

遗传算法在图书采购决策中的应用

林 笠, 马颖慧

(暨南大学 数学系, 广东 广州 510632)

摘要:在图书馆采购图书决策问题中,一方面为了更好地满足读者的要求,另一方面经费又有限.所以,如何决定订购图书的策略,就成为图书馆一个重要的问题.本文提出“碰集”的概念,并用遗传算法计算碰集来解决这个问题.将每位读者要求订购的书目用一个集合来表示,全部读者的要求就是一个集合簇(集合的集合),用遗传算法计算该集合簇的碰集,这些碰集就是决定购买的图书.它可以满足以下二点:若某本书有较多的读者要求订购,则该书被选购的可能性就大;若某个读者选购的图书较少,则该读者的要求被全部满足的可能性就大.并用计算机实现了文中提出的算法.

关键词:图书采购;碰集;遗传算法

中图分类号:G253.1;TP18

文献标识:A

文章编号:1007-7162(2003)03-0023-05

在图书馆的实际购书运作中,有时会因为经费不足,对读者要求采购的图书不能全部满足时,不得不进行一定的选购策略来进行筛选^[1].现有的筛选方法一般是根据每本书有多少读者要求购买的数量来决定.

策略1 图书A有X位读者要求购买图书B有Y位读者要求购买,如果 $X > Y$,则决定购买图书A,否则,购买图书B.

但是该策略有一个缺点:“热门”专业的书容易被订到,即使该书并不是很重要,而非“热门”专业的书不易被订到.“热门”专业即读者数量较多的专业.该策略会导致“热门”专业书过多,而非“热门”专业的书过少.

例1:设有《Borland C++》、《Visual C》、《C++》、《C》等书待购.由于这些书是“热门”专业的图书,每本图书分别有1000、900、15、12位读者要求购买.而《凝聚态物理》由于是非“热门”专业,只有10位读者提出购买.若按策略1,则这些书的购买顺序应该是:《Borland C++》、《Visual C》、《C++》、《C》、《凝聚态物理》,而可能由于经费不足,导致后面2本书放弃不买.但是,从以上统计数据可以看出,《C++》并不是重要的书,只是由于是“热门”专业而订到,对大多读者来讲并不重要,所有订购《C++》的读者的同时也要求订购《Borland C++》和《Visual C》,而要求订购《凝聚态物理》的读者只要求订该一本书.从分析可知,这些书的采购顺序应该是《Borland C++》、《C++》、《Visual C》、《凝聚态物理》.由此看出,策略1并不是最佳的方案,容易导致专业订书不均衡.

为了解决这个问题,下面引入集合的“碰集”策略.

1 碰集算法

定义1 碰集(文[2], Hitting set) 设C是集合簇(set cluster, 由集合作成的集合).C的一个

收稿日期:2002-03-06

基金项目:广东省自然科学基金资助项目(011162)

作者简介:林笠(1968-),男,讲师,博士,主要研究方向为人工智能、计算机图形学.

碰集为 $H \subseteq U_{i \in C}$, 使得对每一个 $s \in C$ 都有 $H \cap s \neq \emptyset$ 记为 $[\cdot]$, 称 C 的一个碰集是最小的 (MHS), 当且仅当它的任意子集都不是碰集.

策略 2 设 C 是读者提出的要求购买的图书的集合簇, H 是 C 的碰集, 则碰集中的图书, 就是可以购买的图书. 碰集即是一个能满足所有读者的方案.

该方案的特点是: (1) 满足尽可能多的读者的要求, 即使得较多的读者的要求被满足; (2) 若某本书被较多的读者的选订, 则该书被订到的可能性就较大; (3) 某读者选订图书数量较多, 则被满足的可能性也相应地变大 (买回其中至少一本的可能性较大); (4) 不论每位读者选取书数量的多少, 均有可能被 (部分) 满足.

这些特点符合一般图书馆的宗旨.

该算法的主要优点是: (1) 某读者订书的数量越多则得到满足的可能性越大; (2) 一本书被越多的读者选订, 则被买回来的可能性就越大; (3) 同时, 也兼顾订书数量比较少的读者, 即若一个读者只订一本书, 且只有该读者一个人订, 则该书也有可能买回. 这正是实际情况中我们所希望的.

