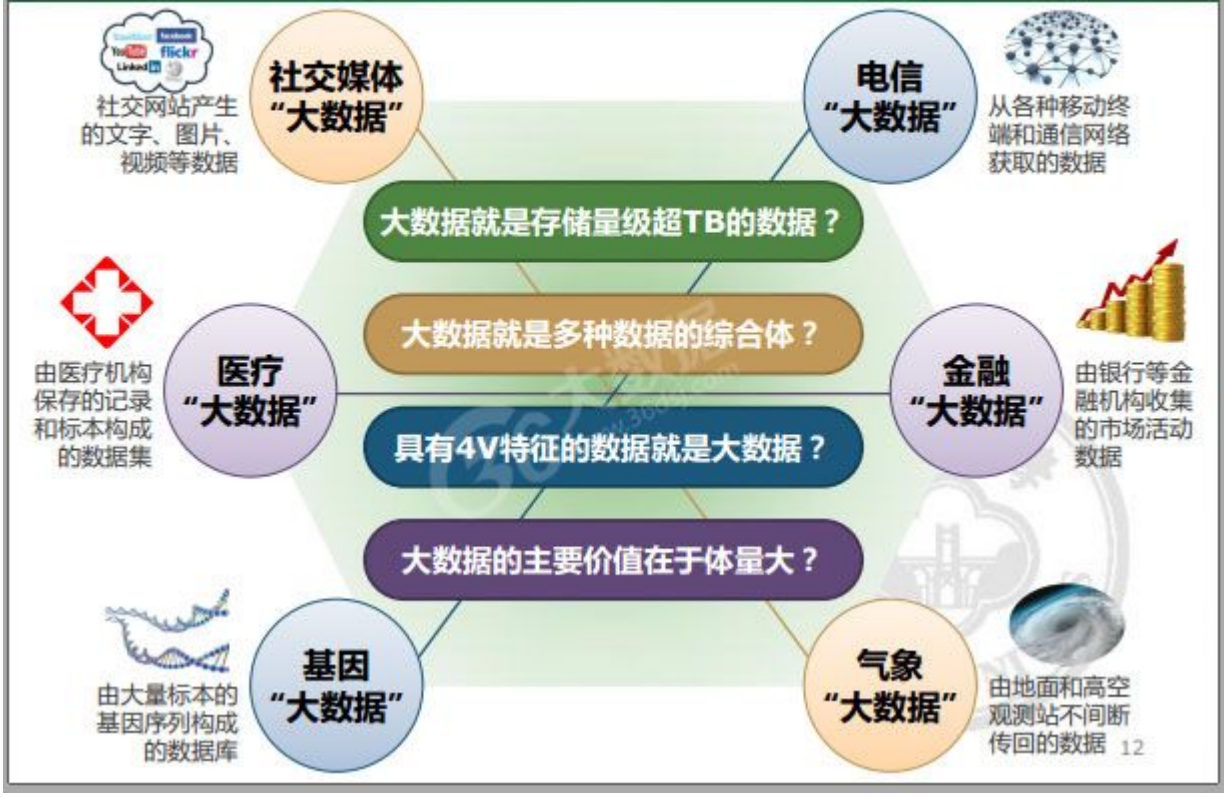

中山大学
SUN YAT-SEN UNIVERSITY
1924
大数据的理解与
分布式进化计算方法
张军
中山大学 超级计算机学院
2014年10月
1

大数据是什么？


社交媒体“大数据”
社交网站产生的文字、图片、视频等数据
电信“大数据”
从各种移动终端和通信网络获取的数据
大数据就是存储量级超TB的数据？
大数据就是多种数据的综合体？
具有4V特征的数据就是大数据？
大数据的主要价值在于体量大？
医疗“大数据”
由医疗机构保存的记录和标本构成的数据集
金融“大数据”
由银行等金融机构收集的市场活动数据
基因“大数据”
由大量标本的基因序列构成的数据库
气象“大数据”
由地面和高空观测站不间断传回的数据

