DOI:10.13657/j.cnki.gxkxyxb.1982.01.005

散列排序算法

广西计算中心　　张正铀

摘　要

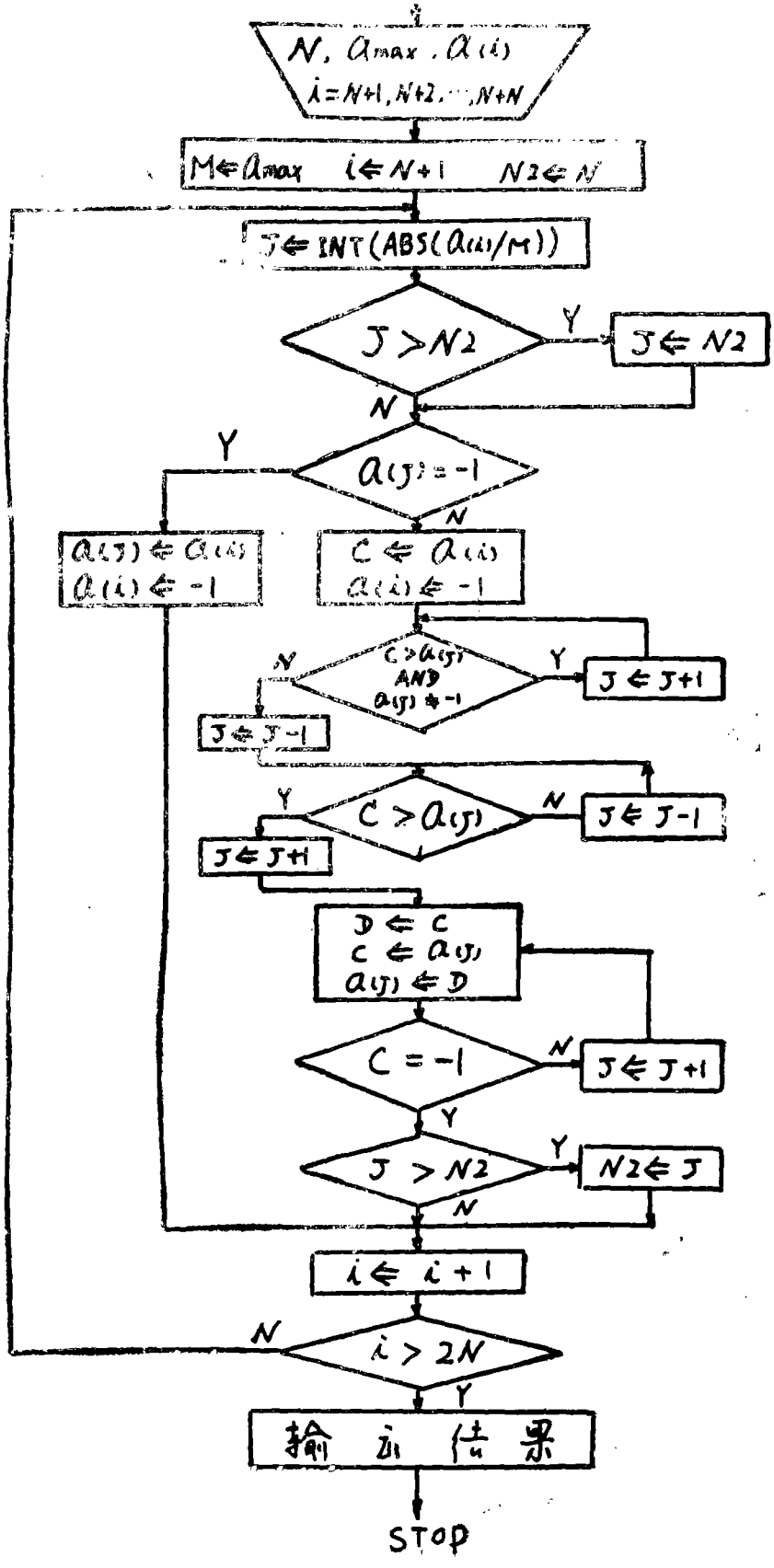
本文认为在排序算法中，决定每个数据在新序列中位置的是它的数值大小。基于这种思想，本文介绍了利用散列函数构造的一种算法复杂性为O（N）的排序算法。

