大数据分析中的计算智能优化方法研究现状与展望

2018年5月12日

主要内容

- 1. 研究方向
- 2. 科研项目
- 3.研究成果
- 4. 计算智能优化算法
- 5. 研究存在问题
- 6.研究实例
- 7.大数据分析中的计算智能优化方法研究现状与展望

1. 研究方向

- 主要研究为智能优化算法的优化理论与方法，生产系统调度以及物流优化调度理论与应用、拓展智能优化算法在计算机网络拓扑优化、数据挖掘、物流运输等领域进行了探索
- 智能优化算法包括算法框架及搜索策略的设计、算法的收敛理论与计算复杂性等方面进行研究，并在生产与运输协调调度实际工程项目中推广应用。
- 大数据分析、机器学习、计算智能优化方法的研究。

2. 科研项目

(1) 主持辽宁省教育厅资助科研<分布估计与分散搜索算法的改进及其在QOS多播路由中的研究> 项目 （课题编号：L2015265，项目起止时间：2015.05-2017.12）。

(2) 主持辽宁省教育厅资助科研<新型智能优化算法的改进及其在组合优化问题中的应用研究> 项目 （课题编号：L2010196，项目起止时间：2010.05-2012.12）。

(3) 主持鞍山科技局资助科研<基于智能优化算法的运输车辆调度的应用研究> 项目 （项目起止时间：2015-2017）。

(4) 主持横向科研项目< 路径优化算法> 项目 （项目起止时间：2017-2019)。

(5) 参加国家自然科学基金，异构信息空间中时间感知的个性化语义实体搜索关键技术研究，项目编号： （课题编号：61402213 ，项目起止时间：2015.01~2017.12）

2 科研项目

(6) 参加辽宁省自然科学基金，面向医疗大数据基于语义认知与演化认知的实体集成关键技术研究，项目编号： （课题编号：20170540471，项目起止时间：2017.05~2020.04）

(7) 主持横向科研项目<物流配送车辆优化调度系统的研究> 项目(项目起止时间：2014-2017)。

(8) 参加鞍山科技局资助科研<以数据为中心的增量式企业数据集成系统研究> 项目 (课题编号：L2010196，项目起止时间：2013.12-2016.12)。

3 学术成果

1. *__Xiaoxia Zhang,__ Lixin Tang. A New Hybrid Ant Colony Optimization Algorithm for the Vehicle Routing Problem. Pattern Recognition Letters, 2009,30:848-855. (SCI收录, ISI:000267415400011)
2. Lixin Tang, *__Xiaoxia Zhang__, Two hybrid metaheuristic algorithms for hot rolling scheduling. ISIJ International. 2009, 49(4) :529-538. (SCI收录, ISI:000265474600009)
3. *__张晓霞__, 唐立新, 一种新的求解TSP问题的ACO&SS算法设计, 控制与决策, 2008, 23 (7):762-766. (国内核心期刊, EI检索:083711540955)
4. *__张晓霞__, 唐立新, 一种新的求解MMKP问题的ACO&PR算法, 控制与决策, 2009, 24(5):729-733 (国内核心期刊, EI检索:20092412122890)
5. *__张晓霞__, 刘哲. 一种新的求解多维背包问题的分散算法，计算机应用研究，2012，29(5):1716-1719. (中文核心)

3. 学术成果

6. *张晓霞，童杰伟，刘哲. 一种新的求解TSP问题的GRASP&PR算法设计, 计算机工程，2012，38(12)：122-124.（中文核心）
7. *张晓霞. 基于位置与连接概率的EDA算法求解PFSP问题, 计算机应用与软件, 2015, 32(12) （中文核心）
8. *张晓霞. 一种求解混合零空闲置换流水车间调度禁忌分布估计算法, 计算机应用与软件, 2017, 34(1) （中文核心）
9. *张晓霞. An Effective Hybrid Ant Colony Optimization for Permutation Flow-Shop Scheduling, The Open Automation and Control Systems Journal, 2014, 6, 62-68 ,EI检索：20143618136477
10. ***Xiaoxia Zhang,** Lixin Tang, New Hybrid Ant colony optimization Algorithm for the traveling salesman problem. ICIC 2008 International Conference, Shanghai, Sept, 2008, pp.17-19. (EI检索:084111631533)

4 智能优化算法　算法分类

- 求解组合最优化问题的数学方法有两类：一类为精确算法，如分支定界、动态规划、线性规划等,精确算法只能解决一些小规模问题；一类为近似算法，它又可分为基于最优化的近似算法、启发式算法和基于智能优化的近似算法。
- 近似算法是指在合理的计算时间内找到一个近似的最优解，基于最优化的近似算法是以数学模型为基础，采用列生成、拉格朗日松弛和状态空间松弛等求解问题。
- 启发式根据求解问题的特点，按照人们经验或某种规则设计的。这种方法比较直观、快速，但有时解的质量不高。

精确算法：优化建模与LINGO软件
WebSphere ILOG CPLEX 的优化软件

实际生产中采用近似算法

4 智能优化算法　算法分类

- 基于智能优化的近似算法是基于一定的优化搜索机制，并具有全局优化性能的一类算法。这类算法常见的有：模拟退火模拟退火(Simulated Annealing，简称SA)、遗传算法(Genetic Algorithm，简称GA)、禁忌搜索禁忌搜索(Tabu Search，简称TS)、蚁群算法(Ant Colony Optimization，简称ACO)、路径重连算法(Path Relinking，简称PR)、分散搜索(Scatter Search，简称SS)、蜂群算法（ Bee Colony Optimization ）等，这些算法也称超启发式算法。虽然基于智能优化的近似算法不能保证求得全局最优解，但因其高效的优化性能、无需问题特殊信息、易于实现且速度较快等优点，受到诸多领域广泛的关注和应用。

4 智能优化算法　算法分类

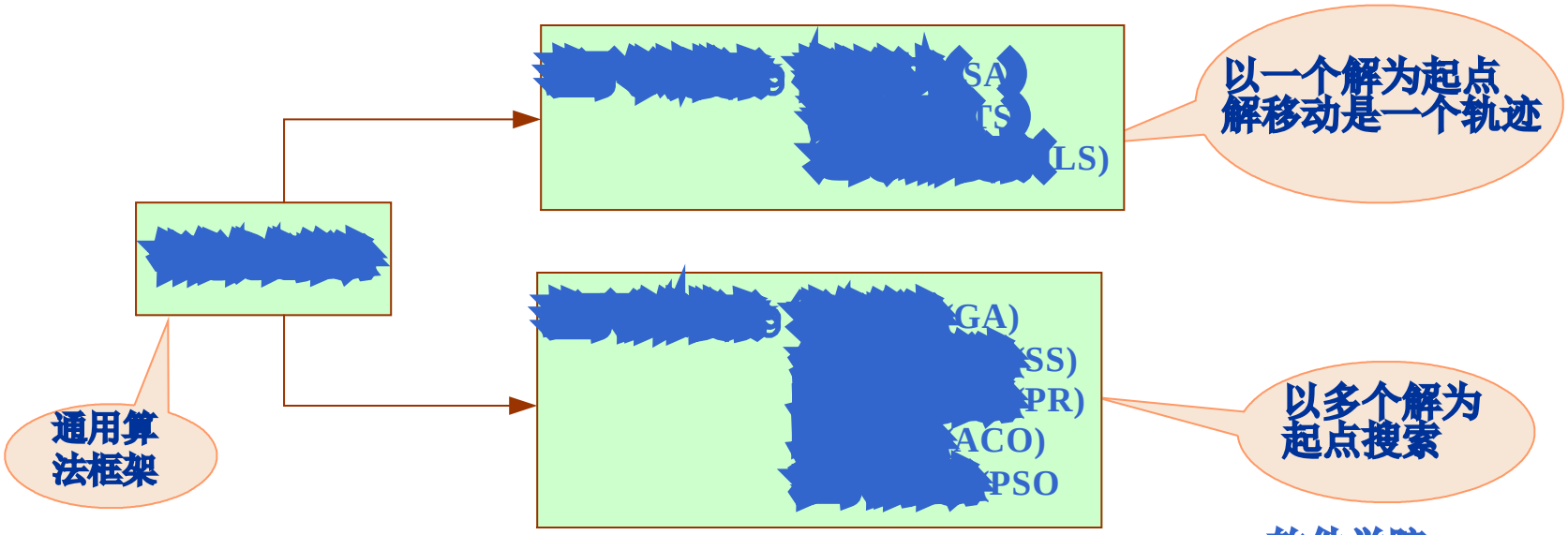
- **智能优化算法的介绍**

随着生物学及信息技术迅速发展，出现遗传算法(GA)、禁忌搜索(TS)、蚁群(ACO)等超启发算法.

- **智能优化算法的分类**

按算法起源、目标函数(动态更新)，邻域(单一、混合).