大数据 = 海量数据？

大数据是海量数据的另一种说法

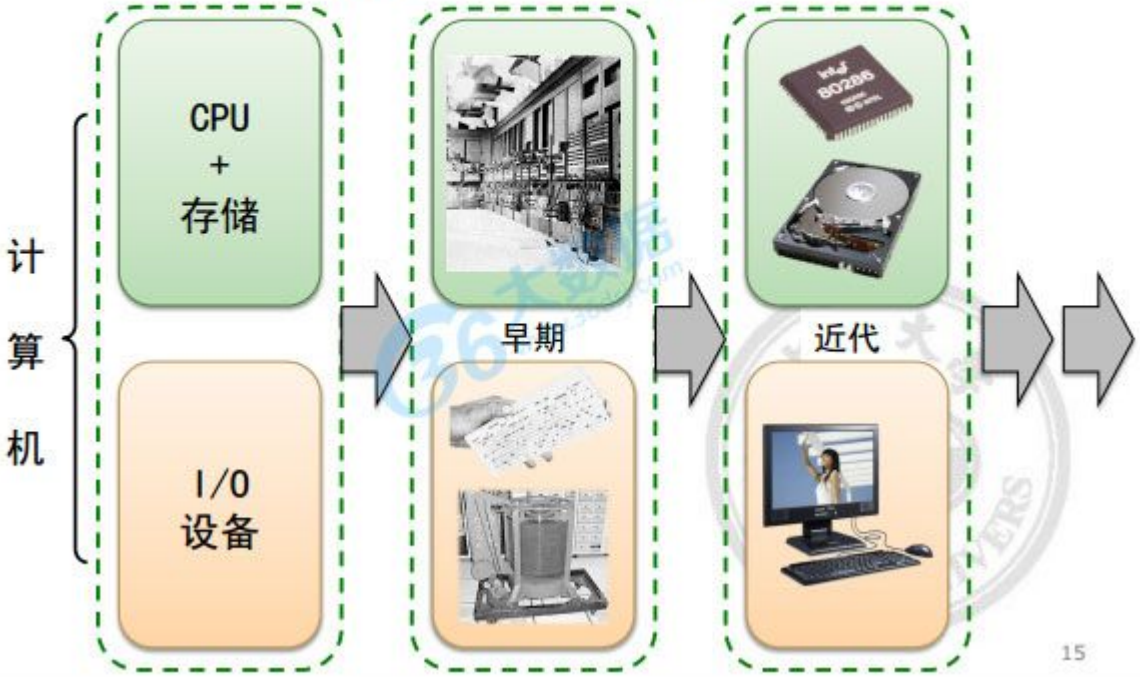
大数据只是体量比海量数据更大

大数据挖掘就是海量数据挖掘

我们的理解

大数据的认识

从计算机的发展历程说起



大数据的认识

从计算机的发展历程说起

计算机

CPU + 存储

超级计算机

云计算平台

神经突触计算机芯片

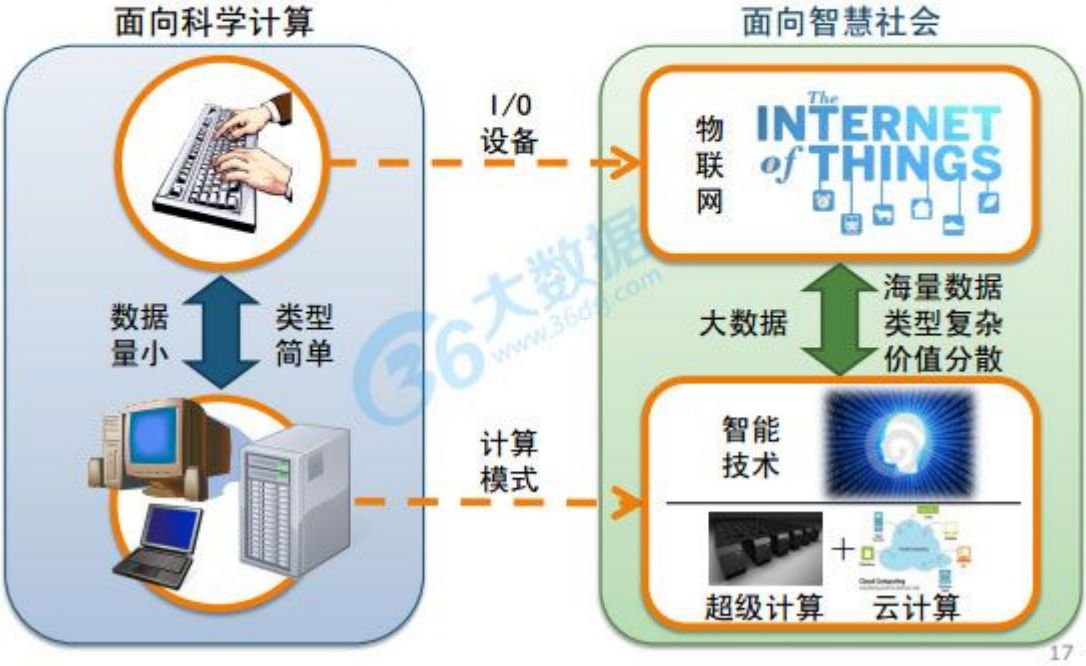
今天

I/O 设备

物联网的广泛应用

大数据的认识

大数据是物联网与新型计算模式发展的产物



大数据的认识

物联网 - 大数据 - 分布式计算

大数据的认识

大数据与智慧城市

建设智慧城市的关键技术

智慧体的感官
物联网技术

智慧体的血脉
大数据技术

智慧体的大脑
高性能计算技术

“智慧城市”是全球城市化建设的目标，也是我国跨越式发展的重要途径

大数据与海量数据的区别

如何解决大数据问题？

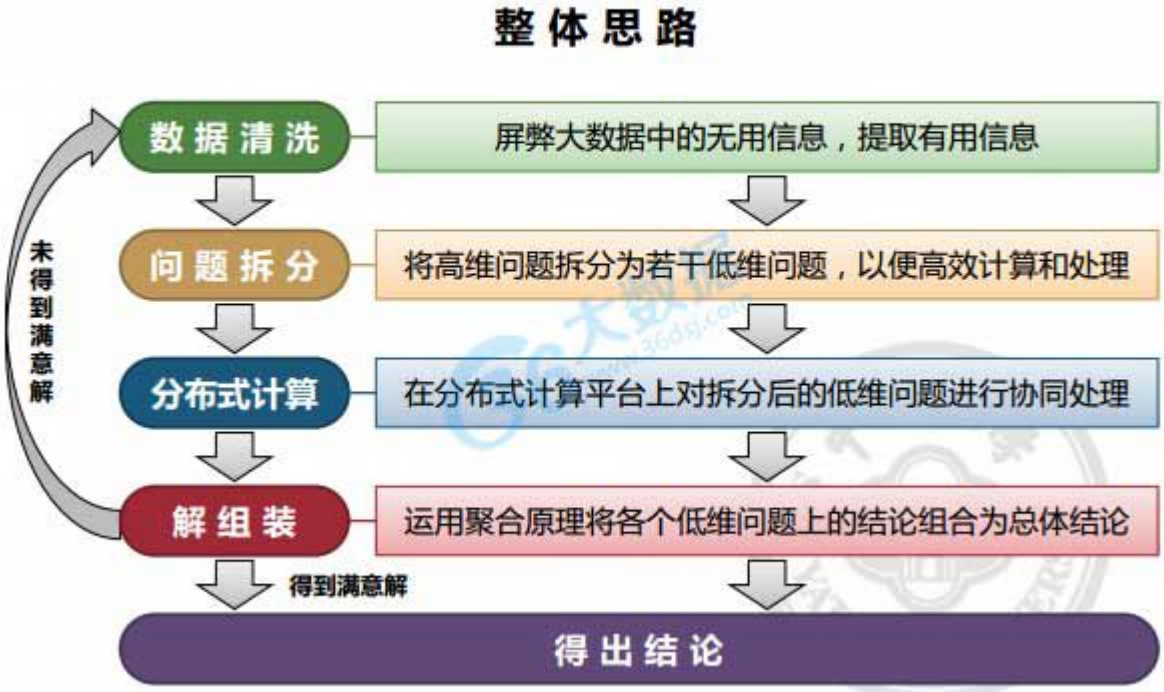

大数据的特点
体量巨大
类型复杂
价值丰富但分布不均
缺乏数学模型
具有容错性，以找到可接受解为目标
基于数学逻辑的人工智能方法难以解决缺乏数学模型的大数据问题
解决
大数据的特点使传统数据分析方法不再适用
新方法的可行途径
人工智能领域的计算智能方法将成为解决大数据问题的主要途径
进化计算
群体智能
深度学习