从对目前所发表的各类排序算法的分析可知，每个数据在新序中的位置，只通过该数据与其它数据比较（或再利用其它数据比较后的信息）才能决定。如对N个数据作排序处理，若用“挑选插入”、“气泡漂浮法”〔Ⅰ、Ⅱ〕等算法，则元素a（i）（i=1,2，…N）必须逐个与其余N－1个元素比较后，才能决定其在新序的位置。显然，这些方法要经过N*（N－1）／2次比较后才能完成排序工作。故其算法时间复杂性为T（N）=O（N^2）。若用“快速分类”〔Ⅲ〕、“合并分类”〔Ⅱ〕、“树型排序”〔Ⅱ〕及“跳跃对分排序”〔Ⅳ〕等算法处理，基本是将原始数据分成若干组后两两比较形成短序列，再将成功的短序进行比较，合并成较长的序，重复此过程直至结束。由于这类算法除了使用数据间的比较外，还注意使用其它数据比较后得到的信息，所以有效地改善了算法的复杂性，达到T（N）=O（$N\log_2 N$）。然而，以上各类算法考虑的重点是数据间的比较，以便决定相应的位置，因而未能使算法复杂性突破O（$N\log_2 N$）。

我们认为，对数据作排序处理，决定数据在新序中位置的决定因素只是其本身值的大小。因此，新序中元素位置是待处理数据值的一个映射。若映射函数J定义为：它产生唯一的一组整数，均有J（a(i)）≠J（a(j)），其中i≠j。按此定义的J将a（i）（i=1,2…，N）落位于相应的位置，则只要作N次求J的计算，然后作N次赋值即可完成全部数据的排序。然而，按此定义的函数，J（a(i)）的取值幅度通常太大，与之对应的工作单元就太大。再则由于原始数据中经常有a（i）=a（j），其中i≠j，这时J的决定也就更复杂。于是，我们考虑选择一个恰当的散列函数，再以线性位移的方式解决“冲突”（或称“溢出”）。此外，为节省工作空间，当处理N个数据时，取2N体积的数组，用a（N＋1），a（N＋2），……a（N＋N）存放待处理的数据，每处理完一个数据a（N＋i），即将此存贮单元置空供新序使用。新序从a（1）起存放数据，从而降低装载率，减少冲突次数，这样，就可以有效地减少比较次数和工作单元、改善算法复杂性。基于此思想，我们构造了散列插入排序算法。

一、算法框图

本文1982年7月21日收到


N, Amax, A(I)
I=N+1, N+2, ..., N+N
M⇐Amax I⇐N+1 N2⇐N
J⇐INT(ABS(A(I)/M))
J>N2
Y
J⇐N2
N
A(J)=-1
Y
N
A(J)⇐A(I)
A(I)⇐-1
C⇐A(I)
A(I)⇐-1
C>A(J) AND A(J)≠-1
N
Y
J⇐J+1
J⇐J-1
C>A(J)
Y
N
J⇐J-1
J⇐J+1
D⇐C
C⇐A(J)
A(J)⇐D
C=-1
N
J⇐J+1
Y
J>N2
Y
N2⇐J
N
I⇐I+1
I>2N
N
Y
輸 出 結 果
STOP

