Revisiting Sequential Search Using Question-sets with Bounded Intersections

Dömötör Pálvölgyi,
Inst. of Mathematics, Comm. Networks Lab,
Eötvös Loránd University, Budapest, Hungary.
Email: dom@cs.elte.hu

24th March 2007

Abstract

The number of yes-no questions needed to find sequentially an unknown element of a set of size n with the restriction that the intersection of any m question-sets can be at most k is denoted by N. Given N, m and k we give the exact value of the biggest n for which N questions are still enough. We also give an almost exact formula for the smallest N if we know n, m and k.

AMS Subject Classification: 90B40

Key Words: Sequential search

1 Introduction

In a search problem, the goal is to find a given element (often called *pivot*) of a finite set using only Yes-No questions (often called *querries*) with some restrictions, using as few questions as possible. Usually only the size of the given set that matters, so let us denote the set by $\{1, \dots, n\} = [n]$. We investigate the sequential case where the questions can depend on the answers to the questions before them, so like in a usual 20 questions game. The goal is, of course, to minimize the number of questions needed to be asked in the worst case. One can easily see that if there are no other restrictions, then the minimum number of the questions needed is $\lceil \log n \rceil$. (Here we note that in this paper the base of the logarithm is always 2.) Since $\log n$ is not always an integer number, this suggests that it might be easier to deal with these kind of problems if we ask the question in the other way around; How big can the number n be at most if we want to find the searched element using only N questions? We denote the above number by $F(N)$. Now it is time to make some restrictions.

Every question can be represented as a subset of $[n]$; the subset of the elements for which the answer to the question is yes. So it can be put in the form „$x \in A$?" where x is the searched element and $A \subset [n]$. Our restriction is that during a game the size of the intersection of any m of these subsets can be at most k, where m and k are given positive integers. We denote the number of

necessary questions in the worst case by $f_k(n,m)$ and we denote the biggest n for which in a set of size n we can still find an element using only N questions by $F_k(N,m)$. We will determine the exact value of the latter in the next section and then we will give estimate of the former using this in Section 3.

This special model was proposed by Katona (1976) hoping that from it we could derive a lower bound for the number of comparisons needed to sort n elements. It was further studied by Sebő (1982). This paper uses some of his notations and his formula of recursion for $F_k(N,m)$. Sebő obtained an exact value for $F_k(N,m)$ from the (exact) formula of recursion only for $k=1$, otherwise he only gave bounds. This paper settles the problem of finding the exact value of $F_k(N,m)$ for all k's and therefore the estimate derived for $f_k(n,m)$ is also stronger. (In Sebő (1982) there was a gap of size n while we have an exact value for most of the n's and a gap of size only 2 for the other sufficiently large n's.)

2 Determining *F*

We start with two useful notations. Let $N_0 := \lfloor \log k \rfloor + 1$ and $\widetilde{k} := 2^{N_0}$. Note that $k < \widetilde{k} \leq 2k$, so it is the smallest power of 2 greater than k.

The special case $m=1$ (where the condition is that the size of each set can be at most k) has been solved by Katona (1966):

Proposition 2.1 *$F_k(N,1) = 2^N$ if $N \leq N_0$ and*
$F_k(N,1) = k(N-N_0) + \widetilde{k}$ if $N \geq N_0$.

Proof. First we will present an algorithm that solves the problem with N questions on $F_k(N,1)$ elements. The $N \leq N_0$ case is trivial. If $N > N_0$, we ask disjoint sets of size k until we have only N_0 questions left. If meanwhile once the answer was *yes*, then we can easily find the element using $\lceil \log k \rceil \leq N_0$ questions. If the answer was *no* all the time, then the searched element is in a $\widetilde{k} = 2^{N_0}$ size subset. We can start asking a set of size 2^i in the step when we have i questions left. If once we get a *yes* answer, then we can find the element with the remaining i questions. If the answer is *no* all the time, then the last element (contained in none of the sets) is the searched one. This was $k(N-N_0) + 2^{N_0-1} + \ldots + 1 + 1 = k(N-N_0) + \widetilde{k}$ (or 2^N if $N \leq N_0$) elements.

We will prove the upper bound using induction on N. If $N \leq N_0$ then the statement trivially holds. Let us suppose that we have N questions and that an optimal algorithm asks a set of size s with its first question. If the answer is *no*, then with the remaining $N-1$ questions we must be able to search a set of size $F_k(N,1) - s$, thus $F_k(N,1) - s \leq F_k(N-1,1)$. On the other hand, if the answer is *yes*, then we must be able to search a set of size s with the remaining N questions, so $s \leq F_k(N-1,1)$. We also know that $s \leq k$, this is the special restriction of the search. Putting this together gives $F_k(N,1) \leq F_k(N-1,1) + min\{k, F_k(N-1,1)\}$. If $N > N_0$, then this gives $F_k(N,1) \leq k(N-1) + \widetilde{k} + k$, while if $N \leq N_0$, then we get $F_k(N,1) \leq 2^{N-1} + 2^{N-1}$, so in both cases we have the statement of the proposition. □

Now we will state our main lemma, that has a similar proof as the proposition before and gives us a very useful recursion.