可行的大数据分析方法

海量数据与大数据分析的异同

海量数据分析		大数据分析
有明确的分析目标	VS	没有具体的分析目标
注重获取因果关系		不仅挖掘因果关系，更注重发现关联关系
追求一个精确的结果		发现未知的知识
		期望快速找到可接受解

可行的大数据分析方法


整体思路
数据清洗
屏弊大数据中的无用信息，提取有用信息
问题拆分
将高维问题拆分为若干低维问题，以便高效计算和处理
分布式计算
在分布式计算平台上对拆分后的低维问题进行协同处理
解组装
运用聚合原理将各个低维问题上的结论组合为总体结论
未得到满意解
得到满意解
得出结论

可行的大数据分析方法

问题拆分方法

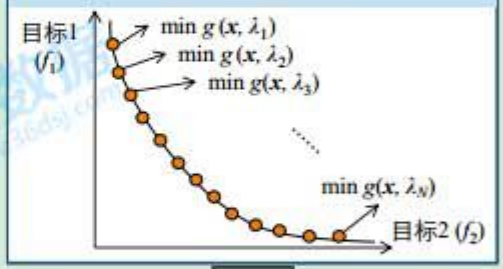
值得研究的拆分方法

- 随机拆分
- 按任务拆分
- 按目标拆分
- ⋮

复杂多目标问题

$$\min g(x,\lambda_i) = \lambda_i^{(1)} f_1(x) + \lambda_i^{(2)} f_2(x)$$

$$\text{where } \lambda_i^{(1)} + \lambda_i^{(2)} = 1$$



分拆

多个单目标问题

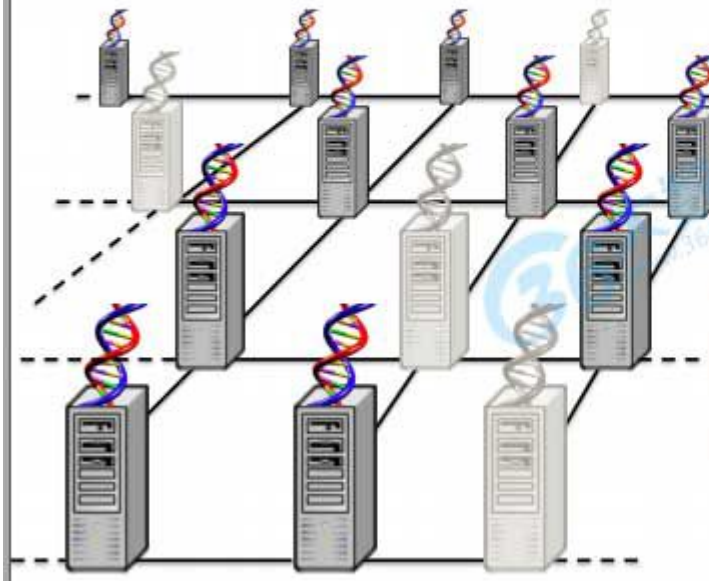
$\min g(x, \lambda_1)$ ⇒ 子系统1

⋮

$\min g(x, \lambda_N)$ ⇒ 子系统N

可行的大数据分析方法

分布的、自组织的智能计算方法



基于云计算平台的分布式进化算法

- 每个计算结点保存一个染色体
- 通过通信网络实现染色体的交叉
- 通过唤醒或删除计算结点来实现种群大小的自适应调整

研究背景与意义

国家需求

- “智慧城市”的建设需求
- 多个领域的智能化需求

智能交通控制

智能数据处理

…

智能传感器网络

智能化需求中的优化问题

- 缺乏精确数学模型
- 具有高维、非线性等特点
- 传统方法难以求解

需求