二、实现算法的程序

```
105 INPUT "NUMBER, MAX=" ; N, M
107 N1=N*2 : M=INT(M/N*1000)/1000
110 DIM A(N1)
120 FOR I=1 TO N
130 A(I+N)=INT(RND(X)*1000)
140 A(I)=-1
145 PRINT A(I+N);
150 NEXT I
155 PRINT
160 TI$="000000" :  N2=N
165 FOR I=N+1 TO N1
170 J=INT(ABS(A(I)/M))+1
175 IF J>N2 THEN J=N2
250 IF A(J)=-1 AND A(I)>=0 THEN A(J)=A(I): A
(I)=-1 : GOTO 365
280 C=A(I) : A(I)=-1
300 IF C>A(J) AND A(J)<>-1 THEN J=J+1:G=G+1:QOLO 300
305 J=J-1
310 FOR LL=J TO 1 STEP -1
320 IF C>A(LL) THEN J=LL+1 : GOTO 340
325 G=G+1
330 NEXT LL
335 J=1
340 D=C : C=A(J) : A(J)=D
345 IF C=-1 THEN 360
350 J=J+1 :  GOTO 340
360 IF J>N2 THEN N2=J
365 NEXT I
370 PRINT "TIME" ; TI$ : T$=TI$ : C=-1E30
380 FOR I=1 TO N2
385 IF A(I)=-1 THEN 400
390 PRINT A(I); :  K=K+1
395 IF C>A(I) THEN PRINT "ERROR!!" ; C;      !A(I):STOP
397 C=A(I)
400 NEXT I
```

```
402  PRINT
405  OPEN1, 4
410  PRINT : PRINT#1, "O. K.  TIME=" ; T$, "K=" ; K, "M=" ; M, "G=" ; G
420  PRINT#1 : CLOSE1 :  K=0 : G=0
```

三、 算 法 分 析

定义：设对N个数据作排序处理，令数据的取值上界为a_{max}，则相应的散列函数为：

$$J(a(i)) = \lfloor |a(i)/(a_{max}/N)| \rfloor + 1$$

根据J（设N2为新序当前大于、等于N的下标，若$J>N_2$，则取$J=N_2$）可找到对应的a（J）。若a（J）为空（如定义此时为－1），则直接将a（i）（i=N+1，N+2，…，N+N）存于a（J），并将a（i）置空。若产生“冲突”，亦即在a(1)，a(2)，…a（J）、…，a（N_2）序列中，a（J）非空。当a(i)>a（J），则a(i)应在a（J）之右侧（设序列最大元素方向为右侧）的a（J+k）（k=1，2…，N_2-J）位。要求a（J+k）满足或a（J+k－1<a(i)≤a（J+k）或a（J+k）为空且a（J+k－1）≤a（i）。若a（J+k）为空单元，则a（J+k）⇐a（i）即可，否则将a（J+k），a（J+k+1），…，a（N_2）右移一位：

1° C⇐a(i)；

2° D⇐a（J+k）；a（J+k）⇐C；C⇐D；

3° 若C=－1，则结束右移工作；

4° k=k+1，转2°。

当a(i)≤a(J)，则a(i)应在a(J)之左侧的a(J－k+1)(k=1，2，…，J)位。要求a(J－k+1）满足a（J－k）<a(i)≤a（J－k+1），则a（J－k+1），a（J－k+2），…，a(N_2)按上述1°～4°规则右移一位。

从散列函数定义可知，若a(i)>a(k)，则J（a(i)）>J（a(k)）。a(i)必然插在a(k)的右侧。所以根据J，就可较快地将所有a(i)排到新序中。由于J比较准确地指出了a（i）的位置，无论N值如何，即使产生“冲突”，也只需与J位的左（右）有限个元素作比较，即可将a(i)定位了。所以，为解决冲突进行比较的次数与N基本无关。而N对算法时间复杂性的影响一是判断J单元是否为空，该比较次数为N；二是解决冲突时的线性位移次数；（由于赋值时间远小于比较时间[V]，对各种方法的赋值时间均未进行分析，所以，在此也从略）。三是冲突次数。由于我们定义存贮空间为总数据量的二倍，其装载率$\beta=N/(2\cdot N)=50\%$。则总“冲突”量一般为1.5N次。对67组随机产生的数排序的结果，其“冲突”次数最多为2.238N，而最少为0.4N，平均产生1.363N次“冲突”（详见附表Ⅰ）。