现在已有多种方法来计算碰集^[2-5], 如: HS-树、BHS-树、递归计算碰集等, 这些算法的共同特点是: (1) 可以计算得到准确的碰集; (2) 这些算法的时间复杂度与空间复杂度较大; (3) 这些算法要求软件能处理“链表”等较复杂的数据结构; (4) 当集合簇的规模较大时 (一般超过 40 个), 就可能因为内存不足, 无法计算出碰集.

而我们考虑实际的应用背景, 对于图书的订购特点, 一是不要求计算出十分准确的解; 二是读者比较多, 接近 5 000 ~ 10 000 人左右, 这样, 集合簇规模就比较大; 三是一般图书管理系统均是数据库系统, 而数据库系统一般不能处理较为复杂的数据结构. 综上所述, 不宜采用“树”的算法. 所以, 下面我们介绍一种简单的遗传算法来计算碰集.

2 遗传算法计算碰集

遗传算法 (Genetic algorithm, 简记 GA)^[6-7] 是当结果不必用精确数学模型表示时, 利用 Darwin 的进化论思想“物竞天择, 适者生存”来进行计算的算法. 以下, 我们将用 GA 算法来计算集合簇碰集.

遗传算法的基本操作是:

(1) 变量编码 (Coding)

在 GA 算法中, 我们用一个一维逻辑型数组 (取值为“0”或“1”) 称为染色体 (Chromosome), 数组的每一位称为一个“基因” (Gene), 每个染色体代表一个个体, 多个个体形成一个“种群” (Population). 每个基因对应唯一的书 (是集合的元素), 若该位为“1”, 则该集合包含该元素; 为“0”, 则不包含该元素.

(2) 群体成员的繁殖 (Reproduction)

群体成员的繁殖即由上一代种群通过“繁殖”得到下一代种群. 若父代不是碰集, 则进化使之成为碰集, “繁殖”操作有“交叉”、“变异”、“倒位”等, 下面分别对它们进行解释.

(3) 个体之间的元素 (基因) 交叉 (Crossover)

假设有 2 个染色体: $S_1 = \{s_{11}, s_{12}, \dots, s_{1n}\}$, $S_2 = \{s_{21}, s_{22}, \dots, s_{2n}\}$, (n = 待购图书数量, 每个“基因”代表一本书), 随机选取一个数 $0 < r < n$, S_3, S_4 是 S_1, S_2 交叉后得到的后代,

$$S_3 = \{s_i \mid \text{if } i \leq r, s_i \in S_1, \text{ else } s_i \in S_2\}, S_4 = \{s_i \mid \text{if } i \leq r, s_i \in S_2, \text{ else } s_i \in S_1\}.$$

(4)个体成员的变异(Mutation)

假设一个染色体 $S_1 = \{s_{11}, s_{12}, \dots, s_{1n}\}$, 随机选取一个整数 $0 < r \leq n$, S_3 是 S_1 变异之后的个体,

$$S_3 = \{s_i \mid \text{if } i \neq r, s_i = s_{1i}, \text{ else } s_i = 1 - s_{1i}\}.$$

(5)个体成员的倒位(Inversion)

设一个染色体 $S_1 = \{s_{11}, s_{12}, \dots, s_{1r}, s_{1, r+1}, \dots, s_{1, r+l}, s_{1, r+l+1}, \dots, s_{1n}\}$, r, l 是 2 个随机的整数, S_2 是倒位后的结果.

$$S_2 = \{s_{11}, s_{12}, \dots, s_{1r}, s_{1, r+l+1}, \dots, s_{1, r+l}, s_{1, r+1}, \dots, s_{1n}\}$$

算法 1 遗传算法.

Step 1. 初始化: 选取 k 个个体的种群, 其中 k 是一个比较大的常数, 是读者数量的 3~4 倍, 每个个体是随机生成的待购图书.

Step 2. 测试是否满足结束条件(进化后代数量, 结果的适应度等), 如果满足则停止, 否则, 继续执行 Step 3.

Step 3. 选择: 适应度高的染色体, 即与它相交的集合的数量较大.

Step 4. 应用遗传操作: “交叉”, “变异”, “倒位”等, 产生它的子代.

Step 5. 将父代与子代种群结合在一起, 从中选择适应度“优秀”的个体, 即与较多的集合相“碰”的个体, 作为新一代种群.

Step 6. 重复 Step 2~5.