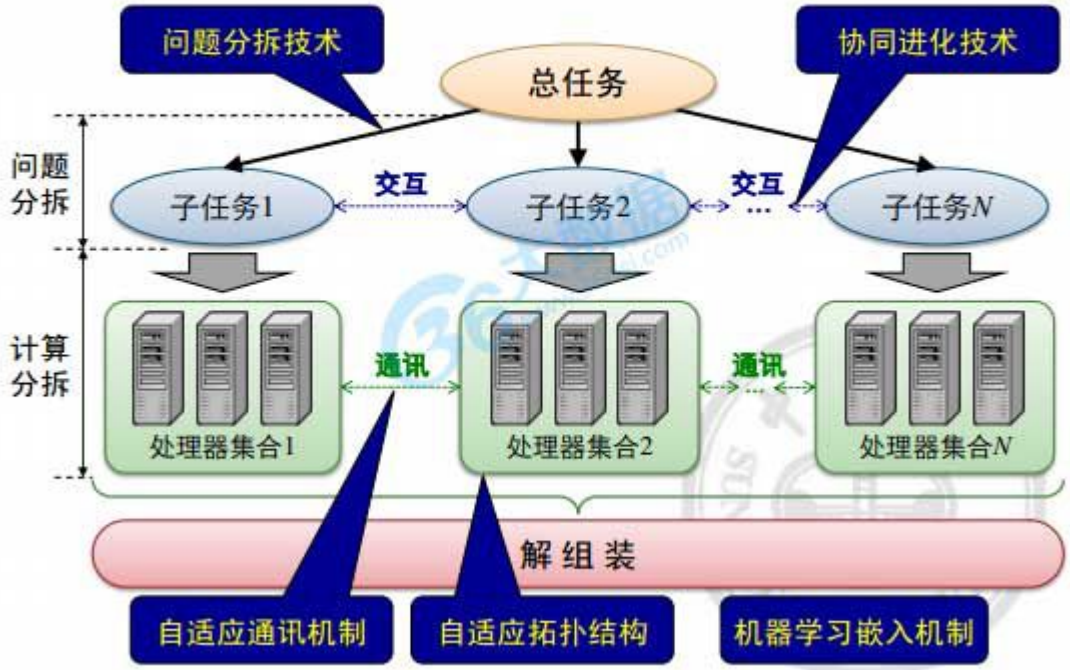
进化计算

- 模拟自然进化的智能方法
- 不依赖问题的数学特性
- 成为解决复杂优化问题的重要途径

- 传统进化计算方法求解大规模复杂优化问题时存在性能瓶颈
- 传统进化计算方法以串行执行为主，无法发挥分布式计算平台优势

分布式进化计算

主要研究内容



研究1：多种群协同进化的多目标PSO

研究背景与意义

问题提出

粒子群优化算法求解多目标问题遭遇效率瓶颈

难点所在

如何克服多个目标冲突导致的赋值困难?

主要研究内容

- 多种群多目标优化框架
- 多种群协同进化的多目标PSO算法(CMPSO)

传统PSO

单个种群同时优化多个冲突目标

赋值困难

VS

CMPSO

多个种群分别优化各个目标 + 种群间进行信息交流与共享

均匀地逼近整个Pareto前沿

研究1：多种群协同进化的多目标PSO

算法思想

主要创新点：

- ☑ 每个种群优化一个目标：**解决了赋值困难，充分探索了目标空间**
- ☑ 种群通过全局Archive共享信息：**促使解均匀覆盖整个Pareto前沿**
- ☑ 上述思想可兼容各类进化算法：**多种群协同优化多目标问题的通用算法框架**

研究1：多种群协同进化的多目标PSO

单个种群的进化策略

◎ 每个种群优化一个目标，进化过程与单目标PSO算法类似。

◎ 每个种群中粒子的位置更新方法：

新位置 = 原位置 + 新速度

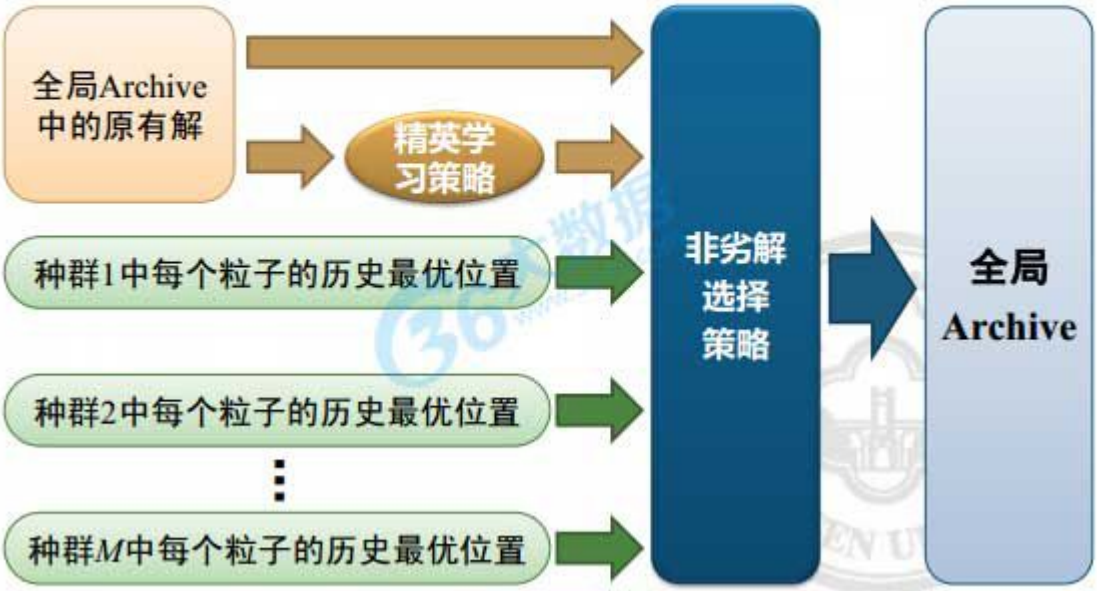
- 原速度的惯性影响
- 个体历史最佳位置的引导
- 种群历史最佳位置的引导
- 全局Archive中非劣解的引导

每个粒子从全局Archive中**随机**选择一个学习对象：

- ☑ **计算量小**
- ☑ **保持多样性**

研究1：多种群协同进化的多目标PSO

全局Archive的更新策略 (1)



全局Archive的更新策略 (2)