因此，该算法最好情况是“冲突”次数为零，即整个排序仅需进行N次判别a（J）是否为空的操作。一般情况下“冲突”次数为1.5N，即除作N次a（J）空否的判别外，还需作1.5N次处理“冲突”的判别。则此数法的时间复杂性$T(N)\doteq 2.5N=O(N)$；其空间复杂性$S(N)=2N=O(N)$。

用该算法对191组随机产生的正数排序，其结果与“挑选排序”、“跳跃对分排序”处理结果均列于附表Ⅱ。从表中可以看出，实际排序时间明显比$O(N^2)$、$O(N\log_2 N)$型算法减少。

若数据中包含少量负数，J的定义不变，效果基本与全部为正数相同。（见附表Ⅱ）若数据全部为负数，只需在求J或输出结果时略加变换即可。

若负数量较大，则可定义：

$$J=\begin{cases} \lfloor a(i)/amax/(N-F)\rfloor+F+1 & a(i)\geqslant 0 \\ \lfloor |a(i)/amax/F+F|\rfloor+1 & a(i)<0 \end{cases}$$

其中F为予估负数个数。

在这种情况下，需判别a(i)是否为负，即增加N次比较，其余不变，所以算法复杂性$T(N)=3.5N$。

定义中的F仅为估计值，若其与实际值有偏差也基本不增加算法的复杂性（见附表Ⅲ的试验结果）。

参加本算法分析及实验的还有刘连芳同志。

参考文献：

〔Ⅰ〕D·E·Knuth, “THE ART OF COMPUTER PROGRAMMING” (Volume 3/Sorting and Searching) P1～329, 1973

〔Ⅱ〕ARNE THESEN, “COMPUTER METHODS IN OPERATIONS RESEARCH” P39～58, 1978

〔Ⅲ〕姚天顺，“数据结构”，P175～186

〔Ⅵ〕张正铀，“排序算法的优化”

附表Ⅰ　冲突率试验结果（67例）

NUMBER	TIME	G	RATE(G/N)
100	13″	40	.4
100	11″	71	.71
100	16″	74	.74
100	13″	83	.83
100	22″	122	1.22
100	14″	128	1.28
200	30″	200	1
200	34″	157	.785
200	50″	277	1.385
300	1′02″	337	1.12333333
300	1′27″	419	1.39666667
400	1′53″	391	.9775
400	1′16″	401	1.0025
500	1′31″	520	1.04

500	1′54″	658	1.316
600	2′26″	664	1.10666667
600	2′26″	753	1.255
700	2′47″	901	1.28714286
700	4′01″	1028	1.46857143
800	2′50″	672	.84
800	4′22″	1334	1.6675
900	5′20″	1380	1.53333333
900	5′44″	1593	1.77
1000	5,41″	1229	1.229
1000	4′31″	1079	1.079
1000	4′53″	1051	1.051
1000	5′33″	1536	1.536
1000	5′46″	1598	1.598
1000	5′34″	1524	1.524
1100	7′40″	1483	1.34818182
1100	6′15″	1525	1.38636364
1200	6′10″	1421	1.18416667
1200	11′12″	2443	2.03583333
1300	7′02″	1378	1.06
1300	10′16″	2146	1.65076923
1400	5′39″	1448	1.03428571
1400	8′55″	1955	1.39642857
1500	6′43″	1931	1.28733333
1500	11′37″	2354	1.56933333
1600	9′52″	2257	1.410625
1600	16′18″	3582	2.23875
1700	9′47″	2376	1.39764706
1700	13′16″	2839	1.67
1700	19′00″	3214	1.89058824
1800	12′04″	2304	1.28
1800	10′00″	2423	1.34611111
1800	15′48″	2660	1.47777778
1900	11′38″	2507	1.31947368
1900	12′53″	2676	1.40842105
2000	13′50″	3009	1.5045
2000	16′35″	3619	1.8095
2100	12′09″	2790	1.32857143
2200	15′46″	3241	1.47318182
2300	13′42″	3021	1.31347826

