# On the Optimality of a Hash-Coding Type Search Algorithm

L. H. Aslanyan

Institute for Informatics and Automation Problems of NAS RA

Yerevan, Armenia

e-mail: lasl@sci.am

H. E. Danoyan

Institute for Informatics and Automation Problems of NAS RA

Yerevan, Armenia

e-mail: haykdanoyan@yandex.ru

## ABSTRACT

The algorithmic problem of nearest neighbors search and the early proposed hash-coding algorithms (Elias algorithm, tree based algorithms) are considered. The problem of optimality characterization of the algorithm - depending on the geometrical structure of blocks in a hash-coding scheme is considered. As a supportive technique for this the structure of consequtive neighbourhood layers to the subsets of binary cube is considered. It is prooved that the function representing the cardinalities of $k^{\text{th}}$ neighborhoods of a set depending on $k$ is concave. Using this fact and the known properties of isoperimetric sets the difference between the mantioned functions (for fixed argument) corresponding to isoperimetric set and an arbitrary set of the same cardinality can be estimated. It is prooved that the algorithm is optimal when the mentioned blocks are isoperimetric sets.

## Keywords

Best match, nearest neighbors, hash-coding, Elias algorithm, k[th] neighborhood, standard placement.

## 1. INTRODUCTION

Let E={0,1}. Consider Cartesian degree $E^n$ which is known as the set of vertexes of n-dimensional unit cube. For $x, y \in E^n$ denote by $d(x,y)$ the Hamming distance between vectors $x$ and $y$. Denote by w(x) the weight of vector x, i.e. $w(x) = d(0^n, x)$. For an $x \in E^n$ denote by $S_r^n(x)$ the sphere of radius $r$ centered at point $x$ i.e. $S_r^n(x) = \{y \in E^n / d(x,y) \le r\}$ and denote by $O_r^n(x)$ the neighbourhood / shell of radius $r$ i.e. $O_r^n(x) = \{y \in E^n / d(x,y) = r\}$. For an $x \in E^n$ and $A \subseteq E^n$ we define the distance $d(x,A)$ as $d(x,A) = \min_{a \in A} d(x,a)$. For $A \subseteq E^n$ denote by $A^{(k)}$ the k[th] neighborhood of $A$, i.e. $A^{(k)} = \{x / d(x,A) = k\}$.

The algorithmic problem of finding the set of all nearest neighbors to a given vector from a given set is known [1]. More precisely, let us we have a nonempty subset $F \subseteq E^n$ and a vector $x$ (query element or query vector). Denote by $F_x$ the set of nearest neighbors to$x$ from $F$ i.e.

$$F_x = \{y \in E^n / d(x,y) = d(x,F)\}.$$

The problem is for a given $F$ and $x$ to find in an optimal way the set $F_x$. We can find the mantioned set by calculating for each $y \in F$ the corresponding distances to $x$but in many cases, for large $F$ it will make the problem practically unsolvable. So, it is required from algorithm to consider as few as possible elements from $F$ in constructing the nearest neighbours. One algorithm solving the mantioned problem is known as Elias algorithm [1]. The main goal of this paper is to provide the proof of the conjecture about optimality of the mentioned algorithm conserning the geometrical shape of code blocks[1].

The paper is organized as follows: In section 2 the descriptions are brought related to the hash-coding schemas and algorithms. In section 3 some auxiliary rezults and definitions are brought about cardinalities of neighborhoods of a cube subset. Finally, in section 4 the theorem is prooved about the optimality of the algorithm.

## 2. ELIAS ALGORITHM

At first, let us describe a particular schema of representataion of subset $F$, called hash-coding schema. Hash function is defined as a function $h: E^n \to V$, where $V$ is a finite set of $N$ elements, $V = \{v_1, \dots, v_N\}$[2]. We consider the case, that $V = E^m$, and $m \le n$. In connection to hash function, $F$ is represented as a union of $N$ distinct linked lists (or buckets) $L_i$, $i = 1, \dots, N$ one for each possible hash value. For $i \in \{1, \dots, N\}$ denote by $B_i$ the set $\{x \in E^n / h(x) = v_i\}$. $B_i$'s are called blocks. The i[th] list stores those vectors belonging to $F$ which have the same hash value i.e. $L_i = \{x \in E^n / h(x) = v_i\}$, or in other words $L_i = F \cap B_i$. Hash coding scheme is called balanced if $|B_i| = 2^n/N$, for $i = 1, \dots, N$.