◎ **精英学习策略**：

新解 = 原解 + 随机扰动向量

随机扰动全局Archive中的解，帮助算法跳出局部最优。

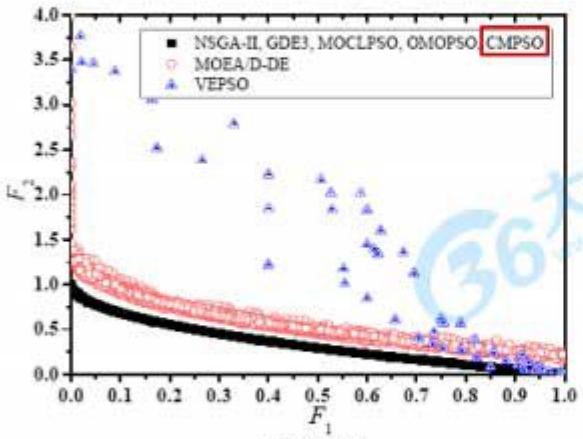
◎ **非劣解选择策略**：

① 运用非支配排序（non-dominated sorting）选出非劣解；

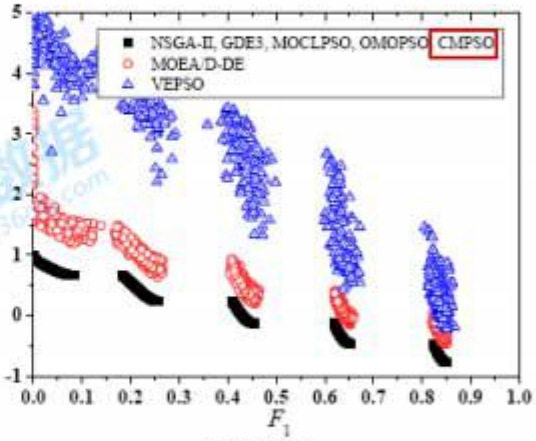
② 如果非劣解的数目超过Archive的容量，根据拥挤距离（crowding-distance）删去分布密度最大的若干解。

研究1：多种群协同进化的多目标PSO

实验结果与讨论



(a) ZDT1

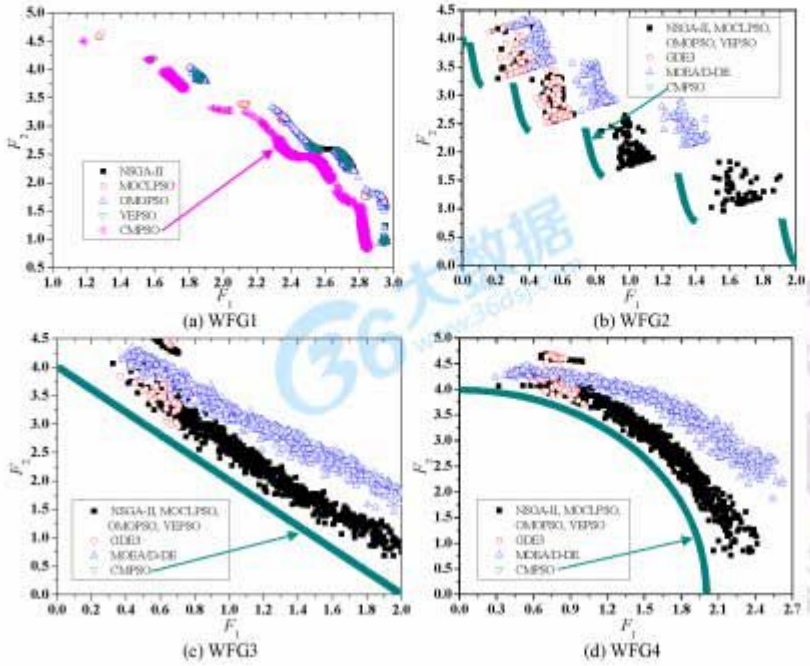


(b) ZDT3

CMPSO在ZDT系列问题上能够很好逼近的Pareto前沿

研究1：多种群协同进化的多目标PSO

实验结果与讨论



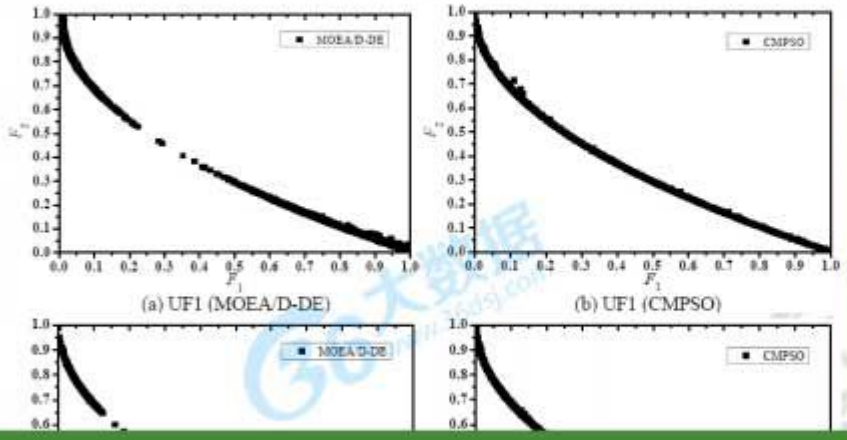
(a) WFG1　(b) WFG2

(c) WFG3　(d) WFG4

CMPSO 在WFG系列问题上的全局收敛能力比其他算法好

研究1：多种群协同进化的多目标PSO

实验结果与讨论



(a) UF1 (MOEA/D-DE)　　(b) UF1 (CMPSO)

CMPSO 在 UF1 和 UF2 上对 Pareto 前沿的覆盖比MOEA/D-DE更均匀

研究成果已发表于国际期刊IEEE Transactions on Cybernetics

Z.-H. Zhan, J. Li, J. Cao, J. Zhang, H. Chung, and Y. H. Shi, "Multiple populations for multiple objectives: A coevolutionary technique for solving multiobjective optimization problems," *IEEE Transactions on Cybernetics*, vol. 43, no. 2, pp. 445 – 463, Apr. 2013.

