第4卷第2期
1993年4月

软 件 学 报
JOURNAL OF SOFTWARE

Vol. 4, No. 2
Apr., 1993

长记录位置不变的排序算法

杨宪泽

(西南民族学院,成都 610041)

THE SORT ALGORITHM OF UNCHANGEABLE LONG-RECORDS PLACE

Yang Xianze

(*Southwest Nationalities College, chengdu* 610041)

Abstract Over the years many sort algorithms have been produced, and the time complexity of unchangeable long-records place algorithms still is $O(N^2)$. It can't satisfy needs of large-scale information treatment. On the basis of Refs. [1,2], the new algorithm is mentioned in this note to enhance the sort speed. In this algorithm, the keywords is mapped to array element subscript and we don't carry out two operations of comparision and exchanges of keywords. Its time complexity is $O(N)$, and the algorithm is appropriate to widely used large-scale information treatment in the future.

摘要 现有的排序算法,长记录位置不变算法时间复杂性还是$O(N^2)$,不能满足大规模信息处理需要.本文在文献[1,2]基础上,提出了一个提高排序速度的新算法.这一算法关键字与数组下标作映射处理,不实施反复比较和交换关键字的操作,时间复杂性达到$O(N)$,适宜今后在大规模信息处理中广泛应用.

§1. 引 言

近来,在文献[1,2]研究基础上,我们构造了信息记录位置不变的一个新排序算法.这一算法采用映射方式,排序过程中不反复比较信息关键字,记录位置不变,只改变逻辑链接位置,并已证明,其时间复杂性优于Hoare快速法,为$O(N)$.

众所周知,随着计算机信息管理自动化的飞速发展,事务工作处理的信息量越来越大.这些信息,有的记录长度很长,含有多个分量.如我国高考成绩记录,7门课程,加上考号、姓名、总分共10个分量;全面质量管理评估记录,所含分量达30多个.这些信息经处理后,都有记录排序名次问题.按一般方法进行排序,不仅排序方法不稳定,而且移动记录所花时间要多于关键字比较所花时间,使平均排序时间大于$O(N^2)$;按链接线性插入排序和计数选择排序,虽能

本文1990年9月25日收到,1991年2月1日定稿.本课题由国家民委科研基金资助.作者杨宪泽,副教授,1982年毕业于西南石油学院,主要研究领域为数据结构、知识工程.

保证记录位置不变，排序方法稳定，但平均排序时间是$O(N^2)$，不能满足大规模信息处理的需要[3]. 如dBASE Ⅲ虽在信息管理中发挥着越来越大的作用，但人们普遍感到排序速度慢. 因此，如何改进长记录位置不变的排序算法，是当今面临的一个较重要课题[4,5]. 本文提出一个新的长记录位置不变的排序算法，在附加一定存贮空间下，排序时间复杂性为$O(N)$，可望在这类问题的排序中广泛应用，满足大规模信息处理需要.

§2. 算法描述与实现

2.1 算法描述

该算法考虑日常事务工作处理的大量信息一般情况：

(1)信息关键字用十进制数$K_i(i=1,2,\cdots,N)$表示，若与实际情况不符，需作转换或在算法中作相应修正.

(2)关键字是正整数，若含小数和负数按文献[1]中方式处理扩充.

(3)关键字最大值可以预知，最小值为0，若不能做到这一点，按文献[1]方法处理.

(4)关键字值不太长，且分布均匀. 若值太大或分布不均，算法初始采用文献[2]子域映射方式.

算法的特点是信息记录不动，按关键字值以映射关系基本排定初步位置.

映射是这样的关系：构造一个数组P，使得关键字值对应P数组元素下标. 如关键值为50，它与数组P(50)对应；为500，则与P(500)相对应.

对于相同关键字的处理，算法附加了三个数组空间：每一记录的链指针R(i)，链首指针空间Q，链当前指针空间W.