- **分类原则：按每次迭代解移动数目的来划分.**
- **基于邻域的搜索算法与基于群体搜索的智能优化算法.**



4 智能优化算法　算法分类

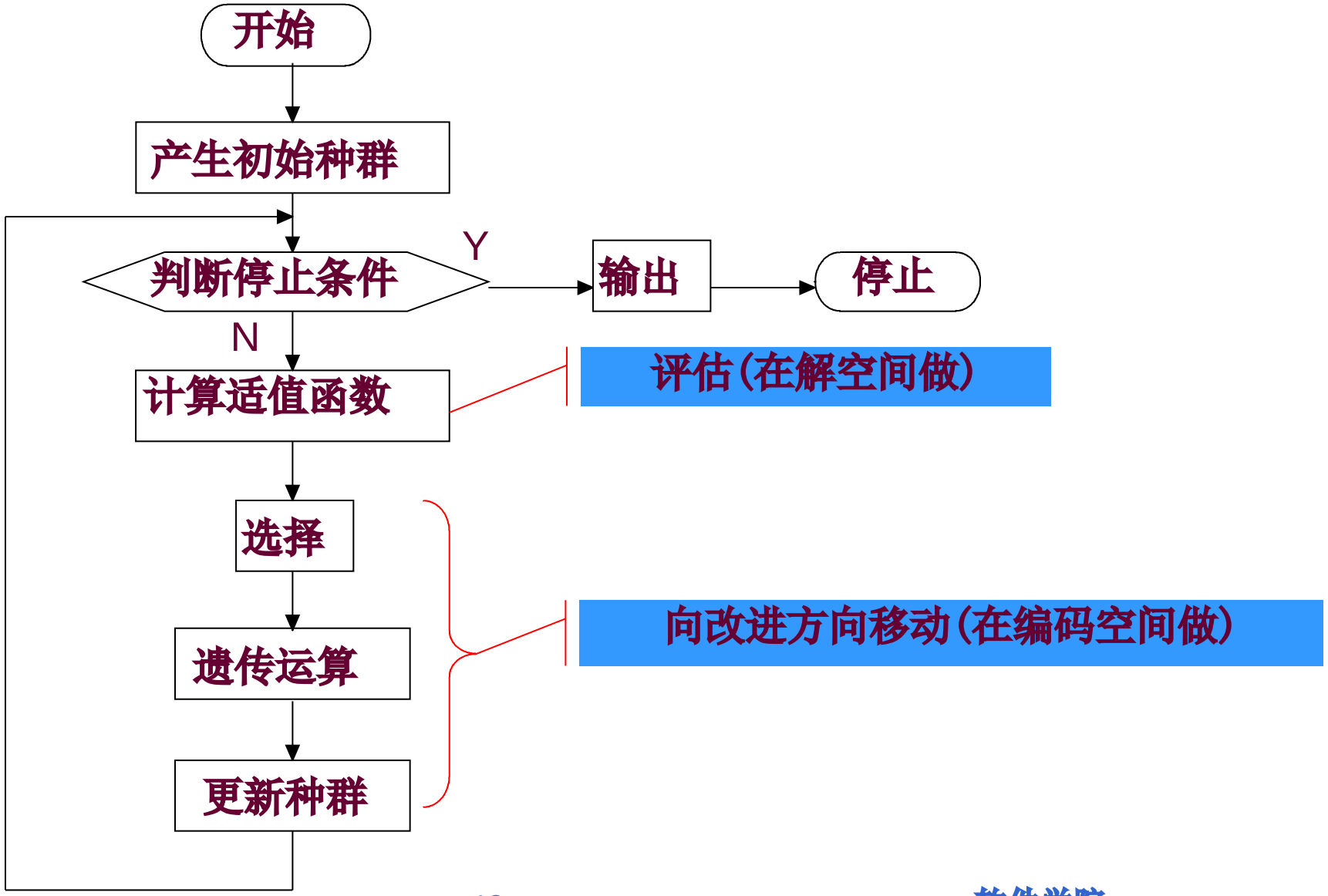
- 近年来，随着生物学、物理学和人工智能及信息技术迅速发展，出现了一些具有全局优化性能且通用性强的超启发算法(智能优化算法)。因其高效的优化性能，受到诸多领域广泛的关注和应用。
- 智能优化算法是一种通用的算法框架，只要根据具体问题特点对这种算法框架结构进行局部修改，就可以直接应用它去解决不同的问题。依算法搜索机制和搜索策略，对智能优化算法进行分类。
- 以基于生物特征的算法起源为依据进行划分，分为基于自然生物特性与非自然生物特性。GA、ACO、PSO算法都属于基于自然生物特性；SA、TS与ILS都属于基于非自然生物特性，但是对带记忆的禁忌搜索算法(TS)很难划分。至今，还没有一个有效的分类方法，把这类智能优化算法的特性都体现出来。下面介绍以每次迭代搜索解的数目为标准进行划分，分为个体搜索与群体搜索算法。

4 智能优化方法　遗传算法(GA)

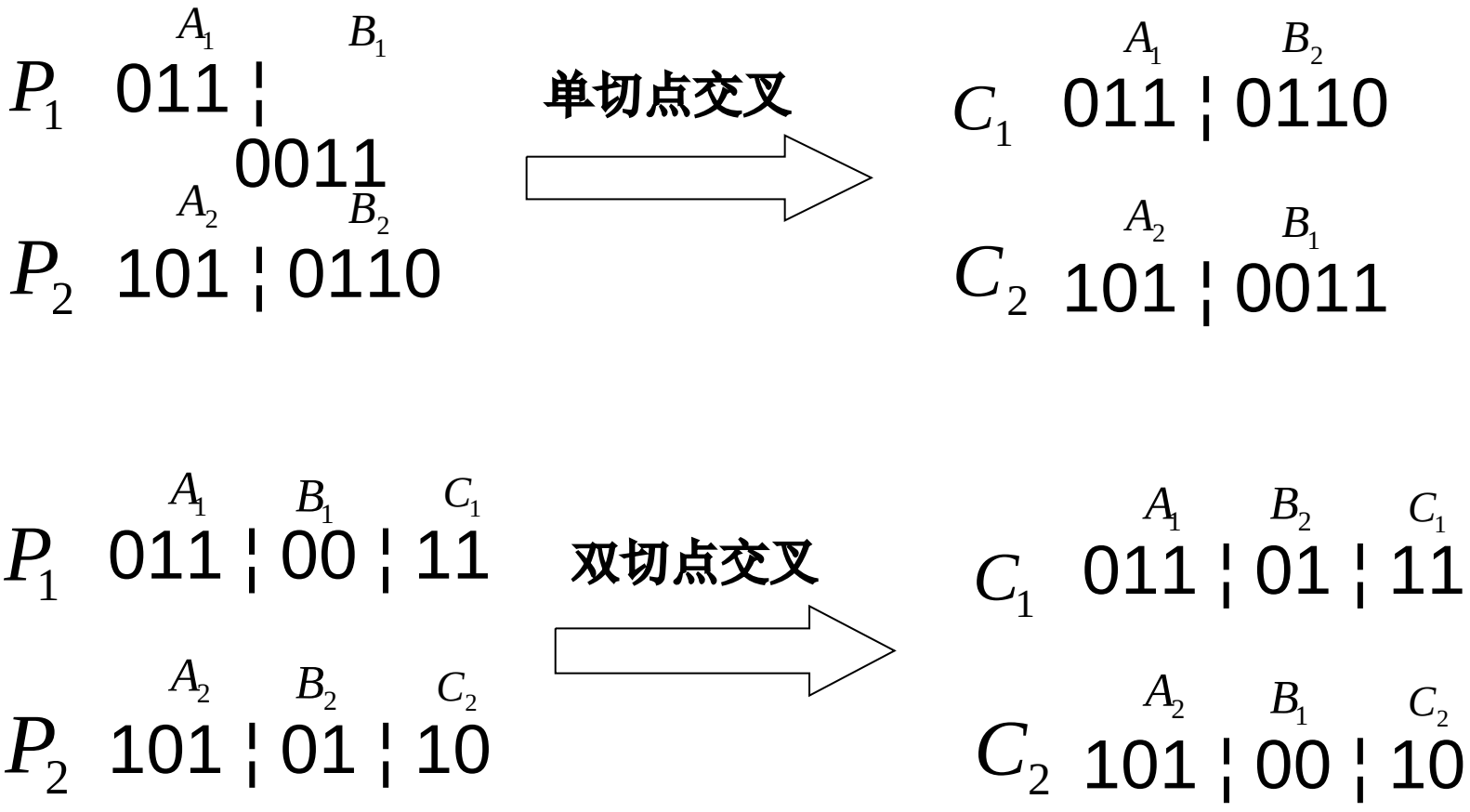
- 遗传算法(Genetic Algorithm，简称GA)是由美国学者Holland于1975年提出的，GA是基于生物自然遗传的基本原理的一种并行、随机和自适应的优化算法。由于GA编码技术与遗传操作简单，并行的全局空间搜索的特点。目前，GA吸引许多研究者，并在组合优化、机器学习]、工程设计、模式识别与神经网络等许多领域得到了应用。
- 遗传算法(GA)的来源：遗传学的“优胜劣汰”“自然选择”的思想。

