基于信息处理的映射排序算法

杨宪泽 (西南民族学院)

摘要:本文以事务管理信息系统为基础,提出了一种映射排序算法。该算法的特点是把记录关键字值映射于数组下标,用记数方式反映关键字值情况,数组元素下标自然把关键字值一次定好了位置,这样,可以不实施反复比较与交换操作。这种映射排序算法与比较交换排序法相比,有较高的效率,适宜在计算机大规模信息处理中广泛采用。

一、引 言

随着计算机的普及,计算机被用于各行各业的事务处理工作中,数据处理,情报资料整理,企业管理等都必须检索,排序可以提高检索的效率。因此,众多领域中许多信息记录的无序序列需要调整成有序序列,致使现今计算机系统中花费在排序上的时间占系统 CPU 运行时间的较大比重。

由于排序技术的重要性,研究各种有效的排序算法成为软件工作者必不可少的重要课题。迄今为止,有代表性的排序算法已提出几十种,但就其全面性能而言,很难提出一种被认为是最好的算法,每一种算法都有各自的优缺点,适合不同的环境使用。映射排序算法亦属分布型排序法,最早是 1956 年由 Isaac,E。J 等提出,称"地址计算排序"[1],近几年,这种算法的研究有许多进展,并投入应用,其原因是:

1.随着我国计算机在事务处理应用中的普及,需要处理的信息量越来越大,采用好的排序算法成为这方面提高系统运行速度的关键技术之一。

2.计算机在事务处理中的某些应用具有特殊性,即关键字值是正整数,且最大值与最小值之差不是很大。如高考分数排序,中文词组作关键字的排序,竞赛名次排序,全面质量管理评估得分排序等,使用映射排序方法可以获得高效率。

3.微电子技术飞速发展,计算机内存容量大大增加,从而使这种算法以牺牲部分存储开销换取速度不会带来问题。

二、基本映射排序算法

1、定义与规定

定义 1 假设含 N 个记录的序列为$\{R_1,R_2...,R_N\}$,对其任定的关键字 $K_1, K_2, ..., K_N$ 的排序$(K_1ER_1,K_2ER_2,...,K_NER_N)$,使 $R_1,R_2, ...,R_N$ 成为一个特定的序列,这就是排序。

定义 2 设 K,L 是两个集合,对于任意 Ki (i=1, 2,..., N),有 $B(K_i)$与 L_i 对应,L_iEL, 则称 K 到 L 内的一个映射。

定义 3 设 B 是集合 K 到集合 L 的映射,E 是符合 K 到 C 的映射,对于任意 K_iEK,规定

$(E.B)(K_i)=E(BCK_i))$

我们称此映射为映射 B 与映射 E 的乘积,映射乘积是一种特殊映射。

规定 1 关键字=记录

这种情况在实际信息处理中少见,具有特殊性,仅作为基本映射排序算法设计之用。一般情况下,关键字≠记录,记录含有多个字段。

规定 2 关键字是十进制正整数。

这种情况在实际信息处理中常见,若不符,可实施转换(见最后讨论)。

规定 3 关键字最大值 Kmax≤N(N 为处理的记录个数)。

这一规定在一类信息处理中可以得到满足,如竞赛总分 100,竞赛人数可能多于 100 人;高考总分 700,高考人数远大于 700 等等。

2.映射排序法基本思想

给定 $K_1,K_2,。..K_N$,可知最大值 Kmax,最小值

© 中国科学院软件研究所 http://www.c-s-a.org.cn

Kmin,那么,开辟一个数组,即可把关键字送入关键字值与 B 数组下标相等的对应元素中。显然,$K_i=50$,它对应 B(50);Kj=500,对应 B(500),相同关键字落在同一数组元素中,用计数方式可知有几个。由于数组元素的下标是有序的,500>50,数组元素的下标自然把关键字一次定好了位置,最后只要按规定的方式调数组非零元素,相同元素按计数值次数调动,排序即完成。

3.算法设计

步骤 1:N 个记录 R(1),R(2),...R(N)已读入内存,最大值 Rmax。

步骤 2:[初值 i=1]一次扫描 R(i),让 T←R(i),P(T)←P(T)+1;以映射关系确定 R(i)的位置,记录相同 R(i)的个数。

步骤 3:i←i+1,直至 i=N,重复步骤 2。

步骤 4:[初值 j=Rmax(递减排序),K=1]二次扫描

(1)若 P(j)=0,转步骤 5;

(2)若 P(j)=1,传送数据 R(K)←j,K←K+1,转步骤 5;

(3)若 P(j)>1,传送数据 R(K)←j,P(j)←P(j)−1,K←K+1,此时,如果 P(j)≠0,仍执行(3)。

步骤 5:j←j−1,直至 j=0,重复步骤 4。

4.算法分析

时间复杂性:步骤 1 至步骤 3 需时间 O(N),步骤 4 至步骤 5 需时间也是 O(N),所以这一算法时间复杂性是 O(N)。

空间复杂性:这一算法附加的存储空间主要是 P 数组,为 Rmax,按规定 3,此算法附加存储空间<N。

三、以名次对应记录的映射排序算法

1.基本思想

这种情况是,信息记录不动,以排列的名次去对应记录。这是实际信息处理工作中经常遇到的,如高考成绩排序

考号	姓名	政治	语文	数学	...	外语	总分	名次
1	...	...	...	...	...	...	...	...
2	...	...	...	...	...	...	...	...
:	:	:	:	:	:	:	:	:
N	...	...	...	...	...	...	...	...

