(Information Processing Letters
Vol. 15, No. 5, pp. 214~219,
1982.12.)

A Note on Upper Bounds for Selection Problem

by Tatsuya MOTOKI

Department of Information Science
Faculty of Engineering
Ibaraki University
4-12-1 Nakanarusawa
Hitachi 316
JAPAN

Keywords :

Minimum comparison selection,

worst-case analysis,

optimum sorting,

computational complexity

1. Introduction

According to Knuth[6], the history of selection problem goes back to Rev.C.L.Dodgson[2] who pointed out in 1883 that the second best player often loses the second prize in lawn-tennis tournaments; about 1930 Hugo Steinhaus posed the problem of finding the minimum number of tennis matches required to select the first- and second-best players from n contestants, assuming a transitive ranking; and now the Steinhaus problem is generalized to our selection problem of finding the worst-case minimum number of comparisons $V_i(n)$ required to select the i-th largest from n distinct numbers. By symmetry, we have

$$V_i(n)=V_{n-i+1}(n).$$

Thus we may assume that $1\le i\le\lceil\frac{n}{2}\rceil$. Throughout this paper $\log_2$ is denoted by lg.

Many people have made efforts to give good upper/lower bounds for the problem. As for lower bounds, Kirkpatrick[5] unified the theory; he showed that

$$V_i(n)\ge\begin{cases}\left\lfloor\frac{3n+i+1}{2}\right\rfloor & \text{if } \frac{n}{3}<i\le\frac{n+1}{2}\\ n+i-3+\sum_{0\le j\le i-2}\left\lceil\lg\frac{n-i+2}{i+j}\right\rceil & \text{if } i\le\frac{n}{3}\end{cases} \qquad (1)$$

and that the result surpasses other results. See Kirkpatrick[5] for a detailed account of lower bound theory.

As for upper bounds, the classical paper Hadian and Sobel[3] showed that

$$V_i(n)\le n-i+(i-1)\lceil\lg(n-i+2)\rceil\ . \qquad (2)$$

Several refinements(e.g. Hyafil[4], Yap[8]) were done by constructing variants of the Hadian-Sobel algorithm, but these are all essentially small improvements on (2); the Hadian-Sobel algorithm and its variants

need $O(n\lg n)$ comparisons when $i\sim\frac{n}{2}$. Until 1972 it was not known whether the selection problem inherently needs $O(n\lg n)$ comparisons; finally, Blum et al.[1] obtained the $O(n)$ upper bound; and further study, due to Schönhage et al.[7], led to a much sharper upper bound for $i=\lceil\frac{n}{2}\rceil$, i.e.

$$V_{\lceil n/2\rceil}(n)\leq 3n+o(n).$$

Since Schönhage et al.'s scheme can be easily generalized for general values of i, we obtain

$$V_i(n)\leq 3n+o(n) \quad \text{for every } i. \tag{3}$$

Let us now consider a question: For what values of i (3) can be asymptotically surpassed ? Blum et al.[1] considered the similar question for their $(\frac{391}{72}n+o(n))$-algorithm and obtained a result

$$\lim_{n\to\infty}\sup\frac{V_{\lfloor q(n-1)\rfloor+1}(n)}{n}\leq 1+\frac{319}{72}\frac{q}{p}+\frac{391}{36}\lceil\lg\frac{p}{q}\rceil q \quad \text{for } 0<q\leq p \tag{4}$$

where $p=0.203688^-$. Thus their $(\frac{391}{72}n+o(n))$-algorithm can be surpassed for $i<pn$. How about the case of Schönhage et al.'s ? After a rough comparison of (2) and (3), the Hadian-Sobel algorithm is seen to give a better upper bound than Schönhage et al.'s only for very small values of i, i.e. for $i<\frac{2n}{\lg n}$. But considering the blum et al.'s result, this ought to be considerably improved.