2400	17′20″	4046	1.68583333
2500	19′55″	4309	1.7236
2600	17′57″	3736	1.43692308
2700	20′17″	4003	1.48259259
2700	21′17″	4811	1.78185185
2800	15′55″	3865	1.38035714
2800	18′37″	3943	1.40821429
2900	24′13″	4145	1.42931034
2900	18′14″	4789	1.65137931
3000	31′33″	4861	1.62033333
3000	30′51″	5701	1.90033333
3050	22″03″	4829	1.58327869
3050	26′21″	3908	1.28131148
		AVERAGE=1.36325156	

* G为“冲突”次数

附表Ⅱ　各算法试验及其比较结果

NO.	N	散列排序	对分跳跃	挑选插入	NO.	N	散列排序	对分跳跃	挑选插入
1	100	18″	52″	1′16″	22	200	34″	2′20″	5′
2	100	17″	1′10″	1′17″	23	300	1′16″	5′01″	11′21″
3	100	24″	1′05″	1′20″	24	300	1′08″	5′08″	11′23″
4	100	21″	1′	1′16″	25	300	1′50″	5′05″	11′39″
5	100	17″	1′	1′17″	26	300	57″	5′01″	11′22″
6	100	13″	47″	1′18″	27	300	1″13″	5′02″	11′29″
7	100	17″	44″	1′18″	28	300	1′27″		
8	100	13″	48″	1′24″	29	300	1′02″		
9	100	14″	42″	1′17″	30	400	1′45″	8′26″	20′16″
10	100	11″	45″	1′18″	31	400	1′15″	8′37″	20′09″
11	100	22″			32	400	1′25″	8′27″	20′27″
12	100	16″			33	400	1′50″	8′38″	20′47″
13	100	13″			34	400	2′08″	8′39″	20′40″
14	200	37″	2′25″	5′06″	35	400	1′16″	8′33″	20′18″
15	200	46″	2′30″	5′06″	36	400	1′53″		
16	200	50″	2′29″	5′12″	37	500	2′36″	13′01″	32′07″
17	200	47″	2′38″	5′13″	38	500	2′43″	13′26″	32′19″
18	200	30″	2′33″	5′13″	39	500	2′18″		
19	200	37″	2′21″	5′03″	40	500	2′15″		
20	200	50″	2′36″	5′19″	41	500	2′17″		
21	200	30″	2′29″	5′11″	42	500	2′15″		

NO、	N	散列排序	对分跳跃	挑选插入	NO、	N	散列排序	对分跳跃	挑选插入
43	500	3′06″			80	900	4′46″		
44	500	3′31″			81	900	3′13″		
45	500	2′03″			82	900	4′35″		
46	500	2′11″			83	900	6′09″		
47	500	1′31″			84	900	5′44″		
48	500	1′54″			85	900	5′02″		
49	500	2′15″			86	900	5′43″		
50	500	2′33″			87	900	7′38″		
51	500	2′12″			88	900	5′56″		
52	600	2′57″	17′58″	45′16″	89	1000	6′37″	47′53″	2：2′46″
53	600	3′			90	1000	5′41″		
54	600	3′07″			91	1000	5′14″		
55	600	3′23″			92	1000	7′32″		
56	600	4′37″			93	1000	5′30″		
57	600	2,26″			94	1000	5′34″		
58	600	2′26″			95	1000	5′41″		
59	700	4′37″			96	1000	8′06″		
60	700	2′52″			97	1000	6′07″		
61	700	3′09″			98	1000	5′39″		
62	700	4′24″			99	1000	5′46″		
63	700	3′36″			100	1000	4′31″		
64	700	4′01″			101	1000	5′33″		
65	700	2′47″			102	1000	4′53″		
66	700	4′04″			103	1100	8′06″		
67	700	3′27″			104	1100	4′26″		
68	700	3′01″			105	1100	4′54″		
69	800	2′41″	32′40″	1：20′2″	106	1100	5′53″		
70	800	4′14″	32′09″	1：20′2″	107	1100	6′15″		
71	800	4′20″			108	1100	7′40″		
72	800	3′37″			109	1100	11′40″		
73	800	3′49″			110	1100	12′55″		
74	800	2′50″			111	1200	9′04″		
75	800	4′22″			112	1200	5′12″		
76	800	6′30″			113	1200	11′12″		
77	800	4′29″			114	1200	6′10″		
78	800	3′55″			115	1200	8′28″		
79	900	3′29″	40′07″	1：41′58″	116	1200	9′49″		