关键字值K_i与P数组元素下标映射关系仅有一次时，$P(K_i)=1$；这时$Q(K_i)\leftarrow i$，记录了具有这唯一对应关系K_i所在信息记录的地址i，并作为最后排序调整位置的首地址. $W(K_i)\leftarrow i$为出现相同关键字提供链接地址作准备. 见图1.

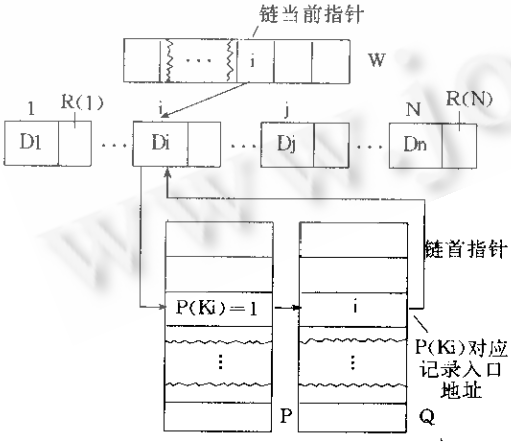

图1

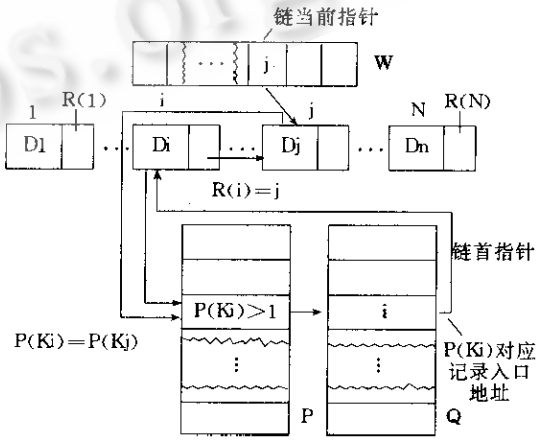

图2

映射时出现相同关键字，如 $K_j=K_i$，这时 $P(K_j)>1$，这时将把 K_i 和 K_j 两个信息记录链接起来，入口地址仍是 $Q(K_i)i$，但 $R(W(K_i))\leftarrow j$，相当于 $R(i)=j$；此外，$W(K_j)\leftarrow j$，为链接出现多个相同关键字作准备. 见图 2.

实施映射和链接处理后，最后根据 P 数组下标值的有序性，P＝0 不实施操作，P≥1 从相对应的 Q 数组下标值作为入口地址调整一次记录位置即完成排序.

算法构造基点：

A1：给定 N 个待排信息记录，含有分量 $D1_i$，$D2_i$，…，Dm_i（m 为分量个数），确定某一分量为关键字，记为 K_i.

A2：开辟链指针空间 R，容量 N；映射记数空间 P，链首指针空间 Q，链当前指针 W，容量均为 K_{max}（K_{max}是关键字最大值）；赋初值 i＝1.

A3：输入 K_i，让 $P(K_i)\leftarrow P(K_i)+1$，即完成映射工作，记录相同关键字出现个数.

A4：若 $P(K_i)=1$，作 $W(K_i)\leftarrow i$ 和 $Q(K_i)\leftarrow i$，转 A6.

A5：若 $P(K_i)>1$，作 $R(W(K_i))\leftarrow i$ 和 $W(K_i)\leftarrow i$.

A6：$i\leftarrow i+1$，直至 i＝N 为止，实施 A3～A5.

A7：(Z＝1)，从 $J=K_{max}$开始，若 P(J)＝0，转 A8；P(J)≠0，作递减排序：

(1)$T\leftarrow Q(J)$；链首指针送 T.

(2)输出 D1(T)，D2(T)，…，Dm(T).

(3)$Z\leftarrow Z+1$，若 $Z\neq P(J)$，$T\leftarrow R(T)$后转(2)；否则，转 A8.

A8：$J\leftarrow J-1$，实施 A7，直至 J＝0 结束.