In this note, we generalize (4) by applying the Blum et al.'s scheme to a general $(cn+o(n))$-algorithm, and show that the generalized result can be surpassed when we apply the scheme to the generalized Schönhage et al.'s $(3n+o(n))$-algorithm. In addition, we show that there exists an asymptotically optimal selection algorithm provided that $i=o(n)$. Explicitly speaking, our results are:

(i) If there exists a $(cn+o(n))$-algorithm for selection,

$$\lim_{n\to\infty}\sup\frac{V_{\lfloor q(n-1)\rfloor+1}(n)}{n}\leq\begin{cases}1 & \text{if } q=0\\ 1+(c-1)/2^{\lceil \lg\frac{c-1}{4cq}\rceil}+\frac{4cq}{c-1}\lceil\lg\frac{c-1}{4cq}\rceil & \text{if } 0<q<\frac{c-1}{4c}\\ c & \text{if } \frac{c-1}{4c}\leq q\leq\frac{1}{2}\end{cases}$$

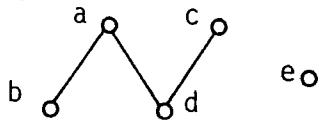
(ii)

$$\lim_{n\to\infty}\sup\frac{V_{\lfloor q(n-1)\rfloor+1}(n)}{n}\leq\begin{cases}1 & \text{if } q=0\\ 1+2^{1-\lceil\lg\frac{1}{5q}\rceil}+5q\lceil\lg\frac{1}{5q}\rceil & \text{if } 0<q<\frac{1}{5}\\ 3 & \text{if } \frac{1}{5}\leq q\leq\frac{1}{2}\end{cases}$$

(iii) If $i=o(n)$, $V_i(n)=n+o(n)$.

2. Algorithm

In this section, we introduce an algorithm, due to Blum et al.[1], which asymptotically surpasses Schönhage et al.'s for $i<\frac{1}{5}n$. The contestants really constitute a total ordered set; but the order is initially not known, and at any stage the algorithm's knowledge of inequality relations between contestants is given by a partial order or equivalently a Hasse diagram; for example



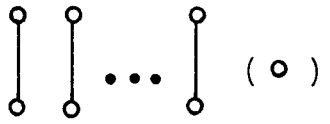
indicates that b<a, d<a and d<c.

The following is an algorithm obtained from Blum et al.'s by interchanging step 1 and step 2.

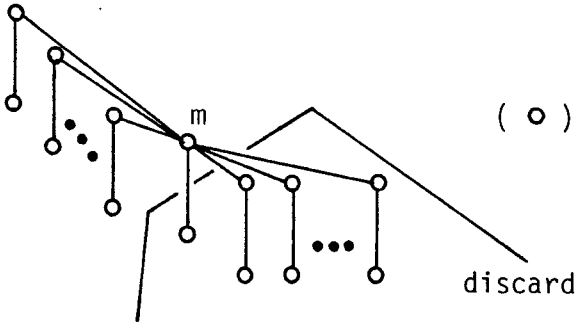
Algorithm SELECT[A,p] (This algorithm selects the i-th largest from n distinct elements. This contains two parameters p and A which will be chosen later; p is any positive constant and A is any algorithm solving the selection problem.)

1. If $i\geq pn$, then select the i-th largest by using algorithm A.

2. Partition n elements into $\lfloor \frac{n}{2} \rfloor$ pairs and possibly one leftover, and compare each pair.



3. Select the i-th largest m from $\lfloor \frac{n}{2} \rfloor$ larger elements of step 2 by using (recursively) SELECT[A,p].



4. Discard all elements known to be larger than m, since these cannot be the i-th largest.

5. Select the i-th largest from the remaining elements by using algorithm A.

3. Analysis

In this section, we investigate how SELECT[A,p] can surpass a general $(cn+o(n))$-algorithm. As a matter of convenience, we introduce a notation

$$V(q)= \lim_{n\to\infty} \sup \frac{V_{\lfloor q(n-1) \rfloor +1}^{(n)}}{n} , \quad 0 \leq q \leq 1. \tag{5}$$

Theorem 1 If there exists a $(cn+o(n))$-algorithm for selection, called PICK(c), then SELECT[PICK(c),$\frac{c-1}{4c}$] brings about a upper bound

$$V(q)\leq\begin{cases}1 & \text{if } q=0\\ 1+(c-1)/2^{\lceil \lg\frac{c-1}{4cq}\rceil}+\frac{4cq}{c-1}\lceil \lg\frac{c-1}{4cq}\rceil & \text{if } 0<q<\frac{c-1}{4c}\\ c & \text{if } \frac{c-1}{4c}\leq q\leq\frac{1}{2}\end{cases}$$