(GA)缺点：无人的主动性；

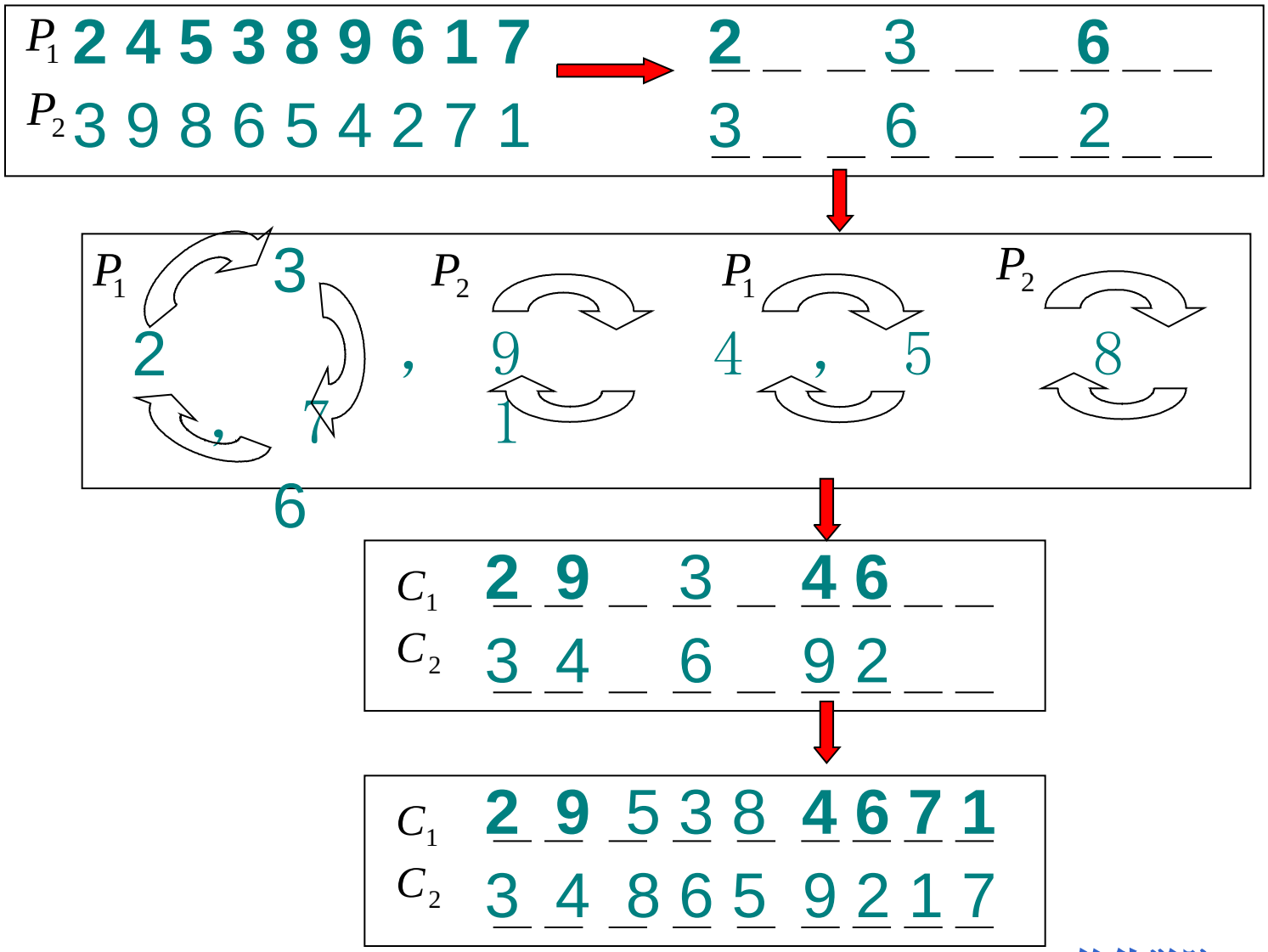
基本GA算法框图


开始
产生初始种群
判断停止条件
Y
输出
停止
N
计算适值函数
评估(在解空间做)
选择
遗传运算
更新种群
向改进方向移动(在编码空间做)

- 交叉(Crossover)



基本GA算法　变形


P1 2 4 5 3 8 9 6 1 7
P2 3 9 8 6 5 4 2 7 1
2 _ _ 3 _ _ 6 _ _
3 _ _ 6 _ _ 2 _ _
P1
P2
P1
P2
C1 2 9 _ 3 _ 4 6 _ _
C2 3 4 _ 6 _ 9 2 _ _
C1 2 9 5 3 8 4 6 7 1
C2 3 4 8 6 5 9 2 1 7

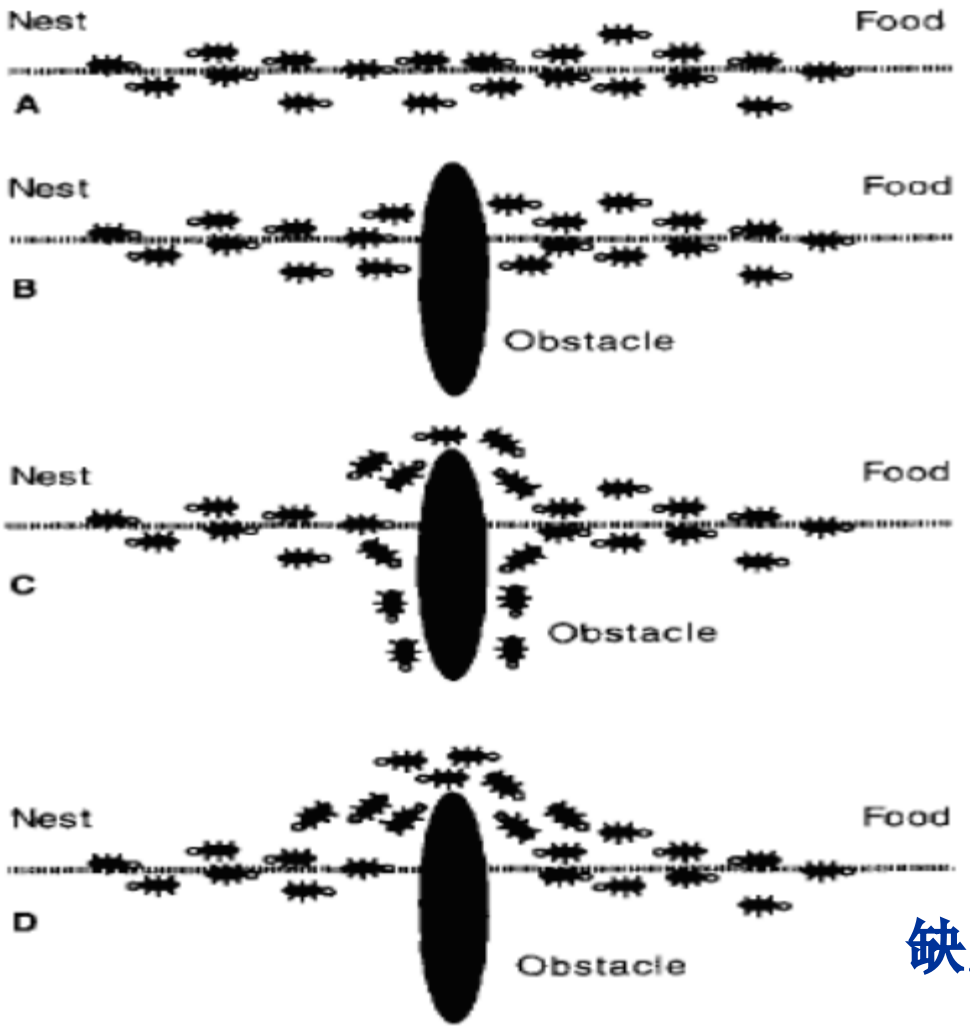
Marco Dorigo (S'92–M'92–SM'96) was born in Milan, Italy, in 1961. He received the Laurea (Master of Technology) degree in industrial technologies engineering in 1986 and the Ph.D. degree in information and systems electronic engineering in 1992 from Politecnico di Milano, Milan, Italy, and the title of Agrégé de l'Enseignement Supérieur, from the Université Libre de Bruxelles, Belgium, in 1995.

From 1992 to 1993 he was a Research Fellow at the International Computer Science Institute of Berkeley, CA. In 1993 he was a NATO-CNR Fellow and from 1994 to 1996 a Marie Curie Fellow. Since 1996 he has been a Research Associate with the FNRS, the Belgian National Fund for Scientific Research. His research areas include evolutionary computation, distributed models of computation, and reinforcement learning. He is interested in applications to autonomous robotics, combinatorial optimization, and telecommunications networks.

Dr. Dorigo is an Associate Editor for the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS and for the IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION. He is a member of the Editorial Board of *Evolutionary Computation* and of *Adaptive Behavior*. He was awarded the 1996 Italian Prize for Artificial Intelligence. He is a member of the Italian Association for Artificial Intelligence (AI*IA).

Luca Maria Gambardella (M'91) was born in Saronno, Italy, in 1962. He received the Laurea degree in computer science in 1985 from the Università degli Studi di Pisa, Facoltà di Scienze Matematiche Fisiche e Naturali.