The Elias algorithm [1] considers blocks $B_i$ of the model, ordering them by their distances at vector $x$. Mention that we must have an efficient method to find all blocks $B_{j_1}, \dots, B_{j_{s(j)}}$ located at distance $j$ from $x$ if such blocks exist. After the step of these ordering the algorithm examines the lists $L_{j_1}, \dots, L_{j_{s(j)}}$ one after the other by increase of $j_t$. Let the best match distance (also its current value in algorithm) be denoted by $\delta$. Due to $F \neq \emptyset$ initialization of $\delta$ will happen on some step and the final algorithmic value will present the real minimal distance $\delta$. Now, if the current values obey $\delta < j$ algorithm stops the work. All blocks with higher distances than $\delta$ at $x$ do not need to be examined. In the reminder case $\delta \ge j$, examining nonempty list $L_{j_i}$ algorithm can change the best match distance $\delta$, also refreshing the current best match set, or the $\delta$ will remain unchanged and the current best match set will be updated.

By the complexity of algorithm we mean the average number of examined lists over all files and queries, supposing that

I. Each vector $x \in E^n$ equally likely can be requsted.

II. Each vector $z \in E^n$ independently appears in $F$ with the same probability $p$. This gives a specific probabilistic distribution over the set of subsets of $E^n$ appearing as the target set $F$.

The pseudocode of the algorithm is brought below:

**Elias Algorithm** /* n is the word length, N is the number of blocks */
input x, F, /* F ≠ ∅ */
integer δ = ∞, /* the current best match distance */
set S = ∅, /* S is the current set of vectors of F located at distance δ from x */
integer j = −1, /* current distance of blocks under consideration from x */
while (j < δ)
{
j + +,
if (s(j) ≠ 0) /*s(j) is the number of blocks in distance j from x*/
for (integer i=0, i < s(j), i++)
{
If ($L_{j_i} \neq \emptyset$) /* start examine the list $L_{j_i}$, i-th list with j distanse block */
If (δ ≤ d(x, $L_{j_i}$)) /*δ is unchanged */
$S = S \cup \left(O_\delta^n(x) \cap L_{j_i}\right)$ /*$O_\delta^n(x)$ is the δ neighborhood of x */
else
{
$S = O_\delta^n(x) \cap L_{j_i}$, /*δ is changed */
$\delta = d(x, L_{j_i})$
}
}
}
return S, comment: S = $F_x$, δ = d(x, F)

For balanced hash coding schemes it is proposed that the algorithm may be optimal when the blocks $B_i$ are isoperimetric sets [1],[3]. Our main goal is to prove the mentioned conjecture.

## 3. ON FUNCTION OF CARDINALITIES OF NEIGHBORHOODS OF A SET

For $A \subseteq E^n$ let us denote

$$\Gamma_A(k) = |A^{(k)}|, k \in \{0, \dots, n\}.$$

It is obvious, that $\sum_{k=0}^{n} \Gamma_A(k) = 2^n$. For a one-point set, $A = \{x\}$, values $\Gamma_A(k)$ are given by combinations $\binom{n}{k}$. For larger sets starting from some point values $\Gamma_A(k)$ become equal to 0. For a one-point set it is evident that $\Gamma_A(k)$ is simply increasing and then decreasing at the point $n/2$. The similar behaviour of the function $\Gamma_A(k)$ is not studied for arbitrary $A \subseteq E^n$ and our nearest goal is this investigation.

**Lemma 3.1.** For arbitrary subset $A$ the function $\Gamma_A(k)$ has one or two maximum points on the domain $\{1, \dots, n\}$ and in case when the maximum points are two then they differ by 1.

**Proof**. Let us prove the lemma by induction. For $n = 3$ one can check that the lemma takes place. Now let the lemma be true in case $\leq n - 1$ and prove it for the case of $n$. Let us fix the i[th] coordinate and consider the partition of unit cube by coordinate $x_i$. Let us denote the corresponding subcubes by $E^n_{x_i=0}$ and $E^n_{x_i=1}$. In connection to this, denote the subsets of the set $A$ belonging to $E^n_{x_i=0}$ and $E^n_{x_i=1}$by $A_{x_i=0}$ and $A_{x_i=1}$ correspondingly. Let us denote $P_0 = A_{x_i=0} \cup {A_{x_i=0}}^{(1)} \cup A_{x_i=1}$ and $P_1 = A_{x_i=1} \cup {A_{x_i=1}}^{(1)} \cup A_{x_i=0}$. Note that $P_0$ and $P_1$ are subsets of $n-1$ dimensional unit cube. It can be shown that (Figure 1)

$$\Gamma_A(k) = \Gamma_{P_0}(k-1) + \Gamma_{P_1}(k-1) \quad (3.1)$$

but for the sake of space here we omit the detail proof of the claim.