算法 1 用类程序语言如下:

```
P ← InitializePopulation();
While (not stop(P)) do
  Parents[1..4] ← SelectParents(P);
  Offspring[1] ← Crossover(Parents[1]);
  Offspring[2] ← Mutation(Parents[2]);
  Offspring[3] ← Inversion(Parents[3]);
  P ← Selection(P, Parents[1..4], Offspring[1..4]);
Endwhile.
```

利用遗传算子“交叉”, “变异”, “倒位”等, 对初始种群进行繁殖, 结果见图 1.

3 算法效率比较

我们以演示程序来比较 HS-树、BHS-树、GA 算法的效果, 见图 1.

通常情况下, GA 算法可以得到 95% 以上的碰集(如图 2), 这在图书采购系统中是完全可以接受的. 且计算时间

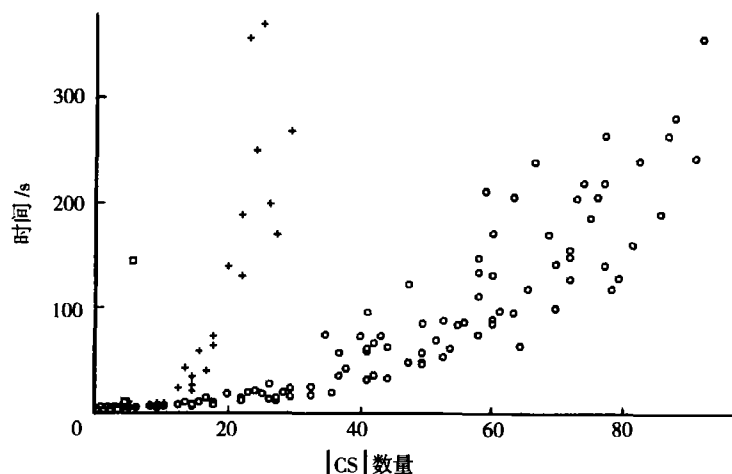


图 1 几种算法计算效率的比较

□ HS-树; + BHS-树; ○ GA

对读者数量的增加并不敏感(如图 3,4),空间的增加也可以用遗传的后代数量来进行控制.

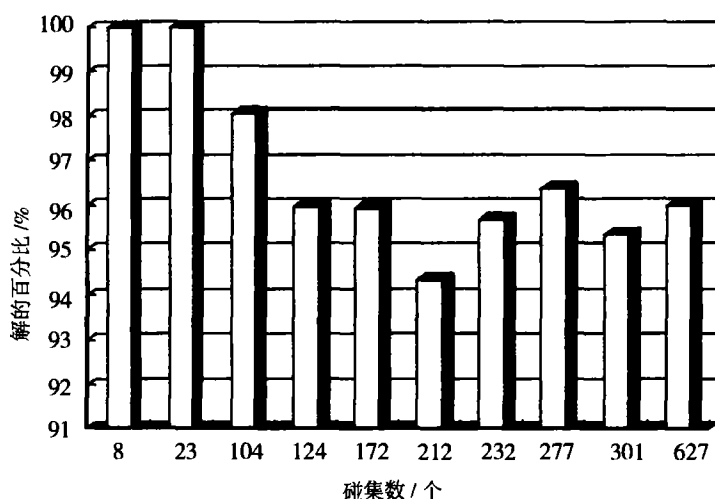


图 2 用 GA 算法可以得到的解的百分比

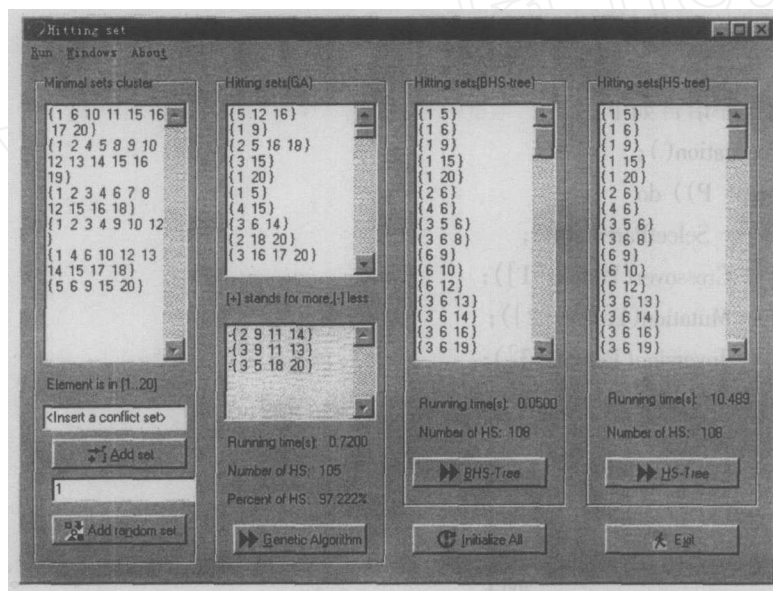


图 3 程序运行的用户界面

4 结论

碰集和遗传算法是人工智能领域十分活跃的算法之一,以前几乎没有注意到在图书馆中的应用.根据图书订购的问题特点:并不需要十分准确及算法的数据结构.我们针对性地提出“碰集”概念和用“遗传算法”来解决该类问题是比较适当的.