Since 1988 he has been Research Director at IDSIA, Istituto Dalle Molle di Studi sull' Intelligenza Artificiale, a private research institute located in Lugano, Switzerland, supported by Canton Ticino and Swiss Confederation. His major research interests are in the area of machine learning and adaptive systems applied to robotics and optimization problems. He is leading several research and industrial projects in the area of collective robotics, cooperation and learning for combinatorial optimization, scheduling, and simulation supported by the Swiss National Fundation and by the Swiss Technology and Innovation Commission.


Nest
Food
A
Nest
Food
B
Obstacle
Nest
Food
C
Obstacle
Nest
Food
D
Obstacle

缺点:容易陷入局部最优

- 特点：每只蚂蚁按概率选择规则选择下一个移动点，重复这个选择过程直到完成一个解的搜索。
- 组成：解的构成与信息素更新。信息素通常有局部与全局更新方式。

$$j \quad \begin{cases} \arg\max_{l \ M_k}\{[\ _{il}][\ _{il}]\ \}, & \text{if } q \quad q_0 \\ S \quad , \quad \text{otherwise} \end{cases} \tag{1}$$

$$P_{ij}^{k} \quad \begin{cases} \dfrac{_{ij}[\ _{ij}]}{\sum_{l \ M_k} {}_{il}[\ _{il}]}, & j \quad M_k \\ 0 \quad , \quad \text{otherwise} \end{cases} \tag{2}$$

蚁群优化算法(ACO)流程

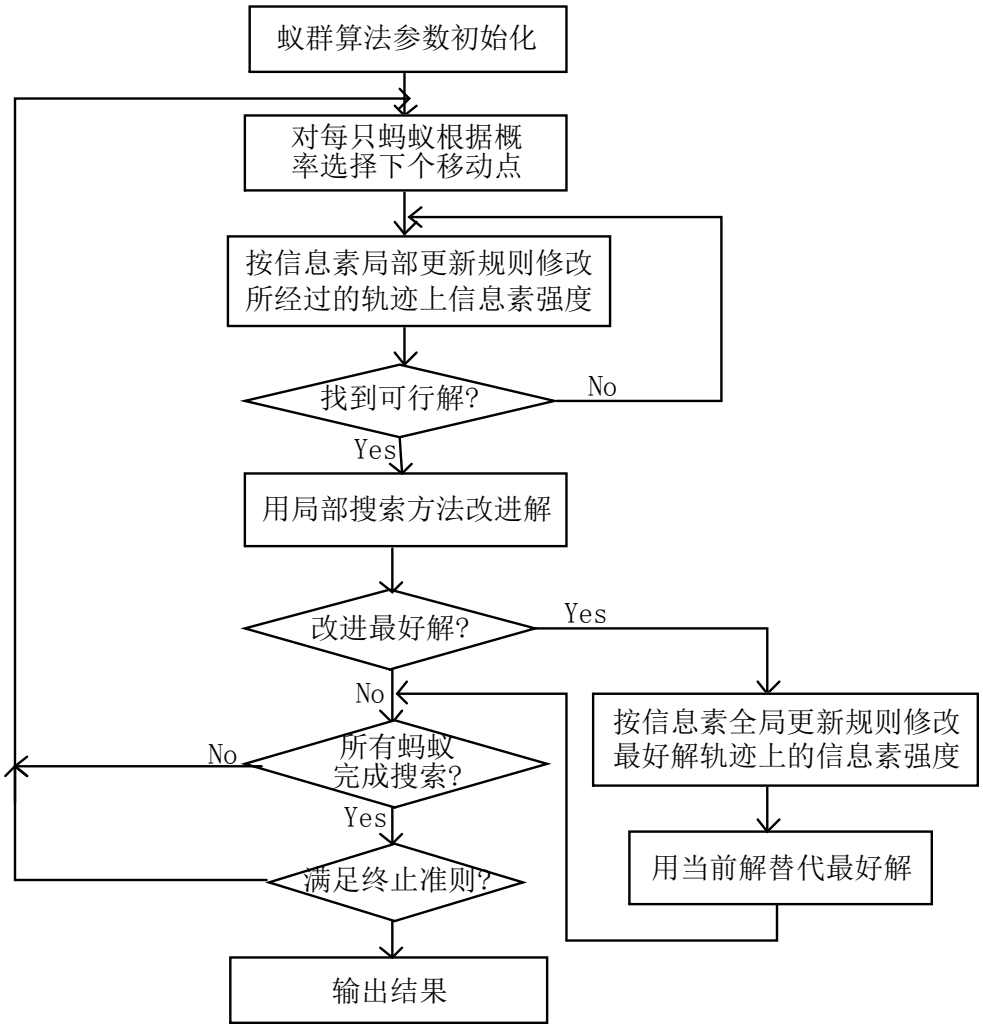


图 3.1 蚁群算法流程图

3 The Canonical Structure of Scatter Search

The process of solving a problem by means of creating progressively new better solutions is organized in the following five components:

1. Diversification method
2. Improvement method
3. Reference set update method
4. Subset generation
5. Solution combination

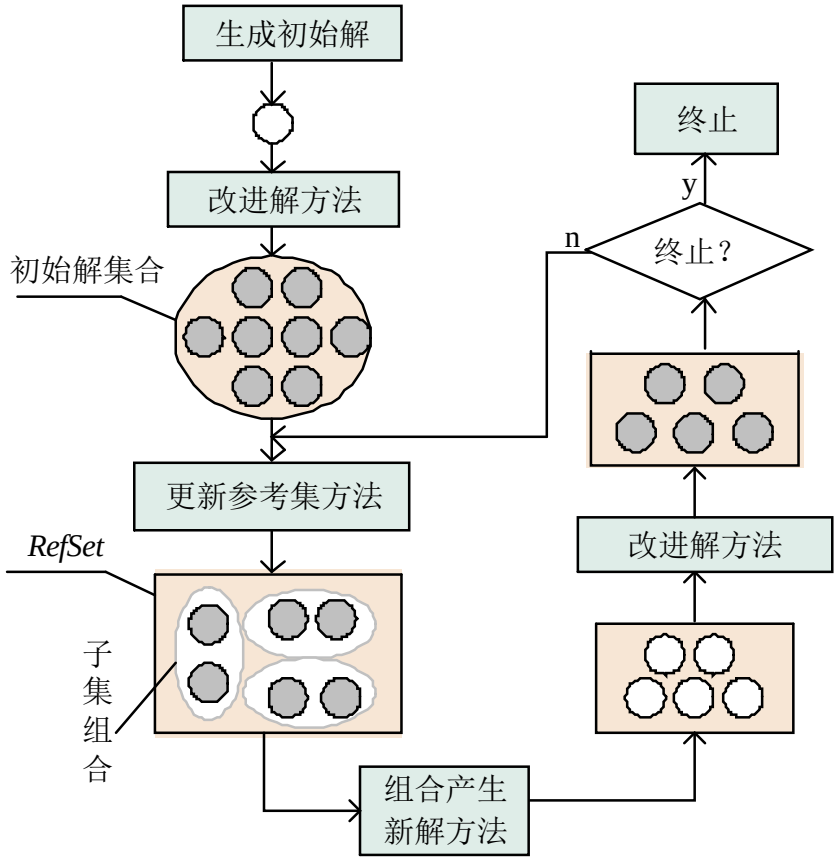
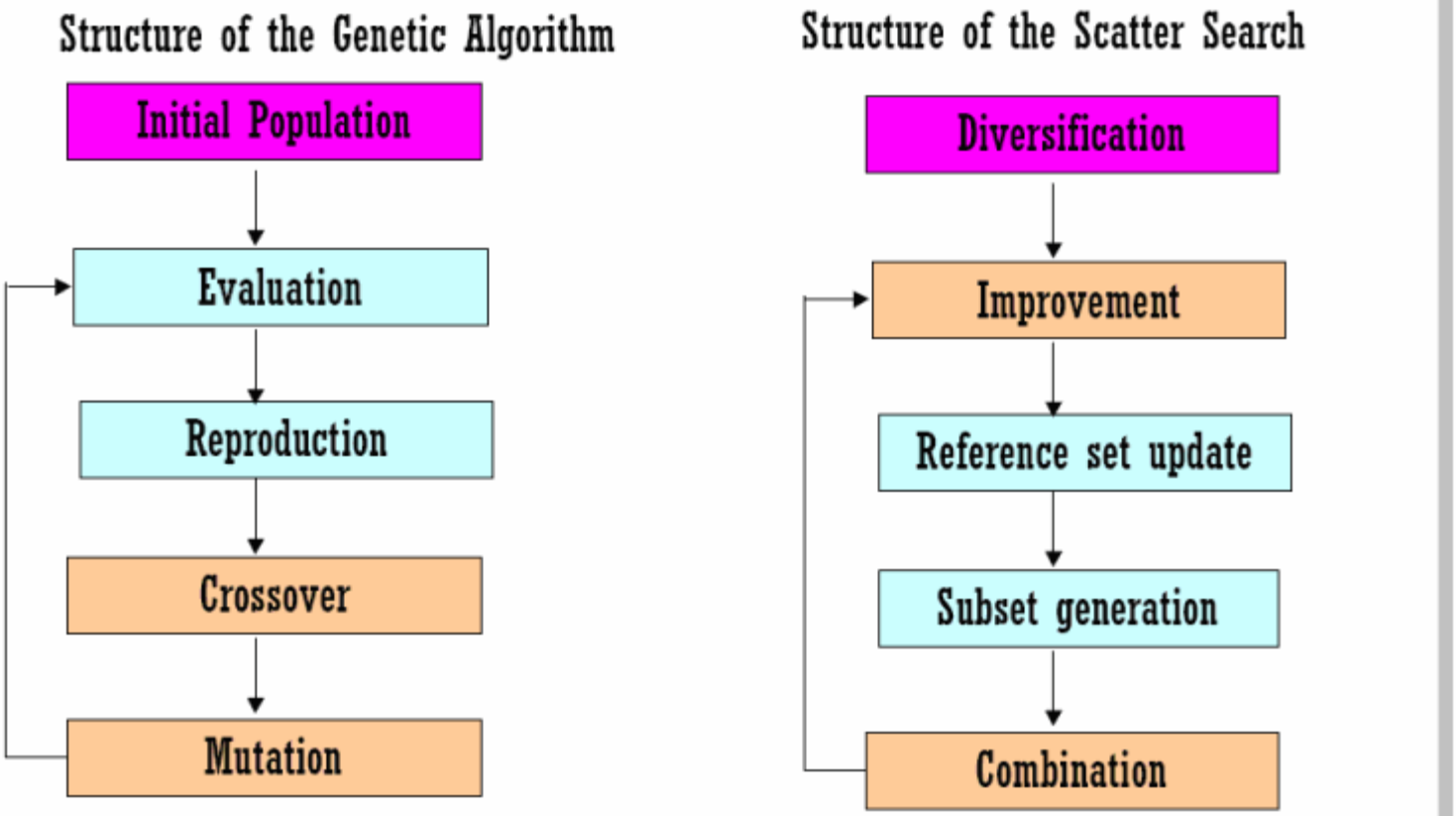