研究2：带双层Archive的多目标进化算法

研究背景与意义

问题提出

多目标优化问题期望得到尽量接近Pareto前沿且均匀分布的一组解

难点所在

Pareto前沿的精确逼近与全面覆盖常难以兼顾

主要研究内容

运用双层Archive结构同时实现Pareto前沿的精确逼近与全面覆盖

现有多目标进化算法

第1类：如NSGA II

基于非支配排序和拥挤距离对问题进行整体求解

优点：全面性较好　**缺点**：对复杂Pareto前沿逼近精度较低

第2类：如MOEA/D

通过权重将原问题拆分为多个单目标问题分别进行优化

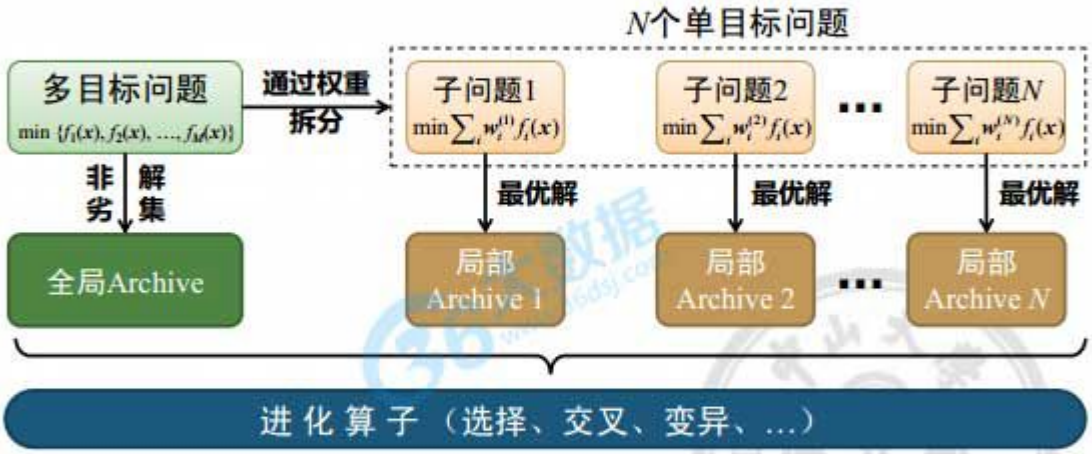
优点：精度较高　**缺点**：解的分布存在缺陷

MOEA-DLA

通过双层Archive结构结合两类算法的优点

研究2：带双层Archive的多目标进化算法

算法思想

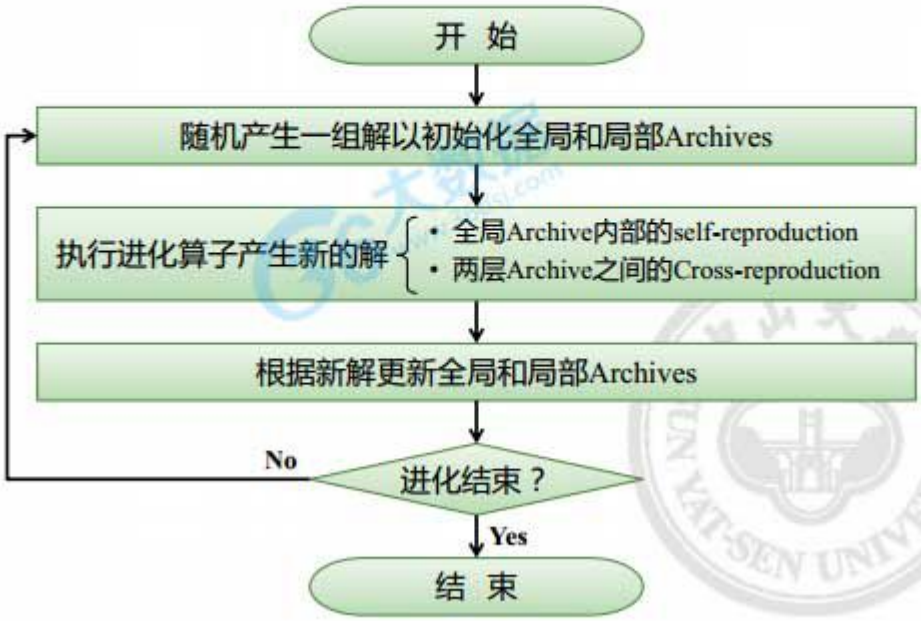


主要创新点：

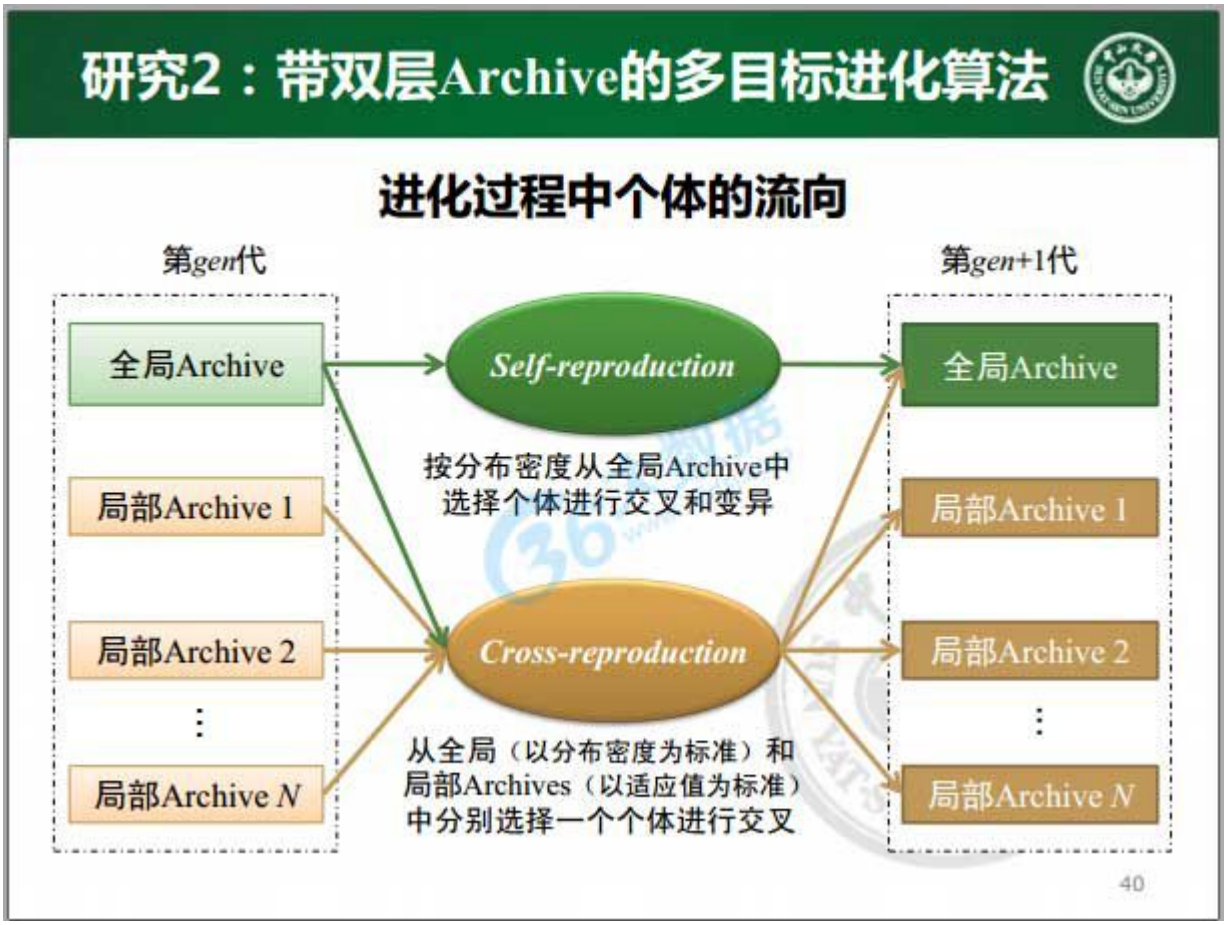
- ☑ 在多目标问题和单目标子问题两个层面上同时进化：**兼顾全面性和搜索精度**
- ☑ 上述算法框架可兼容多种进化算子：**灵活性强，可根据实际问题进行拓展**