Lemma 2.2 *If $m \geq 2$, then $F_k(N,m) = F_k(N-1,m-1) + F_k(N-1,m)$.*

Proof. First we will prove the upper bound using induction on N. If $N = 1$ then the statement ($2 \leq 1 + 1$) trivially holds. Let us suppose that we already know $F_k(N-1,m)$ and $F_k(N-1,m-1)$ and one of the optimal strategies for both of them. Now let us examine the first step of an algorithm that solves the case of k, m. It asks a set S of size s, then it will have $N-1$ questions left. If the answer is *yes*, then in the rest of the game we can assume that it asks only subsets of S. These subsets used in the rest of the search must satisfy the condition with $m-1$ intersections. So from the induction hypothesis we know that with the $N-1$ questions left a set of size at most $F_k(N-1,m-1)$ can be searched, thus $s \leq F_k(N-1,m-1)$. Similarly, if the answer is *no*, it can be supposed that the algorithm will ask only sets disjoint to S. These subsets must satisfy the condition for m intersections, so with the $N-1$ questions left a set of size at most $F_k(N-1,m)$ can be searched by using again the induction hypothesis, so $F_k(N-1,m) - s \leq F_k(N-1,m)$. This together gives us $F_k(N,m) \leq F_k(N-1,m-1) + F_k(N-1,m)$.

To see the lower bound, note that if we ask a set of size $F_k(N-1,m-1)$, we can solve the search problem using induction in $F_k(N-1,m-1)+F_k(N-1,m)$ questions. □

This recursion is suspiciously similar to the one of binomial coefficients. This helps us to conjecture our main theorem and after knowing the formula, it can be easily proved using induction. (Compare this theorem with the formerly known result of Sebő: $\frac{k}{m!}(N-N_0-m+2)^m < F_k(N,m) \leq \frac{k}{m!}(N-N_0+3)^m$ if $N \geq N_0 + m - 1$.)

Theorem 2.3 *$F_k(N,m) = 2^N$ if $N \leq N_0 + m - 1$ and*
$F_k(N,m) = k\binom{N-N_0}{m} + \widetilde{k}\binom{N-N_0}{m-1} + \widetilde{k}\binom{N-N_0}{m-2} + \ldots + \widetilde{k}$ if $N \geq N_0 + m - 1$.

Proof. First note that for $N = N_0 + m - 1$ the two quantities are equal because $\widetilde{k}\sum_{i=0}^{m-1}\binom{m-1}{i} = \widetilde{k}2^{m-1} = 2^{N_0+m-1} = 2^N$.

Now we prove the statement by induction on m. For $m = 1$ the equality holds because of Proposition 2.1. For $m \geq 2$ using Lemma 2.2 we have $F_k(N,m) = F_k(N-1,m-1)+F_k(N-1,m)$. If $N \leq N_0+m-1$, then $N-1 \leq N_0+(m-1)-1$ and $N-1 \leq N_0+m-1$ as well, so we get $F_k(N,m) = 2^{N-1} + 2^{N-1} = 2^N$. If $N > N_0 + m - 1$, then $N-1 \geq N_0 + (m-1) - 1$ and $N-1 \geq N_0 + m - 1$, so we get $F_k(N,m) = (k\binom{N-1-N_0}{m-1} + \widetilde{k}\binom{N-1-N_0}{m-1-1} + \ldots + \widetilde{k}) + (k\binom{N-1-N_0}{m} + \widetilde{k}\binom{N-1-N_0}{m-1} + \ldots + \widetilde{k}) = k\binom{N-N_0}{m} + \widetilde{k}\binom{N-N_0}{m-1} + \ldots + \widetilde{k}$ using the recursion for binomial coefficients. This completes our proof. □

3 Estimating *f*

We already have a not too explicit formula for f: $f_k(n,m) = \min\{N : F_k(N,m) \geq n\}$. For a given n, one can of course easily decide how much $f_k(n,m)$ is by using the formula for F. However, we would like to have an explicit function of n. First, we give the exact value for very small n's, then we determine the exact value of the function except a 0-density set for which there are two possible values, except for a finite set (for small n's).

Proposition 3.1 *If $n \leq 2^{N_0+m-1}$, then $f_k(n,m) = \lceil \log n \rceil$.*

Proof. This follows immediately from the first part of the formula for F. □

Theorem 3.2 *If $n \to \infty$, then* $f_k(n,m) = \left\lceil (\frac{nm!}{k})^{\frac{1}{m}} + N_0 + \frac{m-1}{2} - \frac{\widetilde{k}}{k} + o(1) \right\rceil$.