图 2.2 分散搜索算法的结构示意图

- **特点：在搜索过程中既保证解的质量，又能保证解分散性 。**
- **组成：参考集合生成、组合产生新解、解改进方法、参考集更新方法。**


Structure of the Genetic Algorithm
Initial Population
Evaluation
Reproduction
Crossover
Mutation
Structure of the Scatter Search
Diversification
Improvement
Reference set update
Subset generation
Combination
"Equivalences" between components
17

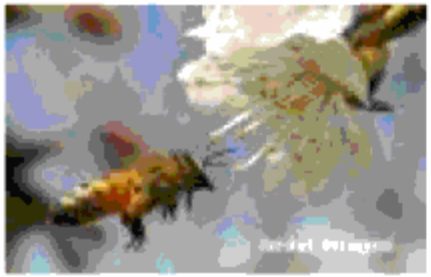
General Comparison of scatter search with Genetic Algorithms

Feature	Scatter Search	Genetic Algorithms
Population Size	Small, usually no more than 20 solutions	Large, usually not less than 100 solutions
Generation of the initial population	Heuristically, based on quality and diversity requirements	Generally at random
Update of the population	Using deterministic rules which combines diversity and quality	Using the principle of the "survival of the fittest"
Randomization	Inexistent or with a very specific character	Profusely used (selection, mutation)
Selection	Systematic	Random
Combination	2 or more solutions	2 solutions
Local Search procedures	Essential	Not used in the general implementation
Memory	Memory oriented	Uses memory in the basic sense

4 智能优化方法　蜂群算法（Bee Colony Optimization）

i) Food Sources: The value of a food source de[illegible] many factors such as its proximity to the nest, its richness or concentration of its energy, and the ease of extracting this energy. For the sake of simplicity, the "profitability" of a food source can be represented with a single quantity [8].

ii) Employed foragers: They are associated with a particular food source which they are currently exploiting or are "employed" at. They carry with them information about this particular source, its distance and direction from the nest, the profitability of the source and share this information with a certain probability.



iii) Unemployed foragers: They are continually at look out for a food source to exploit. There are two types of unemployed foragers: scouts, searching the environment surrounding the nest for new food sources and onlookers waiting in the nest and

4 智能优化方法　智能优化方法一览表

时间	名称	来源
950-1955	模式搜索	
1960-1965	随机搜索	
1975	遗传算法	
1990	文化基因算法	
1990-1995	蚁群算法	模拟蚁群觅食过程
1995	粒子群算法	鸟类和鱼类群体运动行为
2000	和声算法/蜂群算法	即兴音乐创作/蜜蜂采蜜过程
2005	人工萤火虫优化算法	萤火虫通过通过荧光进行信息交流
2009	布谷鸟算法	布谷鸟孵育行为

5. 研究存在问题

- 智能优化算法相对比较容易掌握

1. 组合优化问题相对比较简单，离专业也比较近。
2. 各位老师编程能力强,解决算法编程问题。
3. 标准实例库，测试比较规范。
4. 智能优化算法是一类优化算法，解决问题多样性。

- 存在问题

1. 做研究都有基金梦，申请国家基金需要研究基础,有高质量论文支撑。
2. 如何发高质量论文?
3. 提升复杂实际生产工艺系统提炼科学问题能力,发现创新点
4. 无高性能软硬件计算平台

精确算法：优化建模与LINGO软件
WebSphere ILOG CPLEX 的优化软件

热轧计划问题的描述

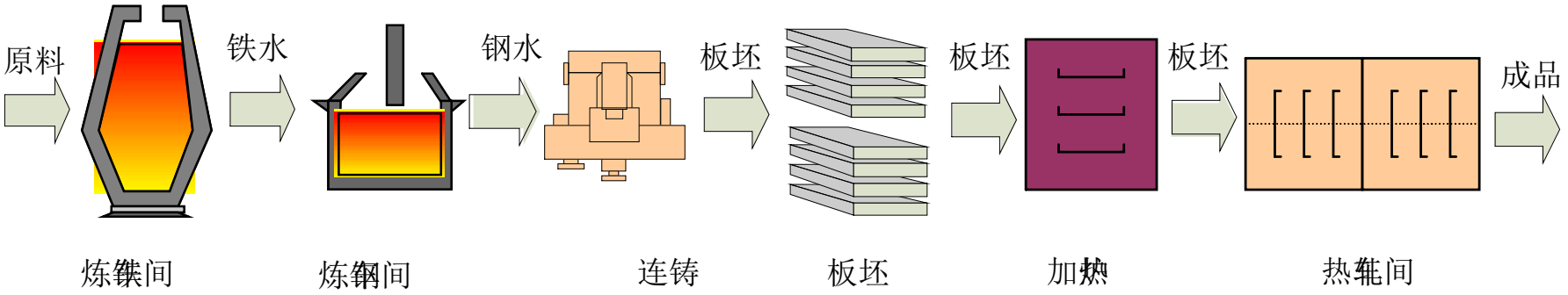


图 6.1 钢铁生产流程

- **问题:**
  定期更换工作辊
  工作辊之间为热轧计划
- **任务:**
  编排热轧计划
  确定计划内板坯顺序
- **文献:侧重理论研究**
- **本文:注重系统实现**
  详细模型与解的方法

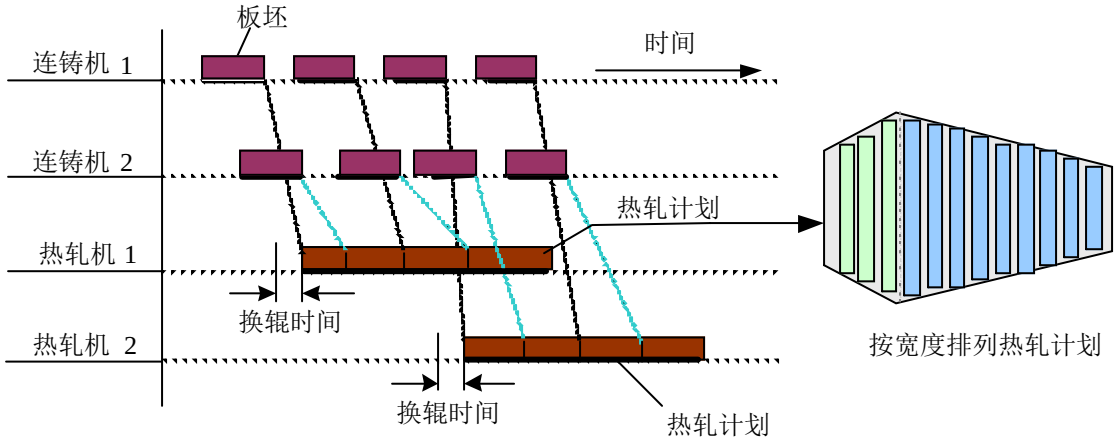


图 6.2 板坯和热轧计划的关系

Methods 74 (2015) 83–89

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

DISEASES: Text mining and data integration of disease–gene associations

Sune Pletscher-Frankild [a], Albert Pallejà [a,b], Kalliopi Tsafou [a], Janos X. Binder [c,d], Lars Juhl Jensen [a,*]

[a] Department of Disease Systems Biology, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
[b] Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
[c] Structural and Computational Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany
[d] Bioinformatics Core Facility, Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Luxembourg

ARTICLE INFO

Article history:
Received 21 January 2014
Received in revised form 15 November 2014
Accepted 25 November 2014
Available online 5 December 2014

Keywords:
Text mining
Named entity recognition
Information extraction
Data integration
Web resource

ABSTRACT

Text mining is a flexible technology that can be applied to numerous different tasks in biology and medicine. We present a system for extracting disease–gene associations from biomedical abstracts. The system consists of a highly efficient dictionary-based tagger for named entity recognition of human genes and diseases, which we combine with a scoring scheme that takes into account co-occurrences both within and between sentences. We show that this approach is able to extract half of all manually curated associations with a false positive rate of only 0.16%. Nonetheless, text mining should not stand alone, but be combined with other types of evidence. For this reason, we have developed the DISEASES resource, which integrates the results from text mining with manually curated disease–gene associations, cancer mutation data, and genome-wide association studies from existing databases. The DISEASES resource is accessible through a web interface at http://diseases.jensenlab.org/, where the text-mining software and all associations are also freely available for download.