研究2：带双层Archive的多目标进化算法

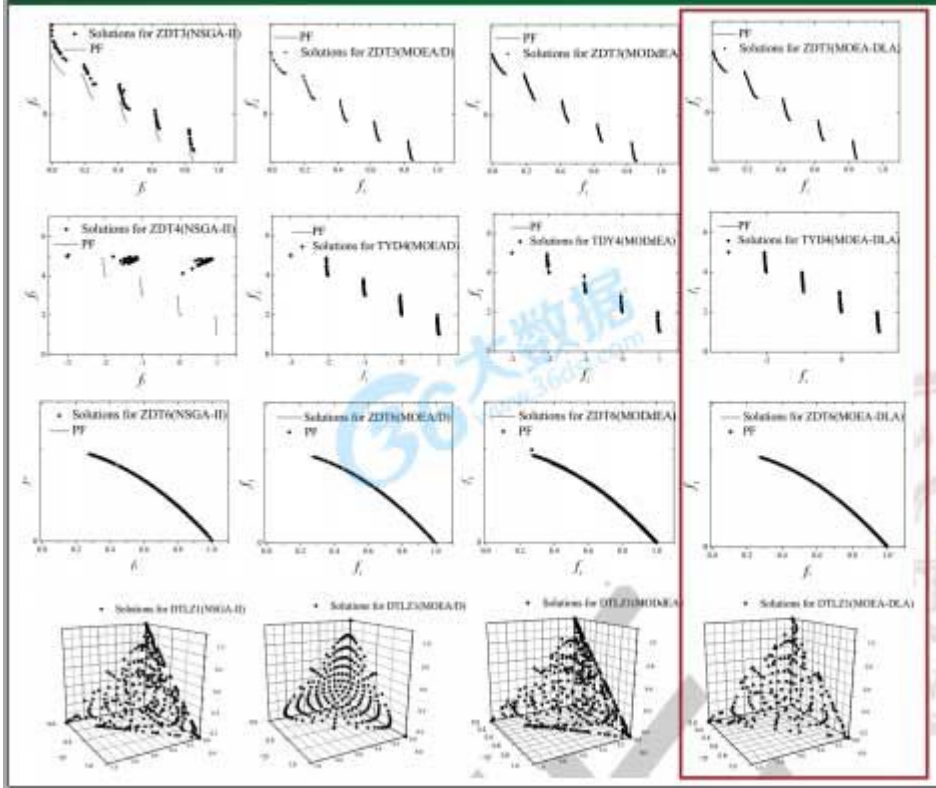
算法流程



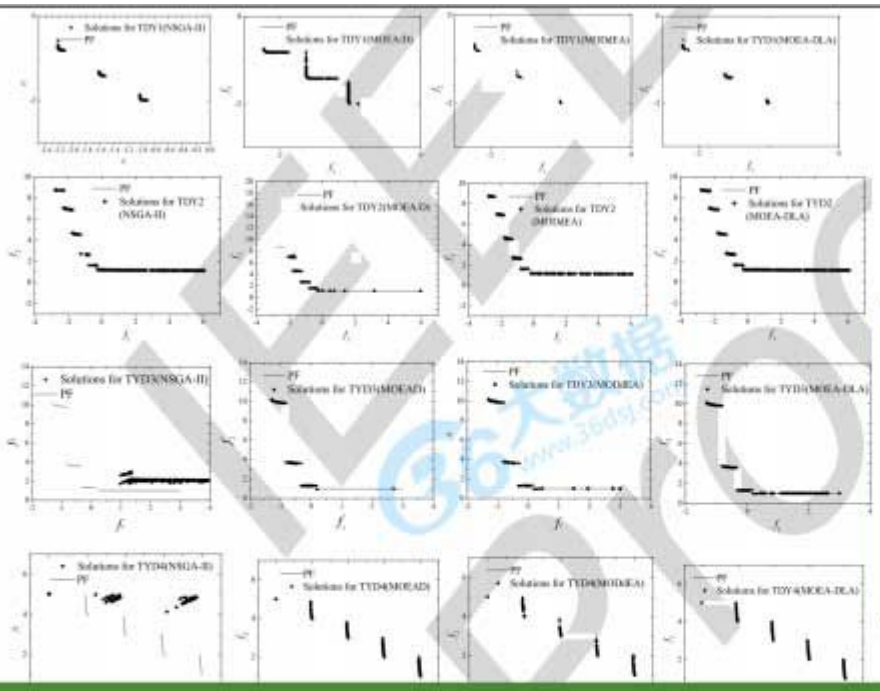
研究2：带双层Archive的多目标进化算法



研究2：带双层Archive的多目标进化算法



在ZDT，WFG和DTLZ问题上，我们的算法均能很好近似帕累托前沿。



在TDY问题上，MOEA-DLA的结果明显优于其它算法，不仅更接近帕累托前沿，覆盖也更均匀。

研究成果已被IEEE Transactions on Cybernetics接收

N. Chen, W.-N. Chen, Y.-J. Gong, Z.-H. Zhan, J. Zhang, Y. Li, & Y.-S. Tan, "An evolutionary algorithm with double-level archives for multiobjective optimization," in press.

研究3：并行粒子群优化算法

研究背景与意义

问题提出

不同版本的PSO算法各有优缺点，不存在一个适用于所有问题的算法

难点所在

如何集成不同版本PSO算法的优点?

主要研究内容

基于多个版本PSO的并行PSO算法（PPSO）

传统PSO

单个处理器使用单一算法

在一些问题上表现好
另一些问题上表现差

VS

PPSO

多个处理器并行执行不同算法 + 处理器之间的通信机制

集成多个算法的优点，普适性更强

研究3：并行粒子群优化算法

算法思想

- $gBest_i$：第i个从处理器上种群找到的历史最优解
- $GBest$：所有$gBest_i$中的最优解

主要创新点：通过不同算法的并行执行和有效通信，提高算法的求解效率和普适性。

研究3：并行粒子群优化算法

处理器间的通信机制

◎ 采用主-从拓扑结构：

从处理器 ➡ 主处理器

- 从处理器的进化过程被划分为若干阶段；
- 每阶段结束时，从处理器向主处理器发送其种群的历史最优解。

主处理器 ➡ 从处理器

- 主处理器从所接收的解中挑选全局最优解，并将其发送给各个从处理器；
- 从处理器将其种群中随机选择的任意一个个体替换为全局最优解。

通讯机制的作用：