The notation $a_n = o(b_n)$ stands for $\frac{a_n}{b_n} \to 0$, $a_n = O(b_n)$ means $\frac{a_n}{b_n}$ is bounded.

Proof. By definition, $f_k(n,m)$ is the smallest N for which $F_k(N,m) \geq n$. So we have to compute N from a value of $F_k(N,m)$. (We can think about $F_k(N,m)$ as if it was n.) We claim that $(\frac{F_k(N,m)m!}{k})^{\frac{1}{m}} \overset{n\to\infty}{\longrightarrow} N_0 + \text{constant}$. First note that $n \to \infty$ is equivalent to $N \to \infty$ and even to $(N - N_0) \to \infty$, so we can use the notation $O(N - N_0)$. Now using Theorem 2.3 we obtain:

$$\frac{F_k(N,m)m!}{k} = \frac{m!}{k}\left(k\binom{N-N_0}{m} + \widetilde{k}\binom{N-N_0}{m-1} + \ldots + \widetilde{k}\right) =$$

$$= (N-N_0)^m - (N-N_0)^{m-1}\binom{m}{2} + \frac{m\widetilde{k}}{k}(N-N_0)^{m-1} + O((N-N_0)^{m-2}).$$

Now let us devide this inequality by $(N - N_0)^m$ and take the $\frac{1}{m}$th power of both sides. We get:

$$\frac{(\frac{F_k(N,m)m!}{k})^{\frac{1}{m}}}{N-N_0} = \left(1 + \frac{\frac{m\widetilde{k}}{k} - \binom{m}{2}}{N-N_0} + O\left(\frac{1}{(N-N_0)^2}\right)\right)^{\frac{1}{m}}.$$

It is well known that for $\alpha > 0$ we have $(1 + x + O(x^2))^\alpha = 1 + \alpha x + O(x^2)$ if $x \to 0$. Using this for $x = \frac{1}{N-N_0}$ and $\alpha = \frac{1}{m}$ we have

$$\frac{(\frac{F_k(N,m)m!}{k})^{\frac{1}{m}}}{N-N_0} = 1 + \frac{1}{m}\frac{\frac{m\widetilde{k}}{k} - \binom{m}{2}}{N-N_0} + O\left(\frac{1}{(N-N_0)^2}\right).$$

Multiplying by $N - N_0$ gives:

$$\left(\frac{F_k(N,m)m!}{k}\right)^{\frac{1}{m}} = N - N_0 + \frac{\widetilde{k}}{k} - \frac{m-1}{2} + o(1).$$

So we obtained an equality for N:

$$N = \left(\frac{F_k(N,m)m!}{k}\right)^{\frac{1}{m}} + N_0 + \frac{m-1}{2} - \frac{\widetilde{k}}{k} + o(1).$$

Supposing $f_k(n,m) = N$ means $F_k(N-1,m) < n \leq F_k(N,m)$, so using the monotonicity of $F_k(N,m)$, the statement of the theorem follows:

$$f_k(n,m) = N = \left\lceil \left(\frac{nm!}{k}\right)^{\frac{1}{m}} + N_0 + \frac{m-1}{2} - \frac{\widetilde{k}}{k} + o(1) \right\rceil$$

□

Formerly only the following estimate was known:

$$\left(\frac{nm!}{k}\right)^{\frac{1}{m}} + \lfloor \log k \rfloor - 2 \leq f_k(n,m) < \left(\frac{nm!}{k}\right)^{\frac{1}{m}} + \lfloor \log k \rfloor + m \text{ if } n > k2^m.$$

This one has a gap of size $m + 2$. This can be compared to the corollary of our theorem (that is obtained by weakening it by getting rid of the notation $\widetilde{k}$):

Corollary 3.3 *For all m, k integers and $\epsilon > 0$ if n is sufficiently large, then*

$$\left\lceil (\frac{nm!}{k})^{\frac{1}{m}} + \lfloor \log k \rfloor + \frac{m-3}{2} - \epsilon \right\rceil < f_k(n,m) < \left\lceil (\frac{nm!}{k})^{\frac{1}{m}} + \lfloor \log k \rfloor + \frac{m-1}{2} + \epsilon \right\rceil .$$

Acknowledgement: I would like to thank Gyula O. H. Katona for raising my awareness to the problem and for his advice regarding the paper.

References

G. Katona, 1966, On separating systems of a finite set, *Journal of Combinatorial Theory*, v. 1, 174-194.

G. Katona, 1976, Search using sets with small intersection, *Colleques Internat, C.N.R.S. No. 276, Theorie de l'information.*

A. Sebő, 1982, Sequential search using question-sets with bounded intersections, *Journal of Statistical Planning and Inference 7*, 139-150.