Proof. Let p be any positive number, and let $Q(i,n,p)$ be the worst-case number of comparisons required in SELECT[PICK(c),p]. Obviously

$$Q(i,n,p)=cn+o(n) \quad \text{for } i\geq pn. \tag{6}$$

For $i<pn$, since SELECT[PICK(c),p] is called

$$t=\lceil \lg\frac{pn}{c}\rceil$$

times and for each j-th calling, $1\leq j\leq t$, the number of contestants is $\lfloor n/2^{j-1}\rfloor$,

$$\begin{aligned}Q(i,n,p)&\leq \underbrace{\textstyle\sum_{j=1}^{t}\lfloor n/2^{j}\rfloor}_{\text{step 2}}+\underbrace{c\lfloor n/2^{t}\rfloor+o(n/2^{t})}_{\text{step 1}}+\underbrace{t(c(2i)+o(i))}_{\text{step 5}}\\ &= n-n/2^{t}+cn/2^{t}+2cti+o(n)\\ &= n+(c-1)n/2^{t}+2cti+o(n).\end{aligned} \tag{7}$$

By combining (6) and (7), we obtain

$$Q(i,n,p)\leq\begin{cases}cn+o(n) & \text{if } i\geq pn\\ n+(c-1)n/2^{\lceil \lg\frac{pn}{i}\rceil}+2c\lceil \lg\frac{pn}{i}\rceil i+o(n) & \text{if } i<pn\end{cases} \tag{8}$$

Thus from (5), (8) and the fact $V_i(n)\leq Q(i,n,p)$

$$V(q)\leq\begin{cases}1 & \text{if } q=0\\ 1+(c-1)/2^{\lceil \lg\frac{p}{q}\rceil}+2cq\ \lg\frac{p}{q} & \text{if } 0<q<p\\ c & \text{if } 0<q\geq p\end{cases}$$

Setting $p=\frac{c-1}{4c}$ gives the theorem. Q.E.D.

It should be noted that setting $p=\frac{c-1}{4c}$ minimizes the right hand side of (9); thus we cannot obtain a better upper bound from (9). To see this, let $F(p,q)$ be the right hand side of (9), and remember that p is an arbitrary positive number in (9). For $q=0$, the proposition is obviously true. For $q>0$, since $\lceil \lg \frac{p}{q} \rceil$ varies over $\{1,2,3,\ldots\}$ with p's varying over $\{p \mid q<p\}$, and since $0<q\geq p$ implies $F(p,q)=c=1+(c-1)/2^0+2cq\cdot 0$, we obtain

$$\min_{p>0} F(p,q)=\min\{1+(c-1)/2^k+2ckq \mid k=0,1,2,\ldots\}.$$

Now by an elementary analysis,

$$\min\{1+(c-1)/2^k+2ckq \mid k=0,1,2,\ldots\}=1+(c-1)/2^k+2ckq$$

if and only if

(i) $k=0$ and $\frac{c-1}{4c} \leq q$

or (ii) $k\geq 1$ and $(c-1)/(c\cdot 2^{k+2}) \leq q \leq (c-1)/(c\cdot 2^{k+1})$.

Hence

$$\min_{p>0} F(p,q)=\begin{cases} c & \text{if } \frac{c-1}{4c} \leq q \\ 1+(c-1)/2^k+2ckq & \end{cases}$$

$$\text{if } k\geq 1 \text{ and } (c-1)/(c\cdot 2^{k+2}) \leq q < (c-1)/(c\cdot 2^{k+1})$$

$$=\begin{cases} c & \text{if } \frac{c-1}{4c} \leq q \\ 1+(c-1)/2^{\lceil \lg \frac{c-1}{4cq} \rceil}+2cq\lceil \lg \frac{c-1}{4cq} \rceil & \text{otherwise} \end{cases}$$

$$=F(\tfrac{c-1}{4c},q)$$

As a corollary of the proof of Theorem 1, we have:

Corollary 2 If $i=o(n)$, $V_i(n)=n+o(n)$.

Proof. Let p be any positive constant. Then for sufficient large values of n, we obtain from (1), (9) and the fact $V_i(n) \leq Q(i,n,p)$,

$$1+\frac{i}{n}+\frac{i}{n}\lg\frac{n}{2i}+\frac{o(n)}{n} \leq \frac{V_i(n)}{n} \leq 1+(c-1)/2^{\lceil \lg \frac{pn}{i} \rceil}+2c\lceil \lg \frac{pn}{i} \rceil \frac{i}{n}+\frac{o(n)}{n}.$$