NO.	N	散列排序	对分跳跃	挑选插入	NO.	N	散列排序	对分跳跃	挑选插入
117	1200	8′59″			155	2100	12′09″		
118	1200	7′34″			156	2200	23′34″		
119	1300	7′02″			157	2200	15′46″		
120	1300	10′16″			158	2200	15′02″		
121	1400	5′39″			159	2200	15′46″		
122	1400	8′55″			160	2300	13′42″		
123	1500	9′23″	1：41′25″	4：31′26″	161	2400	17′20″		
124	1500	11′40″	1：45′20″	4：21′29″	162	2500	15′48″		
125	1500	17′13″			163	2500	23′10″		
126	1500	11′37″			164	2500	29′15″		
127	1500	6′43″			165	2500	16′34″		
128	1600	9′52″			166	2500	17′21″		
129	1600	16′18″			167	2500	19′55″		
130	1700	9′50″			168	2600	17′57″		
131	1700	14′52″			169	2700	20′17″		
132	1700	14′56″			170	2700	21′48″		
133	1700	13′16″			171	2800	15′55″		
134	1700	9′47″			172	2800	18′37″		
135	1700	19′			173	2900	24′13″		
136	1800	12′56″			174	2900	18′14″		
137	1800	13′07″			175	3000	29′54″		
138	1800	13′07″			176	3000	38′26″		
139	1800	12′04″			177	3000	46′24″		
140	1800	10′			178	3000	22′41″		
141	1800	15′48″			179	3000	20′28″		
142	1900	12′53″			180	3000	41′24″		
143	1900	11′38″			181	3000	31′33″		
144	2000	16′02″			182	3000	30′51″		
145	2000	11′32″			183	3000	51′51″		
146	2000	18′32″			184	3000	40′25″		
147	2000	13′50″			185	3000	31′14″		
148	2000	16′35″			186	3050	21′14″		
149	2000	12′27″			187	3050	28′32″		
150	2000	18′02″			188	3050	25′25″		
151	2000	15′34″			189	3050	42′30″		
152	2100	14′21″			190	3050	22′03″		
153	2100	18′03″			191	3050	26′21″		
154	2100	15′19″							

附表Ⅱ 包含负数的算法试验结果

NO.	N	时间	予估负数率	实际负数率	NO.	N	时间	予估负数率	实际负数率
1	100	13″	8%	8%					
2	100	14″	8%	8%	13	200	46″	10%	10%
3	100	14″	10%	10%	14	200	46″	10%	10%
4	100	14″	10%	10%	15	200	49″	15%	15%
5	100	17″	30%	30%	16	200	53″	20%	20%
6	100	23″	20%	30%	17	300	58″	3%	3%
7	100	15″	20%	30%	18	300	57″	3%	3%
8	100	32″	40%	50%	19	300	1′08″	3%	3%
9	100	28″	40%	50%	20	300	1′17″	20%	20%
10	100	34″	40%	50%	21	500	2′	10%	10%
11	100	27″	40%	50%	22	500	1′55″	10%	10%
12	100	36″	50%	60%	23	500	3″	20%	20%