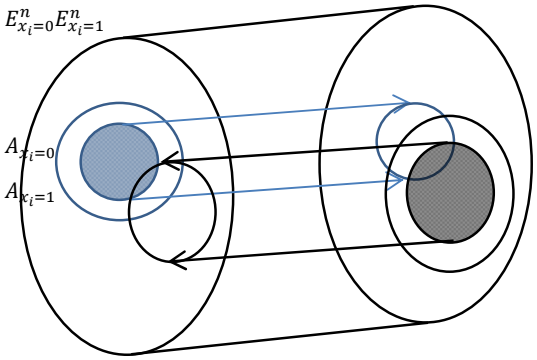


Figure 1.

As $P_0$ and $P_1$ are subsets of $n-1$ dimensional unit cube, then for functions $\Gamma_{P_0}(k)$ and $\Gamma_{P_1}(k)$the lemma is true by the induction hypothesis. Now let $k_0$ be the smallest integer such that one of the mentioned functions (without loosing generality assume that it is the $\Gamma_{P_0}(k)$) decreases. Then from the definition of the set $P_1$ it follows that at most at point $k_0 + 1$ the function $\Gamma_{P_1}(k)$ also decreases. The lemma follows purporting this and equation (3.1).

Now let us have a number $0 \leq a \leq 2^n$ and a point $c \in E^n$. We define a set of cardinality $a$ called standard placement [3]. At first we define a linear order on $E^n$ by the following way [1],[3]. Let $u, v \in E^n$, then:

I. if $d(u,c) < w(t(v,c))$ then $u \prec v$,

II. if $d(u,c) = d(v,c)$ then $u \prec v$ if there exists $i$, $1 \leq i \leq n$ such that $u_j = v_j$, $j = 0, .., i-1$ and $u_i + c_i = 1$, $v_i + c_i = 0$.

The standard placement with centre c and cardinality $a$ is denoted by $L^n_a(c)$. $L^n_a(c)$ is the set of first $a$ elements by means of linear order defined above [1],[3].

**Theorem 3.1 [1],[3]**. For $A \subseteq E^n$, $|A| = a$ and $c \in E^n$ the following takes place:

$$\Gamma_{L^n_a(c)}(1) \leq \Gamma_A(1). \quad (3.2)$$

This result is the origin of our investigation. It is known as the base Discrete Isoperimetry Theorem. More properties of Discrete Isoperimetry were investigated, see [3],[4], but this research concerns only the set and its 1-neighbourehood. Consequtive neighbourhoods raise more questions, and here our target is understand the functional behaviour of the series of neighbourhoods, and

their sizes in particular. Now let us have an arbitrary subset $A$, $|A| = a$ and the standard placement $L^n{}_a(c)$.

**Lemma 3.2**. There exists an integer $s$ ($1 \le s \le n$) such that

$$\Gamma_{L^n_a(c)}(k) \le \Gamma_A(k), \text{ when } k \le s \tag{3.3}$$

and

$$\Gamma_{L^n_a(c)}(k) \ge \Gamma_A(k), \text{ when } k > s. \tag{3.4}$$

The proof technique is similar to the proof of Lemma 3.1 and is omitted due to the space limitation. For geometrical interpretation of lemma 3.2 see figure 2.

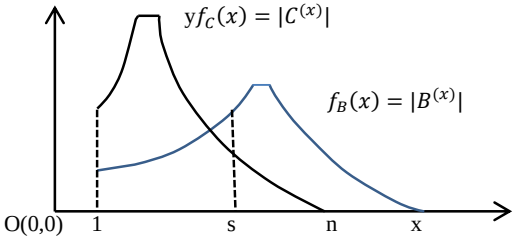


Figure 2.

## 4. OPTIMALITY OF THE HASH-CODING TYPE SEARCH ALGORITHM

Let us consider the following problem: which property must satisfy blocks of balanced hash-coding schema, so the algorithm to be optimal.