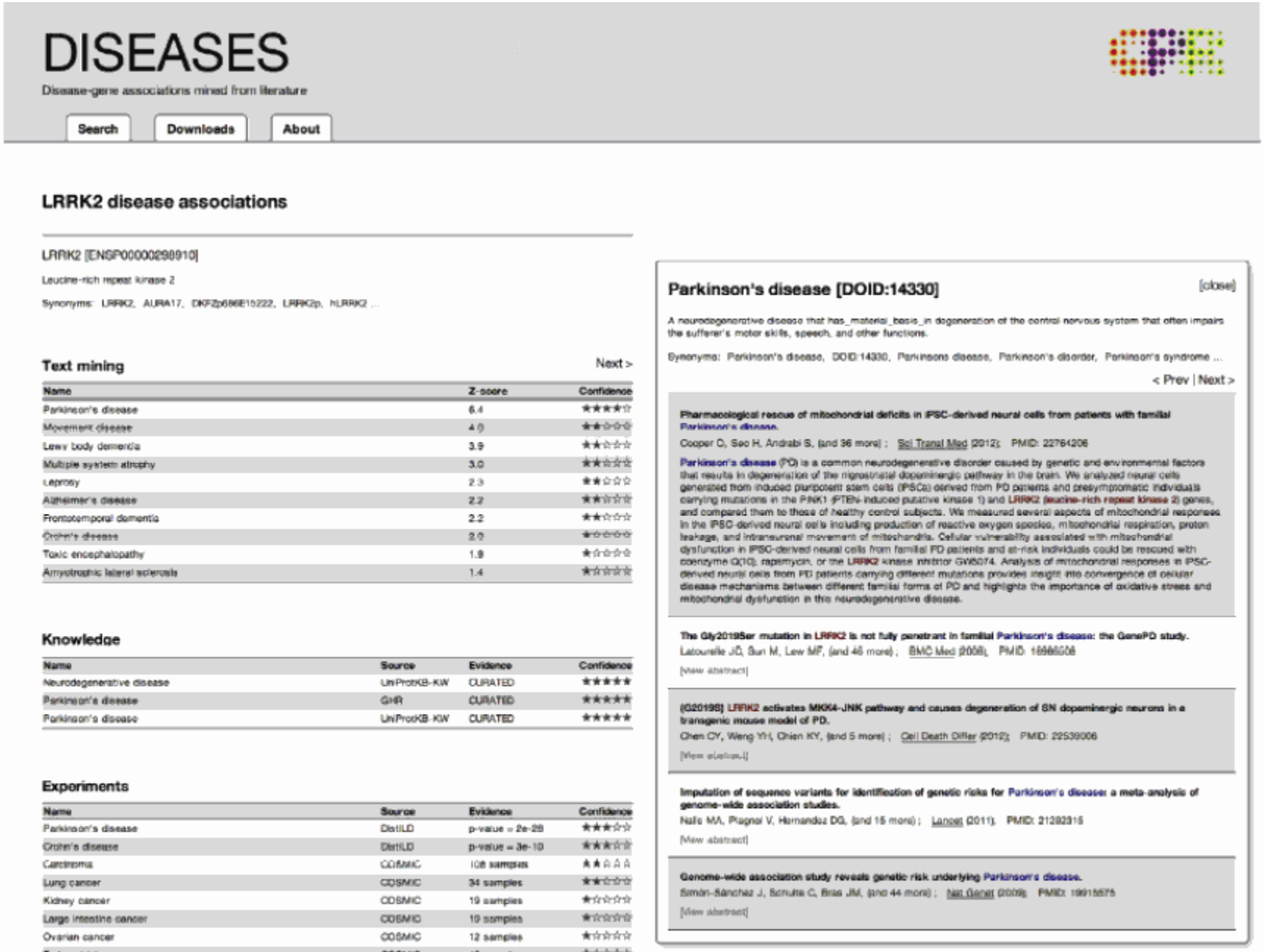


Fig. 2. The DISEASES web resource. The figure shows how the disease–gene associations are presented in the web interface, exemplified by the LRRK2 gene. The three tables provide the user with an overview of the evidence from text mining, curated knowledge, and experimental data. Clicking on an association, e.g. to Parkinson's disease, in the text mining table gives access to the underlying abstracts with the co-occurring gene and disease highlighted. The two other tables provide hyperlinks to the relevant entries

大数据,机器学习,数据挖掘关系

- 是三个平行的概念。大数据侧重描述数据，数据挖掘侧重描述应用，机器学习侧重描述方法。数据是基础，是挖掘和学习的“燃料”（深度学习像火箭，计算是引擎，数据是燃料）。

- 大数据的存储、传输、计算、处理、分析等，都是传统方式难以应对的，相关的技术就要升级，新的技术通常基于分布式架构解决，而分布式架构又带来一致性、资源调度、性能优化等多种问题。

- 数据挖掘是指从大量数据中挖掘出有价值的潜藏规律和知识。实施过程涉及机器学习、模式识别、统计学、分布式存储、分布式计算、可视化等。

- 机器学习是从数据中获取经验进而改善系统性能的一类重要方法，“学习”的意义就是求解最逼近真相的经验。数据挖掘经常需要采用机器学习方法，但目前机器学习主要是想实现某种程度的人工智

未来这些技术会
向office办公软件一样普
及

- 软件学报 *Journal of Software*,2015,26(11):3010−3025

大数据分析中的计算智能研究现状与展望

郭 平, 王 可, 罗阿理, 薛明志

1(北京理工大学 计算机学院, 北京 100081)

2(北京师范大学 图形图像与模式识别实验室)

3(中国科学院 国家天文台 光学天文重点实验室)

通讯作者: 郭平, E-mail: pguo@bit.edu.cn,

7. 大数据分析中的计算智能优化方法研究现状与展望

- 随着科学界数据量的爆炸式增长, 大数据技术和应用吸引了众多的关注. 如何分析大数据, 充分挖掘大数据的潜在价值, 成为需要深入探讨的科学问题.
- 计算智能是科学研究和工程实践中解决复杂问题的有效手段, 是人工智能和信息科学的重要研究方向,应用计算智能方法进行大数据分析具有巨大的潜力.
- 对大数据分析中的计算智能方法进行综述, 结合大数据的特征, 讨论了大数据分析中计算智能研究存在的问题和进一步的研究方向.

- 随着互联网、移动智能终端、物联网等信息与通信技术的迅猛发展, 以及计算机存储和计算能力的不断提升, 大数据时代悄然而至, 成为各界共同关注的热点. 大数据将导致人们的工作和生活方式发生巨大变革.

- 2012 年3 月, 美国政府发布“大数据研究和发展计划”.

- 2015 年初, 国务院下发《关于促进云计算创新发展培育信息产业新业态的意见》, 明确指出:将着力突破大数据挖掘分析等关键技术, 推动大数据挖掘、分析、应用和服务。

- 为了充分挖掘大数据的潜在价值, 必须解决一系列技术问题, 如数据采集、信息抽取和清理、数据集成、大规模并行计算、云计算、数据分析以及解释和部署, 需要计算机软、硬件的综合解决方案。

- 计算智能是人工智能发展的新阶段, 是受到大自然智慧启发而设计出的一类解决复杂问题方法的统称.计算智能的最大特点是不需要建立问题本身的精确(数学或逻辑)模型. 这一特点非常适合于解决大数据分析中那些由于难以建立有效的形式化模型. 近年来。

- 计算智能在图像处理、模式识别、知识获取、生物医学、等许多领域广泛应用, 取得了一系列令人鼓舞的研究成果. 同时, 大数据也给计算智能发展带来新的挑战与机遇.

- 本文结合大数据的特征, 针对大数据的分析方法, 从人工神经网络、模糊系统、演化计算和群体智能这3个方面梳理大数据环境下计算智能的相关研究, 总结大数据分析中计算智能面临的主要问题;在此基础上, 给出进一步深入研究的方向, 并阐述了数据源共享与数据密集型科学问题.