2.2 算法编程框图

本文给出类 BASIC 语言实现的较详细的算法编程框图（见图 3），以帮助读者深入考察算法，减少实际应用障碍. 这一流程也可供用 PASCAL 和其它语言实现该算法时参考.

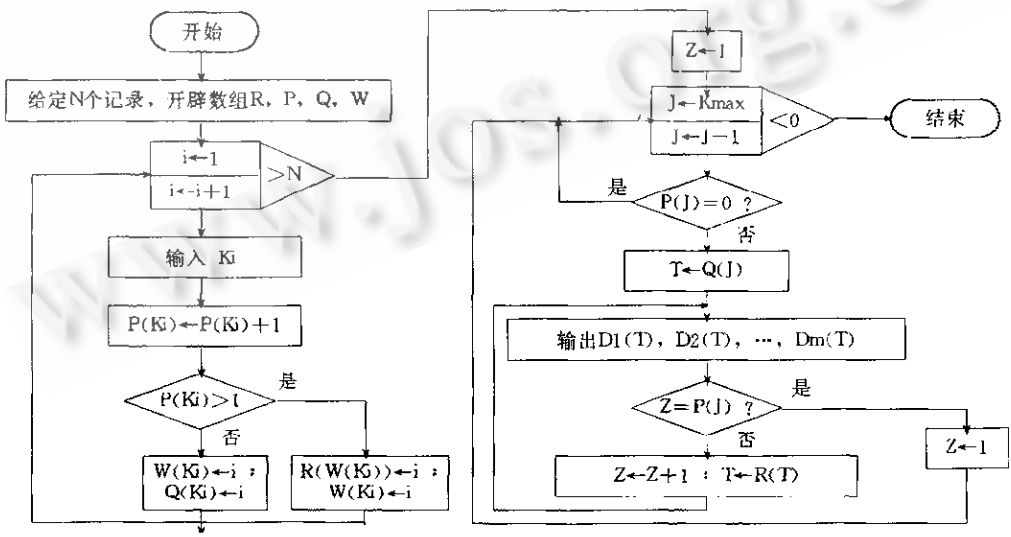

图 3

§3. 算法分析

3.1 空间复杂性

该算法附加的存贮空间是链指针空间R,链首指针空间Q,链当前指针空间W,记数空间P,合计 $3K_{max}+N$. 由于 K_{max} 在一般常见问题中不大,如高考总分 $K_{max}=650$,而大规模信息处理中N很大,所以可以认为附加存贮空间小于2N. 如果N很小,计算机存贮空间富裕,附加量不会带来问题.

3.2 算法稳定性

该算法是确定待排信息关键字的映射关系,然后构造调整信息记录链接指针,排序过程中信息记录位置不变,且关键字也很少移动,不出现关键字之间反复比较和交换操作,因此算法是稳定的.

3.3 时间复杂性

该算法所需时间分析如下:

设 T_C 为两数比较所需时间,T_C 为传送一个数据所需时间,T_A 为数加1所需时间.

因为A3中所需时间:$T_3=NT_L+NT_A$;A4中所需时间:$T_4\leqslant NT_C+(N+2K_{max})T_L$;A5中所需时间:$T_5\leqslant NT_C+3NT_L$;A6中所需时间:$T_6=NT_A+NT_C$;A7中所需时间:$T_7<3K_{max}T_C+3NT_L+NT_A$;A8中所需时间:$T_8=K_{max}T_A+K_{max}T_C$. 而 $T=T_3+T_4+T_5+T_6+T_7+T_8$,所以 $T<(8N+2K_{max})T_L+(3N+4K_{max})T_C+(3N+K_{max})T_A$.