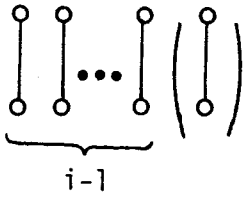
Thus

$$\lim_{n\to\infty} \frac{V_i(n)}{n} = 1. \qquad \text{Q.E.D.}$$

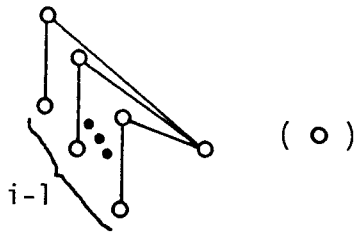
From (3) and Theorem 1, we obtain

$$V(q) \le \begin{cases} 1 & \text{if } q=0 \\ 1+2^{1-\lceil \lg \frac{1}{6q} \rceil} + 6\lceil \lg \frac{1}{6q} \rceil & \text{if } 0<q<\frac{1}{6} \\ 3 & \text{if } \frac{1}{6} \le q \le \frac{1}{2} \end{cases}$$

Can this be surpassed ? Let us now reconsider the proof of Theorem 1. Suppose that p is any positive number, and that GS denotes the generalized Schönhage et al.'s algorithm. GS is roughly given in Fig.1. The initial step of GS is a pairing step; especially for GS invoked at step 5 of SELECT[GS,p], the initial pairing step is to form a Hasse diagram



But we can save i-1 comparisons out of these i (or i-1) comparisons, since after step 4 of SELECT[GS,p] the remaining $2i$ (or $2i$-1) elements constitute a Hasse diagram



Thus (7) can be improved in SELECT[GS,p]:

$$Q(i,n,p) \le \underbrace{\sum_{j=1}^{t} \lfloor n/2^j \rfloor}_{\text{step 2}} + \underbrace{3\lfloor n/2^t \rfloor + o(n/2^t)}_{\text{step 1}} + \underbrace{t(3(2i)-(i-1)+o(i))}_{\text{step 5}}$$

$$= n+2^{1-t}n+5ti+o(n)$$

This leads to the following theorem in the same manner that (7) leads to Theorem 1.

Theorem 3

$$V(q) \le \begin{cases} 1 & \text{if } q=0 \\ 1+2^{1-\lceil \lg \frac{1}{5q} \rceil} + 5\lceil \lg \frac{1}{5q} \rceil & \text{if } 0<q<\frac{1}{5} \\ 3 & \text{if } \frac{1}{5} \le q \le \frac{1}{2} \end{cases}$$

In conclusion, the bounds for $V(q)$, due to Theorem 3 and (1), are illustrated in Fig.2. Note that for $(2^k+1)i-(2^k+1) < n \le (2^{k+1}+1)i-1$, $0 \le k$,

$$\left\lceil \lg \frac{n-i+1}{i+j} \right\rceil = \begin{cases} k+1 & \text{if } 0 \le j < \{n-(2^k+1)i+1\}/2^k \\ k & \text{if } \{n-(2^k+1)i+1\}/2^k \le j \le i-2 \end{cases}$$

Thus

$$\sum_{0 \le j \le i-2} \left\lceil \lg \frac{n-i+1}{i+j} \right\rceil = \begin{cases} k(i-1) + \lfloor \{n-(2^k+1)i\}/2^k \rfloor + 1 \\ \qquad \text{if } (2^k+1)i-1 \le n \le (2^{k+1}+1)i-(2^{k+1}+1),\ k \ge 0 \\ k(i-1) \quad \text{if } (2^k+1)i-(2^k+1) < n < (2^{k+1}+1)i-1,\ k \ge 0 \end{cases}$$

This leads to a corollary of (1):

$$V(q) \ge \begin{cases} \frac{3+q}{2} & \text{if } \frac{1}{3} \le q \le \frac{1}{2} \\ 1+2^{-k}+(k-2^{-k})q & \text{if } 1/(2^{k+1}+1) \le q < 1/(2^k+1),\ k \ge 1 \end{cases}$$

Acknowledgments

I wish to thank the referee for his kind and detailed comments, and Prof. J.Takeda for his careful reading.

References

[1] M.Blum, R.W.Floyd, V.Pratt, R.L.Rivest and R.E.Tarjan, Time Bounds for Selection, J.Comput.Systems Sci. 7(1973) 448-461.

[2] C.L.Dodgson, St.James's Gazette, August 1(1883) 5-6.

[3] A.Hadian and M.Sobel, Selecting the t-th Largest Using Binary Errorless Comparisons, Tech.Rep.121, Dept. of Statist., Univ. of Minneapolis, 1969.

[4] L.Hyafil, Bounds for Selection, SIAM J.Comput. 5(1976) 109-114.

[5] D.G.Kirkpatrick, A Unified Lower Bound for Selection and Set Partitioning Problems, J.ACM 28(1981) 150-165.