显然,这种情况下考号 1,2 所获名次不一定是 1,2 名,将根据他(她)所得分数确定。将要设计的算法规定,相同分数者名次并列。若有两个第 1 名,就没有第 2 名,第 3 名有 5 个,就没有 4,5,6,7 名,以此类推。

2.算法设计

步骤 1:N 个记录 $R_1,R_2,...,R_NR_i(i=1,2,...,N)$含有 Y 个字段 D1i D2i,...,Dri,其中某一字段为关键字,记为 K(i),关键字最大值 Kmax。

步骤 2:[初值 i=1]一次扫描 K(i),让 T←K(i),P(T)←P(T)+1。

步骤 3:i←i+1,直至 i=N,重复步骤 2。

步骤 4:[初值 j=Kmax,F(j)←1]对计数器 P(j)扫描

(1)若 P(j)=0,转步骤 5。

(2)F(j−1)←F(j)+P(j);确定名次,显然 F(Kmax)为第 1 名,而 F(Kmax−1)不存在即没有,否则,将根据第 1 名有几个才能确定名次,以此类推。

步骤 5:j←j−1,直到 j=1,重复步骤 4。

步骤 6:[初值 i=1]最后一次扫描,输出 Ri,即输出:D1(i),D2(i),。..,F(Kci)).F(Kci))为 Ri 的名次。

步骤 7:i←i+1,直至 i=N,重复步骤 6。

3.算法分析

时间复杂性:步骤 1 至步骤 3 需时间 O(N);步骤 4 至步骤 5 最多需时间 O(N);步骤 6 至步骤 7 需时间 O(N),因此,这一算法时间复杂性是 O(N)。

空间复杂性:这一算法附加存储空间主要是 P 数组和 F 数组,为 2Kmax。按规定 3,此算法附加存储空间<2N。

四、多字段记录的映射排序算法

1.基本思想

一般情况下,关键字≠信息记录。因此,基本映射排序算法适用范围有很大局限性。如果仅考虑关键字=信息记录情况,关键字在多次移动后(记录没有一起移动),不再与记录有一一对应关系,这样的算法简化了排序任务[8]。

日常事务工作所处理的信息,有的记录长度很长,含有多个字段,如高考成绩记录,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。

这些信息记录按一般方法排序,由于多次比较关键字大小而交换记录,不仅排序算法不稳定,而且移动记录所花时间要多于关键字比较所花时间。

本节排序思想是,信息记录不动,按关键字值以映射

© 中国科学院软件研究所 http://www.c-s-a.org.cn

关系作一次扫描基本确定位置;二次扫描按关键字最大值到最小值的计数个数,统计记录 R_1—R_N 确切位置;最后一次扫描移动记录到位。

2.算法设计

步骤 1:N 个记录 R_1,R_2。..,R_N 已读入内存,Ri(i=1,2,...,N)含有 r 个字段 $D1_i,D2_i,Dr_i$ 其中,某一字段为关键字,记为 K(i),关键字最大值为 Kmax。

步骤 2:[初值 i'=1]一次扫描 K(i),让 T←K(i),P(T)←P(T)+1。

步骤 3:i←i+1,直至 i=N,重复步骤 2。

步骤 4:[初值 j=Kmax,F(j)←1]对计数器 P(j)扫描

(1)若 P(j)=0,转步骤 5。

(2)F(j)←F(j)+P(j);确定记录入口,显然,F(Kmax)对应的记录位置为第一,而 F(Kmax-1)不存在即没有,否则,将根据第一的记录有几个才能确定 F(Kmax-1)对应记录入口,以此类推。

步骤 6:[初值 i=1,最后一次扫描,移动记录 Ri 到位]T←K(i)(关键字映射值),T←F(T))(记录位置入口),Q1(T)←D1(i),Q2(T)←D2(i),...,Qr(T)←Dr(i).F(T)←F(T)+1(有相同关键字时依次排列,入口位置加 1)。

步骤 7:i←i+1,直至 i=N,重复步骤 6。

3.算法分析

时间复杂性:步骤 1 至步骤 3 需时间 O(N);步骤 4 至步骤 5 最多需时间 O(N);步骤 6 至步骤 7 需时间 O(N),因此,这一算法的时间复杂性是 O(N)。