在大规模信息处理中,假定 $K_{max}<N$,有

$$T<10NT_L+7NT_C+4NT_A \tag{1}$$

从文献[3]可知,Hoare快速排序法时间复杂性 T_H 介于 $O(N\log_2 N)\sim O(N^2)$ 之间,作为期望时间

$$T_H\leqslant 1.4N\log_2 NT_C+1.4N\log_2 NT_L+1.4N\log_2 NT_A \tag{2}$$

为比较(1)和(2)式,我们按机器指令执行时间关系[6],有 $T_C\approx 1.5T_L$,$T_A\approx 0.75T_L$.

这样

$$T<23NT_L \tag{3}$$

$$T_H<4.55N\log_2 NT_L \tag{4}$$

从(3)和(4)式分析可知,由于T的常数因子不大,该算法N大于64时就优于Hoare算法. 实际上,该算法已经达到时间复杂性下界O(N),只要不出现 $K_{max}\gg N$ 的情况,这一结论不会有问题.

§4. 实验结果

为了验证算法的效率,我们选用链接线性插入排序法和Hoare快速排序法与之对比,采用PASCAL语言在IBM-PC上用机器产生五个分量的随机数据做了实验,实验结果见表1.

此外,我们利用此算法采用BASIC调dBASE Ⅲ文件记录方式进行排序,以90年西藏高考成绩进行处理实验,结果是这种方式比直接用dBASE Ⅲ中排序命令排序效率高七倍以上.

这一算法为什么有高效率,其原因是它不实施对信息关键字反复比较和交换两种操作,附加了一定的存贮空间换取速度,以映射关系完成排序. 实验结果除了证实算法理论的正确性外,还预示了广泛的应用前景.

表1　排序方法的时间对比(单位:S)

数据N	A算法	Hoare法	链接线性插入法
5000	5.54	31.22	103.59
7000	8.76	36.67	151.24
9000	13.11	57.39	183.11
11000	15.99	79.12	208.95

作者对我院鄢德英讲师在算法采用BASIC调dBASEⅢ文件记录的排序实验中所做大量工作深表感谢！同时向审稿者对本文初稿提出的宝贵意见致谢！

参考文献

[1]杨宪泽,分级快速排序法研究,科学通报,34:11(1989),871-873.

[2]杨宪泽,子域映射快速排序法研究,科学通报,35:15(1990),1199-1200.

[3]王本颜,方慧昌,数据结构技术,清华大学出版社,1988.

[4]Mc Culloch, C. M, Quickshunt-A Distributive Sorting Algorithm, Computer Journal 25 : 1 (1982), 102-104.

[5]Barstow D. R, Remarks on A Syntheis of Serveral Sorting Algorithms, Actu Information 13 : 3 (1980), 225-227.

[6]张怀莲,宏汇编语言程序设计,电子工业出版社,1989.

四通4S高级科技编排系统简介

四通4S(Super Science Setting System)高级科技书刊编排系统,是专门为科技类书刊文献的编排而设计的,排版功能实用性很强,采用即打即排的操作界面,直观形象,从根本上解决了科技书刊排版的难题。

4S系统具有文字、图片、数学、化学、乐谱、表格、杂志、造字等直观排版功能,可在24针打印机、不同精度激光印机字及激光照排机上输出与版式一致的样张,目前可提供包括宋、楷、黑、仿宋、魏碑、隶书、中圆、细圆、标宋等丰富的汉字字体。

四通4S先后荣获第37届尤里卡世界发明博览金奖、北京首届国际博览会金奖、广州第二届国际专利技术及新产品展金奖、美国纽约第14届国际发明展金奖、北京市发明展金奖、科技进步一等奖等20多项大奖,获中国发明专利和美国发明专利授权。

四通4S系统自1987年7月问世以来,受到广大用户的欢迎,取得了令人满意的使用效果。为了使4S系统更加完善,北京四通集团公司于1990年6月在珠海创办了开发生产基地,先后推出了4S—KNM、4S—9218等性能价格比极优的排版机,及国内一流的向量汉卡及激光照排机。该公司拥有一支经验丰富的技术开发队伍及遍布全国各地的销售、服务网点,可为广大用户提供优质服务。