# Completeness of the word set and structured coding in problems of discrete mathematics

Evdokimov Alexandr Andreevich

Candidate of Physico-mathematical Sciences, Full Professor, Head of Laboratory,

Sobolev Institute of Mathematics SB RAS, Novosibirsk, Russia

**Abstract**. The article provides a brief overview of the known results, as well as new results on the problem of completeness of the set of words and the coding of information, which preserves the structural properties of the set being encoded. Some problems are well known in discrete mathematics, but new and unsolved old problems are presented. Algorithmic issues of recognition of the completeness of a set of words and its complexity are considered.

**Keywords.** discrete mathematics, symbolic sequences, completeness, words, boolean n-dimensional cube, structured coding, discrete functions, embeddings, codes.

A set *S* of words in a finite alphabet *A* is called complete if any infinite sequence of letters from *A* contains at least one word from *S* as a subword. In this case, it is said that the set of prohibition words *S* blocks any infinite sequence of letters of the alphabet *A*. Investigations of the completeness of a set of words and problems on avoiding by symbolic sequences of a set of "forbidden" subwords were initiated by the author back in the 70s, and in general form the problem of completeness was formulated in a small note [1] and in the talk “Non-repetitive sequences” made at the Joint Russian-German mathematical colloquium in 1979 in Rostock (Greiswald). In [1], the relationship between the formulated completeness problem and problems from various areas of discrete mathematics is noted. The usefulness of studying combinatorial problems about symbolic sequences in their geometric interpretation on de Bruijn graphs is noted. In particular, for efficient algorithmic recognition the completeness properties of a finite set of words. A result of V.A. Krainev on the boundedness of the length of non-repetitive in the strong sense binary sequences is presented (more details below). Guidelines for further research are given.

The author's interest in this topic and the general formulation of the completeness problem arose after he had solved two combinatorial problems. The first is the problem of the Hungarian mathematician P. Erdös about the existence of an infinite symbolic sequence in a finite alphabet that does not contain two consecutive repeating segments with the same composition of letters in them. The construction of an infinite non-repetitive symbolic sequence was found, which avoids the prohibition of such repetitions [2,3]. For the problems of non-repetitive sequences, we managed to make rather deep progress in understanding the difficulties of studying the border of transition from completeness to incompleteness. In this respect, an interesting result on the finiteness of the length of words in a binary alphabet for a set of stronger prohibitions, when the absence of repetitions of subwords equal in the frequency composition of letters in them is required, and this is true for any number of repetitions [4]. The geometric interpretation of these problems is interesting. In particular, the question posed by the author back in the 70s. Is there an infinite path in the positive direction of the unit vectors of an integer lattice of dimension n that does not contain k points lying on one straight line? For n = 2 and any k> 2, the answer is no. For an alphabet of cardinality n> 3 and arbitrary k, the problem has not been finally solved.

Research on the combinatorics of nonrepeated sequences has a long history. Thus, the studies of the Norwegian mathematician A. Thue on problems of algorithmic solvability in associative calculus were carried out at the beginning of the last century. In the 1920s, these are the works of M. Morse on topological dynamics. In these works, the possibility of a simple constructive assignment of endless non-repetitive sequences was used. Later, in the 60s, interest was revived in the study of various variations of the property of "strong non-periodicity",

irregularity, pseudo-randomness of symbolic sequences, the complexity of their definability and computability, which is associated with intensive research during this period of the mathematical foundations of Computer Science, cybernetics and applied problems of computer science [5, 6].

The second problem, which served as a source of interest and formulation of the completeness problem, is known in the mathematical literature as the "Snake-in-the-Box Problem". A snake is a simple path (or cycle) in a boolean n-cube that does not contain chords, that is, it is a generated subgraph of an n-cube. The question of the maximum length of such a path arose in the study of local algorithms for minimizing Boolean functions represented by formulas in disjunctive normal form [6]. The problem of the maximum length of a chain and a cycle was investigated by the author in the interpretation of words with prohibitions, when paths in a hypercube of dimension n are encoded by symbolic sequences in an n-letter alphabet with prohibitions on subwords. Moreover, finding the order of magnitude of the maximum cycle length and its construction are essentially based on the existence and construction of an infinite "control" sequence, also with restrictions-prohibitions of a special form on its subwords [7]. A lot of new information about publications and applications on the Snake-in-the-Box problem can be found on the Google search engine. Including the search with the help of modern computer technologies for "long snakes" in hypercubes of small dimensions.