用碰集概念的优点是:(1)可以保证被多数读者选订的书优先被选中;(2)可以兼顾订书少的“冷门”专业的读者,这是其它算法不能满足的.

基于这 2 个特点,在资金紧张的情况下,可以用碰集来解决不同类型读者的要求.

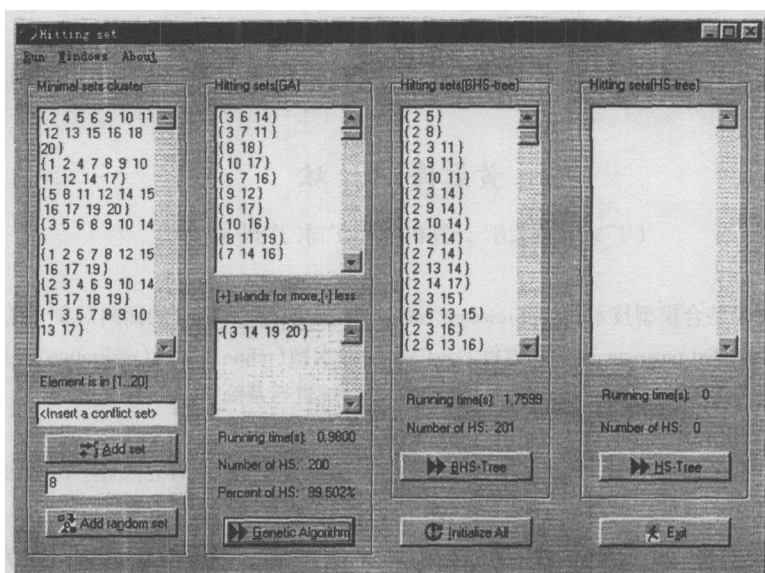


图4 程序运行的用户界面(HS树算法已不能计算碰集)

遗传算法计算碰集的优点是:(1)不需要复杂的数据结构;(2)可计算出95%以上的解;(3)算法的时间复杂度与空间复杂度均比较优。

参考文献:

- [1] 廖向群.高校图书馆采购问题及对策[J].现代图书情报技术,2000,(86):142-144.
- [2] R Reiter. A Theory of Diagnosis from First Principles[J]. Artificial Intelligence, 1987, 32(1): 57-96.
- [3] 林笠.基于模型诊断中用逻辑数组计算最小碰集[J].暨南大学学报,2002,22(1):20-24.
- [4] 林笠.递归建立HS-树计算最小碰集[J].微电子学与计算机.2002,19(2):7-10.
- [5] L Lin, Y F Jiang. Computing Minimal Hitting Sets with Genetic Algorithm[C]. In Proc. 13th International Workshop on Principles of Diagnosis. Austria. 2002. 77-80
- [6] J Holland. Adaptation in Natural and Artificial Systems [M]. Ann Arbor: The University of Michigan Press, 1975.
- [7] D Goldberg. Genetic Algorithms in Search, Optimization, and Machine Learning [M]. Reading, MA: Addison-Wesley Publishing Company, 1989.

The Application of Genetic Algorithm in Book-Purchased Decision

LIN Li, MA Ying-hui

(Dept. of Maths., Jinan University, Guangzhou 510632, China)

Abstract: In book-purchased decision, on the one hand, the readers' demands are satisfied more and more, on the other hand, the less of money is often occurred. So, the book-purchased decision becomes an important problem in a library. In this paper, both "Hitting-set" and "Genetic Algorithm" are used in this problem. Set clusters are used to represent the books that are selected by the readers; the hittingsets are the books that can be purchased. If these books are purchased then it will satisfy more readers. The hitting sets can be computed by Genetic Algorithm, which is simply and memory saved. The program was also given.

Key words: book-purchased; hitting set; genetic algorithm