[6] D.E.Knuth, "The Art of Computer Programming", vol.3 Sorting and Searching, Addison-Wesley, 1973.

[7] A.Schönhage, M.Paterson and N.Pippenger, Finding the Median, J.Comput. Systems Sci. 13(1976) 184-199.

[8] C.K.Yap, New Upper Bounds for Selection, Commun.ACM 19(1976) 501-508.

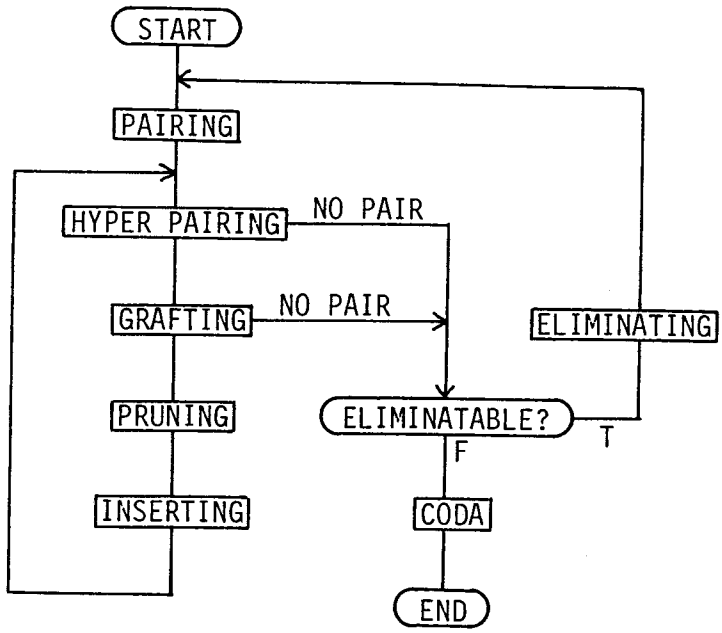

START
PAIRING
HYPER PAIRING
NO PAIR
GRAFTING
NO PAIR
ELIMINATING
PRUNING
ELIMINATABLE?
T
F
INSERTING
CODA
END

Fig.1.

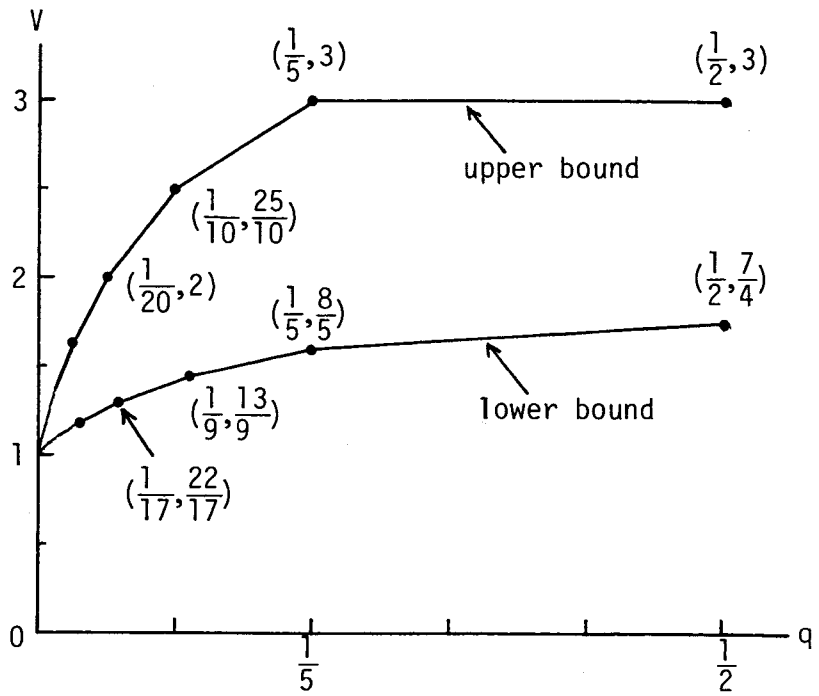

V
3
2
1
0
q
$\frac{1}{5}$
$\frac{1}{2}$
$(\frac{1}{5},3)$
$(\frac{1}{2},3)$
upper bound
$(\frac{1}{10},\frac{25}{10})$
$(\frac{1}{20},2)$
$(\frac{1}{2},\frac{7}{4})$
$(\frac{1}{5},\frac{8}{5})$
$(\frac{1}{9},\frac{13}{9})$
lower bound
$(\frac{1}{17},\frac{22}{17})$

Fig.2.