**Theorem 4.1.**The algorithm is optimal, when the blocks $B_i$, $i = 1, \dots, N$, of hash-coding schama are standard placements (spheres in particular cases).

**Proof.**As it is known [1], the complexity of the algorithm is

$$\alpha(h) = \sum_{i=1}^{N} \Phi(B_i), \tag{4.1}$$

where by $\Phi(B_i)$ is denoted the average probability aver all query vectors that the block $B_i$ will be considered. It is known [1], that

$$\Phi(B_i) = \frac{1}{2^n} \sum_{j=0}^{n} \left|B_i^{(j)}\right| \Psi(n, p, j-1), \tag{4.2}$$

where $\Psi(n, p, j-1)$ is the probability that the sphere of radius $j-1$ does not contain points belonging to subset $F$. Now let $B$ be a standard placement and $C$ be an arbitrary set, such that $|B| = |C|$. Let us consider the following difference

$$\Phi(C) - \Phi(B) = \tag{4.3}$$

$$= \frac{1}{2^n} \sum_{i=1}^{n} \left(\left|C^{(i)}\right| - \left|B^{(i)}\right|\right) \cdot \Psi(n, p, i-1)$$

From theorem 3.2 and lemma 3.2 it follows that there exists an integers, such that

$$\Phi(C) - \Phi(B) =$$

$$= \frac{1}{2^n} \sum_{i=1}^{s} \left(\left|C^{(i)}\right| - \left|B^{(i)}\right|\right) \cdot \Psi(n, p, i-1)$$
$$+ \frac{1}{2^n} \sum_{i=s+1}^{n} \left(\left|C^{(i)}\right| - \left|B^{(i)}\right|\right) \cdot \Psi(n, p, i-1)$$
$$\ge \frac{1}{2^n} \sum_{i=m}^{s} \left(\left|C^{(i)}\right| - \left|B^{(i)}\right|\right) \cdot \Psi(n, p, s-1) - \tag{4.3}$$
$$- \frac{1}{2^n} \sum_{i=s+1}^{n} \left(\left|B^{(i)}\right| - \left|C^{(i)}\right|\right) \cdot \Psi(n, p, s)$$

As $\Psi(n, p, s) \le \Psi(n, p, s-1)$ and $\sum_{i=1}^{s}\left(\left|C^{(i)}\right| - \left|B^{(i)}\right|\right) = \sum_{i=s+1}^{n}\left(\left|B^{(i)}\right| - \left|C^{(i)}\right|\right)$, then from (4.3) follows that

$$\Phi(C) - \Phi(B) \ge 0.$$

## CONCLUSION

The nearest neighbors search is an important part of many procedures such as in Pettern Recognition, Natural Language Error Correction, and others. Two types of algorithms are well known in this area: the tree based algorithms and the dynamic programming style algorithms. Our interest is devoted to the second one. First of all we use the hypotheses that the optimality of this type algorithms is related to the well known Discrete Isoperimetry problem. Then using the known results in this area and giving some generalization we obtain combinatorial postulates that help to prove the optimality of these algorithms. Results achieved are interesting for both – the Discrete Isoperimetry domain and in Divide and Conquer type srearch. The complementary domain of inverstigation that considers the issues how to partition the discrete space to the Isoperimetry type blocks is also under consideration but this is a separate broad research direction that is addressed [5], [6].

## REFERENCES


[1] R. L. Rivest, "On the optimality of Elias's algorithm for performing best-match searches", Information Processing, pp. 678-681, 1974.

[2] D. E. Knuth, "The Art of Computer Programming, vol. 3 / Sorting and Searching", Eddison-Wesley, p. 800, 1998.

[3] L. H. Aslanyan, "The Discrete Isoperimetry Problem and Related Extremal Problems for Discrete Spaces", ProblemiKibernetiki, pp. 85-128, 1979.

[4] L. H. Harper, Global Methods for Combinatorial Isoperimetric Problems, Cambridge University Press, 2010, 250p.

[5] L. H. Aslanyan, H. E. Danoyan, Complexity of Elias algorithm based on codes with covering radius three, Proceedings of the Yerevan state university, 2013 №1, pp. 44-50.

[6] L. H. Aslanyan, H. E. Danoyan, Complexity of Elias algorithm based on Hamming and extended Hamming codes, Reports of NAS RA, vol. 113, no. 2, pp. 151-158, 2013.