- 计算智能:人工神经网络、模糊系统、演化计算和群体智能

- **人工神经网络**

人工神经网络是一种模仿动物神经系统行为特征进行分布式并行信息处理的数学模型, 具有高度的非线性映射能力、良好的容错性、自适应能力以及分布存储等优良特性, 是一类重要的计算智能方法.

Hinton 等人的论文引发了深度学习的研究热潮. 深度学习直接从大数据中学习特征, 能够更深刻地刻画出海量数据中蕴藏的丰富信息. 计算机计算能力的提升使得训练大规模深度神经网络成为可能, 因此近几年来, 深度学习以其强大的学习能力在**图像识别、语音识别、自然语言理解**等诸多应用领域取得了令人振奋的成果.

- **演化计算和群体智能**

以遗传算法(genetic algorithm,简称GA)为代表的演化计算和以粒子群优化(particle swarm optimization,简称PSO)、蚁群优化(ant colony optimization,简称ACO)等为代表的群体智能算法是解决复杂优化问题的常用方法.这类智能算法的主要意义在于:一方面,可以快速近似求解一些难解的问题,比如NP 难问题;另一方面,还可用于约简问题的规模,从而解决那些由于数据量太大而不易解决的问题.

Hinton 等人[24]的自然语言理解[30]等诸多应用领域取得了令人振奋的成果.

- **遗传算法对医疗大数据挖掘**

数据挖掘技术在医疗大数据中的应用价值十分明显。基于遗传算法的数据挖掘技术在医疗大数据中的应用，在实际的医疗大数据挖掘中，可以对分类算法、聚类算法、实践序列和的关联规则和回归预测等方法进行应用，从而完成对医疗大数据的有效挖掘，进而获取准确的数据信息，保障医院医疗服务的质量和相关决策的效率。

结合基于遗传算法的k-means 算法优化，实现对医疗大数据的数据挖掘技术的应用。在血液系统疾病费用的统计分析中，借助优化后的算法可以完成对不同血液疾病的病例数和治疗费用等信息的获取，以达到降低医院治疗成本的效果，再为提高医疗服务质量提供有效的数据信息，推动医院医疗质量水平的提升。为医院决策提供数据信息，这是符合现代医院大数据应用的需求。

- **文本分类是数据挖掘领域中重要研究方向**

文本分类通过训练和分类两个步骤完成模型建立与问题处理，训练过程与分类过程的系统流程见图 2-1[5]。

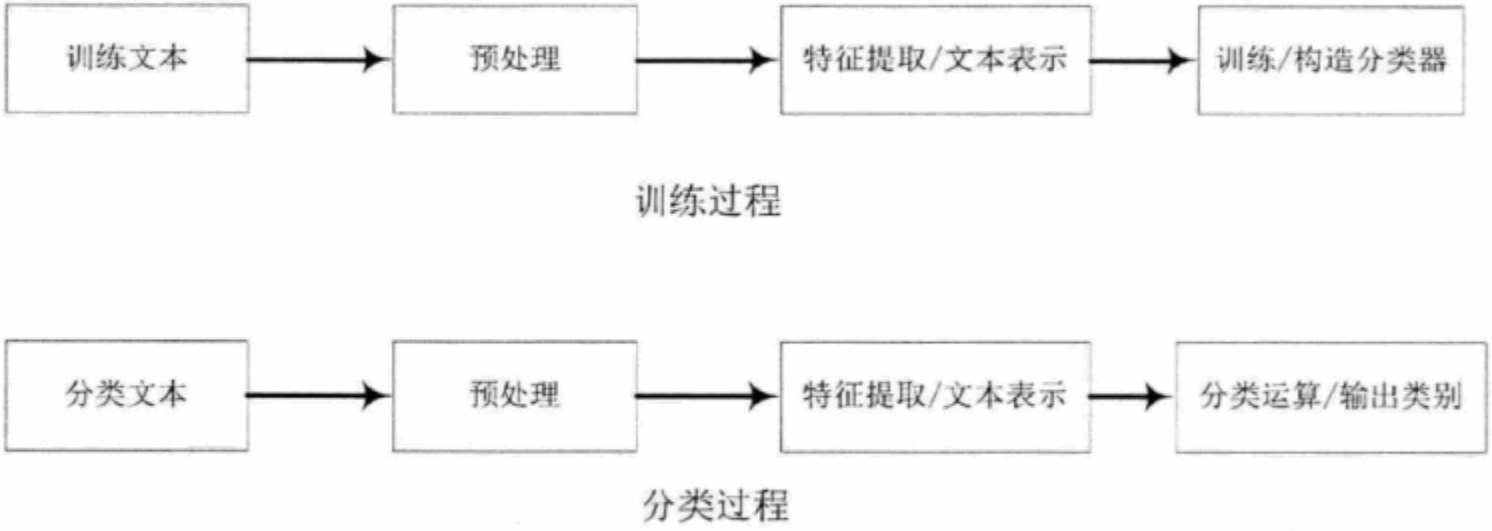


图 2-1　文本分类的一般过程

- **文本分类是数据挖掘领域中重要研究方向**

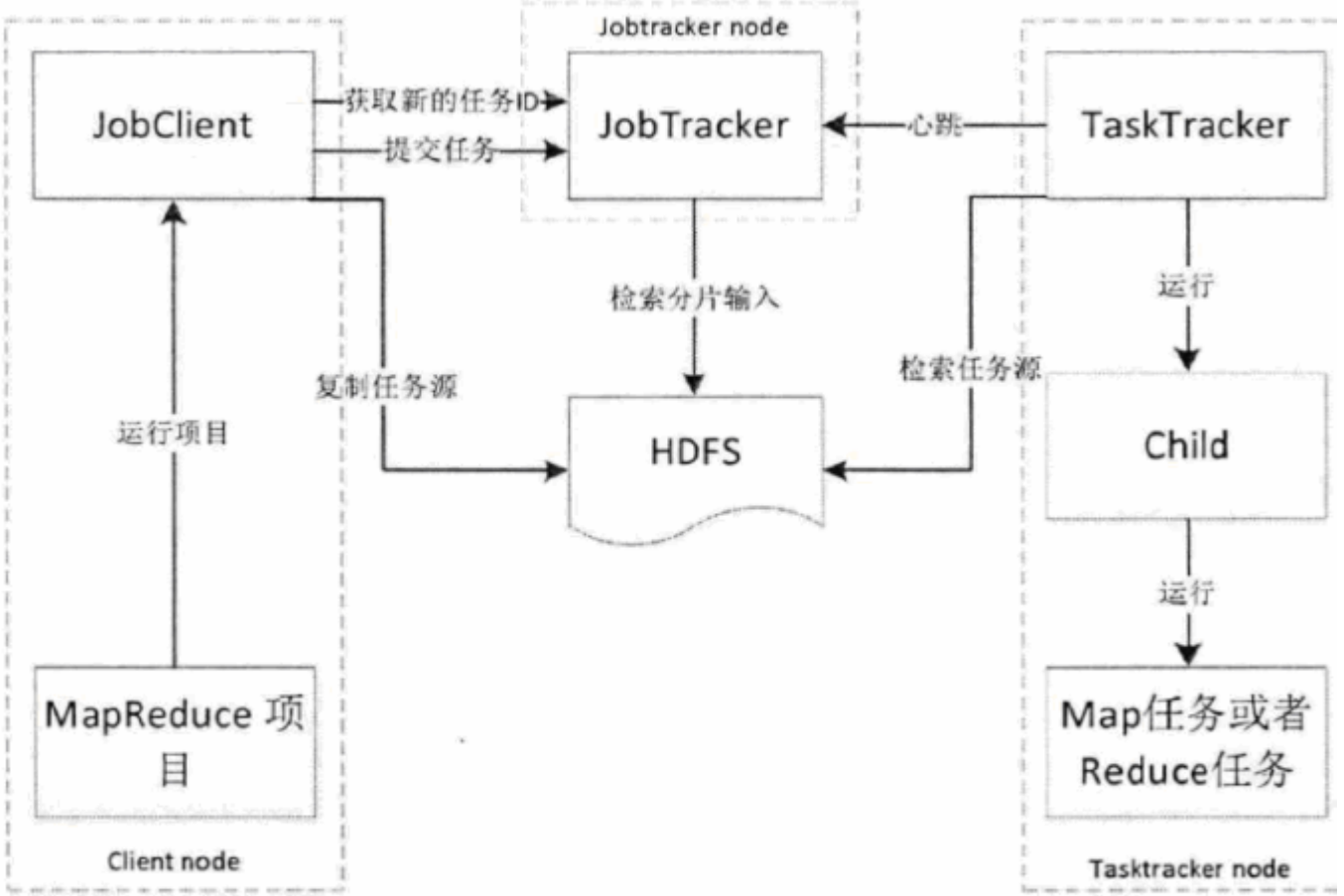
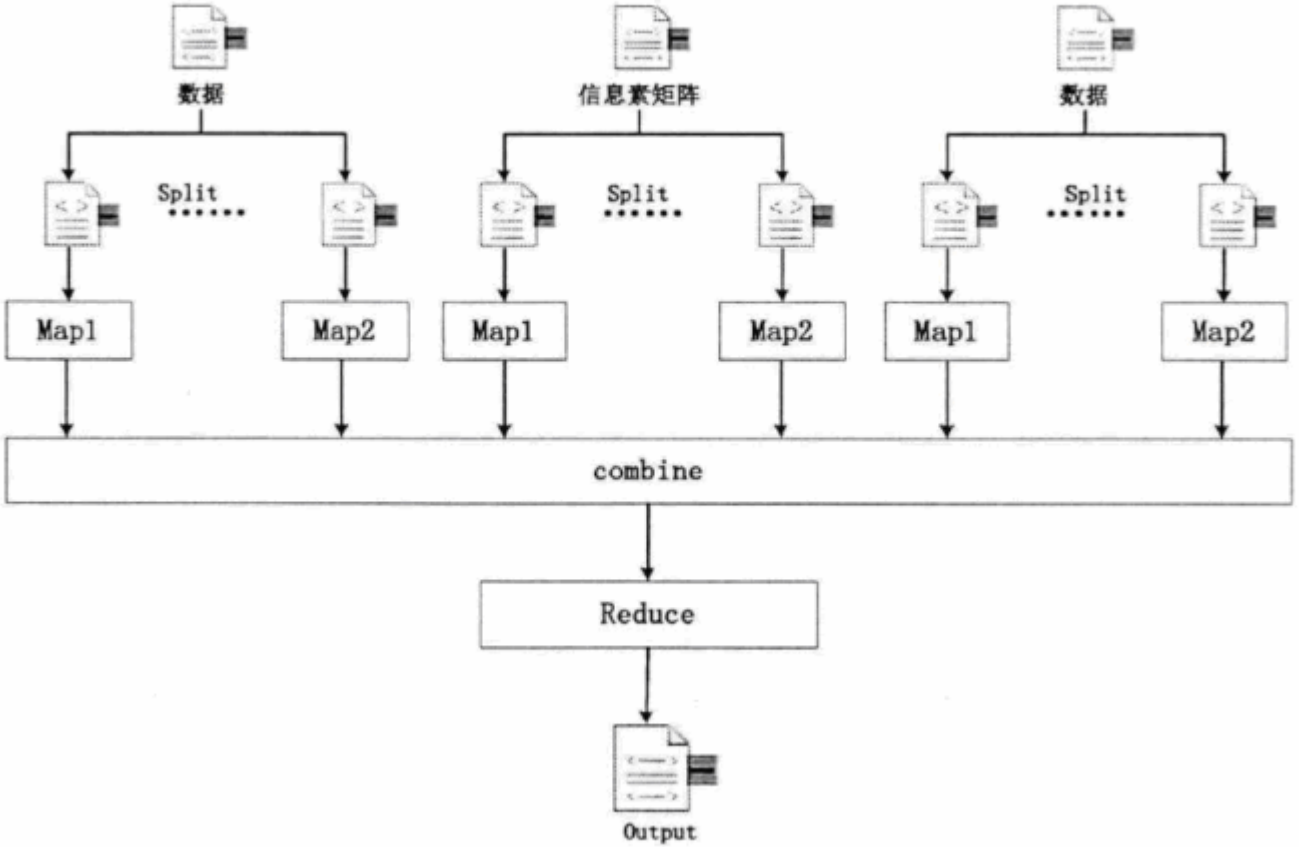