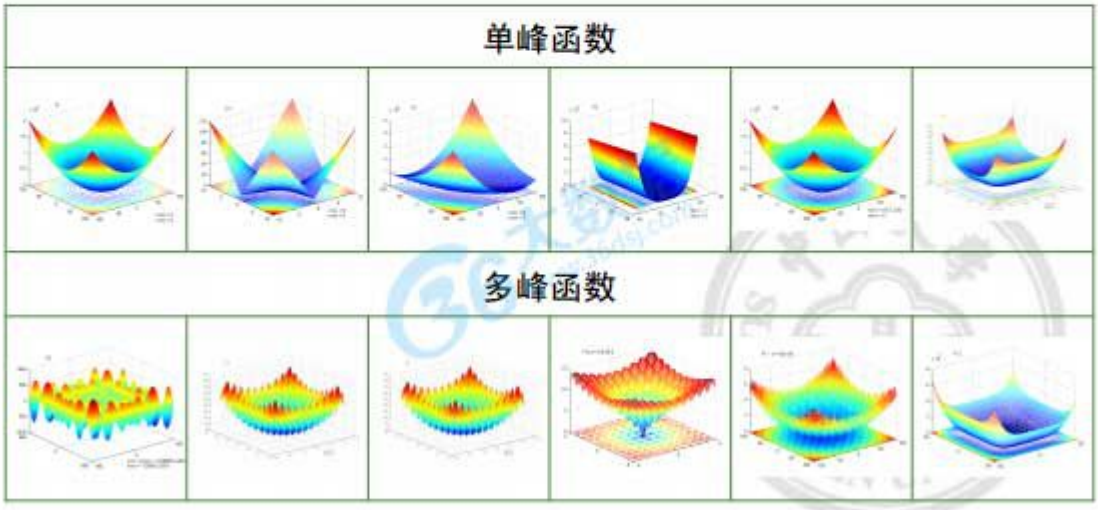
☑ 从处理器分阶段多次向主处理器传递最优解：**充分发挥不同算法的搜索能力**

☑ 主处理器将全局最优解返回给从处理器：**保持多样性，提供更好的引导向量**

研究3：并行粒子群优化算法

实验结果与讨论

◎ 在12个30维标准测试函数上进行了测试



研究3：并行粒子群优化算法

实验结果与讨论

（30次运行的平均结果，**红色粗体**表示最佳结果，***蓝色斜体***表示次优结果）

F	GPSO	LPSO	BPSO	CLPSO	PPSO
f_1	$3.76*10^{-77}$	$3.09*10^{-18}$	**$3.35*10^{-218}$**	$1.19*10^{-74}$	***$3.75*10^{-216}$***
f_2	$2.00*10^{-41}$	$2.36*10^{-13}$	**$2.39*10^{-152}$**	$4.16*10^{-47}$	***$9.33*10^{-152}$***
f_3	$1.66*10^{-3}$	499.85	**$4.83*10^{-13}$**	$8.14*10^{-2}$	***$1.37*10^{-12}$***
f_4	0.22	9.39	***$7.41*10^{-05}$***	0.27	**$1.48*10^{-05}$**
f_5	44.60	37.02	***10.08***	32.72	**2.85**
f_6	**0**	**0**	***3.1***	**0**	**0**
f_1	3089.33	4242.06	1342.33	***43.44***	**$1.34*10^{-2}$**
f_2	26.47	37.63	75.92	***7.96***	**6.87**

研究3的阶段性成果已发表于国际会议IES 2014

G.-W. Zhang, Z.-H. Zhan, K.-J. Du, Y. Lin, W.-N. Chen, J.-J. Li, & J. Zhang, "Parallel particle swarm optimization using message passing interface," in *Proceeding of IES 2014*, Singapore.

谢谢！