空间复杂性:这一算法主要附加存储开销是 P 数组和 F 数组空间,它们均为 Kmax,按规定 3,它们之和≤2N。

五、多字段记录的链式映射排序算法

1.基本思想

上节算法处理多字段记录,采用了统计相同关键字值的个数,计算记录应到的位置的做法。本节提出的算法,借助于基数排序法组桶的思想,落入每一桶的关键字为相同关键字,采用链接方式把它们链接起来。显然,这时还需要两个指针,一是链首指针,提供桶的入口地址;二是链当前指针,为进入桶里的当前关键字提供链接地址。

L(i): i=1,2,..,N。每一记录的链指针。

Q(j): j=1,2,...,Kmax.桶的入口地址(链首指针)。

W(j): j=1,2,...,Kmax,链当前指针。

关键字值 K(i)与 P 数组元素下标映射的关系仅有一次时,P(K(i))=1;这时 Q(K(i))←i,记录了具有这唯一对应关系 K(i)所在信息记录的地址(也提供了第 i 个记录进入第 K(i)桶),并作为最后排序调整位置的首地址。W(K(i))←i,为这一桶出现相同关键字提供链接地址准备。

映射时出现相同关键字,如 K(j)=K(i),这时 P(Kcj))>1,将把 K(i)和 K(j)对应的两个信息记录链接起来,入口地址仍是 Q(K(i))←i,但有 L(W(K(i)))←j,相当于 L(i)←j;此外,W(K(j))←j,为链接一个桶里进入多个相同关键字作准备。

2.算法设计

步骤 1:N 个待排序信息记录 $R_1,R_2,...,R_N$ 已读入内存,R_i(i=1,2,...,N)含有 r 个字段 $D1_i,D2_i,...,D_{ri}$,其中某一字段为关键字,记为 K(i),关键字最大值 Kmax。

步骤 2:[初值 i=1]输入 K(i),让 T←K(i),P(T)←P(T)+1,即完成映射工作,记录相同关键字个数。

步骤 3:若 P(T)=1,作 W(T)←i 和 Q(T)←i,转步骤 5。

步骤 4:若 P(T)>1,作 b←W(T),L(b)←i 和 W(T)←i。

步骤 5: i←i+1,直至 i=N 为止,实施步骤 2~4。

步骤 6: [z=1,a=1]从 J=Kmax 开始,若 P(T)=0 转步骤 7;若 P(T)≠0 作递减排序

(1)T←Q(J);链首指针送 T。

(2)让 S1(a)←D1(T),S2(a)←D2(T),...,

Sr(a)←Dr(T);a←a+1。

(3)z←z+1,若 z≤P(J),T←L(T)后转(2);否则转步骤 7。

步骤 7: J←J-1,Z=1,实施步骤 6,至 J=0 为止。

3.算法分析

空间复杂性:这一算法主要需链指针数组 L,链首指针数组 Q,链当前指针数组 W,记数数组 P,合计 3Kmax+N。因此,这一算法适宜 Kmax《N 的问题中使用,如高考总分 Kmax=700,而 N(高考人数)远大于 700。

时间复杂性:可以证明,这一算法需时间与第四节算法为同一数量级,即 O(N)(本文略)。

六、总 结

1.映射排序算法效率很高,因为算法的基本特征是空间换取时间,不实施反复比较记录关键字而交换的操作,一次扫描以映射关系确定基本确定记录位置。

© 中国科学院软件研究所 http://www.c-s-a.org.cn

2.映射排序算法采用确定特排信息关键字的映射关系,或者再调整信息记录的链接指针,整个过程信息记录位置不变,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.映射排序算法在关键字值分布极不均匀的情况下效率下降,这主要是牺牲很多时间判断 P(L)=0。因此,映射排序算法的意义在于一类特殊问题的应用,而这类特殊问题在信息处理中大量存在,即要求 Kmax≤N。

4.本文所述算法全部经过应用实践,效果良好。

5.实施映射排序算法若关键字值为负或者 Kmin≠0 又较大,可实施变换 $K_i=K-Kmin$。例如某排序要处理的关键字最大值 Kmax=13000,Kmin=10000,使用这种变换映射数组元素只需 3000 个,若按基本方式实施映射需数组元素 13000 个,而且有 10000 个是无用零元素,不仅占用空间,而且占用时间来区别零元素;又如某排序要处理的关键字 Kmin=−100,若数组下标不能为负,采用变换形式可使数组下标变为 0～Kmax+|Kmin|个元素(如果 Kmax 也为负,处理可采用全部关键字都乘−1 进行)。

如果关键字值是小数,或者要分析关键字值具体情况,把记录直接送入基本正确的位置,也可实施变换:$K_i=aK+b$(a,b 为大于 1 的正整数)。

如果关键字是字符,可按文献[3]方式把字符换算成 ASSII 码十进制数组。

参考文献:

[1]Isacc.E.J,and Singleton,R.C,"Sorting by Address Calculation,J.ACM,No.3,PP.169−174,1956。

[2]杨宪泽,研究排序算法应该注意的一个问题,软件报,1990,7.14。

[3]杨宪泽,直接映射式字符检索算法,中文信息学报,第 5 卷,1991 年 3 期。

© 中国科学院软件研究所 http://www.c-s-a.org.cn