图 4-1　MapReduce 流程图

- **文本分类是数据挖掘领域中重要研究方向**

（4）Get Splits 实现方法



- **蚁群算法(AntFColonyFAlgorithm)和MapReduce框架**

引入蚁群算法(ACO)和MapReduce框架,对上述文中所提及的短文本分类算法参数优化,可以通过信息素迭代处理文本相似度进行参数比较。同时,通过蚁群算法的迭代特性优化比对短文本选择的过程,尽可能减少短文本分类过程中不必要的比对次数,以提升算法的运行效率,优化短文本分类效果

标准.

Table 1 Summary of computational intelligence methods in big data analytics

表 1 大数据分析中的计算智能方法

大数据特征及挑战	问题	方法	类别	实例	数据规模	运行环境
数据持续产生,不断变化,无法用传统的批处理方式(variability, velocity);	在线学习	感知器	人工神经网络	投票感知器[9]	70 000	单核 SGI MIPS R10000 CPU (194MHZ)
				均值感知器[11]	240 000	–
				权重多数感知器[12]	–	–
				被动主动感知器[14]	252 800 275	–
				置信度权重感知器[17]	581 012	–
				核感知器[19]	199 523	–
数据维度高,包含冗余属性(volume)	数据约简	神经网络	人工神经网络	基于网络结构约简的特征选择[21]	290	–
				FSOM[25]	41 386	–
		混合方法	人工神经网络、演化计算	SAGA[48]	10 000 维	Intel Pentium D CPU (3.40GHz);2GB RAM
		遗传算法、随机投影	演化计算	基于遗传算法的异常检测[49]	100 000	233MHZ CPU; 100MB RAM
传统的浅层学习无法揭示大数据复杂的规律,大量无标记、多模态数据难以直接利用(variety)	深度学习	深度神经网络	人工神经网络	深度神经网络[24]	804 414	–
				无监督特征学习[25]	10 000 000	1 000 台计算机(16 000 核)的集群
数据规模大,计算时间长(volume)	可扩展算法	随机采样、增量处理	模糊系统	rseFCM[32] spkFCM[32] okFCM[32]	5 000 000 000	普通 PC
		随机梯度下降、增量处理		SGFC[35]	70 000	四核 Inter I5-2400 CPU;24GB RAM
		分治、数据压缩		FAR-HD[41]	651 425	AMD Athlon X2 4200+ CPU;2GB RAM
		数据压缩		FARC-HD[42]	19 020	四核 Pentium Core 2 CPU (2.5GHz);4GB RAM
数据中存在噪声、错误及缺值(veracity)	鲁棒算法	数据预处理	模糊系统	KFCM-FSVM[36]	6 435	双核 Intel Xeon CPU (2GHZ);4GB RAM
数据维度高,算法性能退化(volume)	协同演化	分治	群体智能	CCPSO[52] CCPSO2[52] DECC-DG[54] CBCC-DG[54]	1 000 维	–
复杂数据空间多重假设(variety)	多目标优化	遗传算法、粒子群优化等	演化计算、群体智能	MODPSO[51]	–	Inter(R) Celeron(R)M CPU 520(1.6GHz), 512MB RAM

计算智能未来的研究方向

- **提高算法的可扩展性**

提高算法的可扩展性. 由于很多计算智能方法并非针对大数据分析, 也缺少这些方法在大数据分析中性能的相关报道. 将原本针对小数据集的计算智能方法移植到大数据集上, 是有待研究的重要问题. 解决这一问题常见策略为在线优化的方法、随机化算法、基于哈希的方法以及通过大规模计算集群实现分布式并行计算. 如何将这些常见策略与计算智能方法相结合, 如何发展具有高可扩展性的计算智能新方法.

计算智能未来的研究方向

- **采用分而治之的策略**

采用分而治之的策略, 实现原始问题的化大为小、化难为易、化繁为简. 分而治之是处理大规模复杂问题的直接、可行的策略, 其关键问题在于如何对问题进行**抽象和划分**,如何由子问题的解推演出全局解.

计算智能未来的研究方向

- **数据进行采样**

大数据分析中是否一定要使用原始数据集中的全部数据?如果不是, 如何采用有效手段找到和问题相关的数据子集, 对数据进行采样, 是值得研究的问题. 通过分析大数据的子集来揭示大数据中蕴含的规律.

计算智能未来的研究方向

- **在线分析优化算法**

对于无法一次性载入内存的大数据集, 每次以一定的概率分布随机选择一个样本作为输入. 针对在线算法, 如何有效地对结果进行融合;当数据分布发生变化时, 如何保证数据分析及算法的稳定性;满足实际应用中对算法实时性的要求.

计算智能未来的研究方向

- **分布式并行计算**

以MapReduce为代表的分布式并行计算模型的出现,有力地推动了大数据在产业界的应用和发展. 针对传统计算智能算法在MapReduce 框架下实现的研究, 如何设计和实现针对包括Hadoop 在内的多种平台下的计算智能算法,仍将是未来的重要研究方向.

结束语

- 大数据在带来巨大机遇的同时, 也给信息技术提出了严峻的挑战.

- 本文结合大数据的特点对大数据分析中计算智能方法的研究进行了归纳和总结, 讨论了大数据分析中计算智能存在的问题和未来的研究方向, 阐述了数据源共享问题, 并建议利用丰富和开放的天文大数据, 开展基于计算智能的大数据分析研究.

- 总之, 计算智能在大数据分析中具有巨大的应用潜力, 尽管目前已经有了一些探索性的研究工作, 但是总体上看, 针对大数据分析的计算智能方法的研究还处于起步阶段, 尚有诸多问题亟待解决.

谢谢大家！