When examining the completeness of a set of words, two types of questions arise. First, it is the study of the completeness of specific finite or infinite sets of prohibitions and the construction of prohibition-avoiding symbolic sequences in the case of incompleteness of the set of words. Second, research on general questions. For example, the construction of algorithms for recognizing completeness, their complexity, studying the completeness of constructively defined classes of infinite sets, describing the set of words and sequences free from prohibitions, its cardinality, estimates of the maximum length of words free from prohibitions, the growth function of the number of words of length n. At present, research is intensively developing in the area called, in the broader context of the problems under consideration, "combinatorics of words" [8, 9]. Words and symbolic sequences are the object of research in the theory of formal languages [10], the theory of information coding and compression, symbolic dynamics, the theory of automata, the mathematical foundations of cryptography and programming. The problems of analyzing the structure and studying the properties of strings of symbols arise in many areas of natural sciences. Finding effective algorithms for solving these problems is a topical area of research, which has been the subject of a large number of publications.

Let $S$ - be a complete set, that is, the set $\widehat{S}$ of prohibition-free words, of course.

Let's introduce the functions

$$L(\widehat{S}) = \max_{X \in S} |X|,$$

$$L(n) = \max_{S} L(\widehat{S}),$$

where, in the first case, max over all words free from prohibitions, and the second maximum over all complete sets of words $S \subset A^n$.

**Theorem.** [11] $L(n) = C(n) + n - 1 = |A|^{n-1} + n - 2$, where $C(n)$ - the greatest length of a simple path without chords in a de Bruijn graph of order $n$.

Consider $M(n) = \min_{S} |S|$, where $S$ - complete set, $S \subset A^n$

**Theorem.** [12] $M(n) = \frac{1}{n} \sum_{d/n} \varphi(\frac{n}{d}) |A|^d$, where $\varphi$ - Euler's function.

Notice, that $M(n) \sim {|A|^n}/{n}$, and if $|A| = 2$, then $M(10) = 108$. Thus, any set of length 10 binary words containing less than 108 words is avoidable, and this boundary is exact, that is, there is a set of 108 words of length 10 blocking any infinite binary sequence.

The next two theorems give algorithms for recognizing the completeness of word sets $S \subseteq A^n$.

**Theorem.** The completeness of any set $S \subseteq A^n$ is recognizable with a complexity of order $|S| \cdot n$.

**Evidence**. Using the interpretation of the completeness problem on de Bruijn graphs $B_m^n$ it is easy to prove the equivalence of the following statements:

1) $S$ full set of words;

2) the set $V(S)$ cuts all the contours of the graph $B_m^n$;

3) the subgraph of the graph $B_m^{n-1}$,formed by the set of arcs $E(B_m^{n-1}) \setminus E(S)$ is acyclic;

where $V(S)$ and $E(S)$ - the sets of vertices of the graph $B_m^n$ and arcs of the graph $B_m^{n-1}$, which correspond to the words of the set $S$. The equivalence of statements 1-3 is based on the correspondence between words $X$ of length $|X| \geq n$ and directed paths in graph $B_m^n$, passing through the vertices corresponding to subwords of length $n$ of word $X$. The equivalence of statements 2 and 3 is true, since the graph $B_m^n$ is an edge graph for $B_m^{n-1}$ by the definition of de Bruijn graphs. Thus, recognition of the completeness of the set of words $S$ can be replaced by checking the absence of directed contours in the subgraph formed by the set of arcs $E(B_m^{n-1}) \setminus E(S)$. Checking the completeness of the set $S$ we can consider $|S| \geq {|A|^n}/{n}$, since otherwise $S$ is incomplete by the definition of the function *M (n)*. Therefore, we have

$$|E(B_m^{n-1}) \setminus E(S)| = |A|^n - |S| \leq O(|S|(n-1)) \quad (1)$$

It remains to note that checking the acyclicity of a directed graph, for example. depth-first search algorithm is possible with a complexity of the order of the sum of the numbers of its vertices and arcs. Together with inequality (1) and Statements 1–3, this proves the theorem.

Let us now clarify the question of the boundary of the lengths of prohibition-free words of the set $\hat{S}$, formulating it in this way. Let $l$ be some natural number, $S$ a complete set of words, that is, the set $\hat{S}$ is bounded. Does $\hat{S}$ have a word that is longer than $l$?

**Theorem.** The problem of recognizing the existence in $\hat{S}$ of a word of length at least $l$ is NP-complete.

Thus, recognition of the boundedness of the lengths of words that avoid the set of forbidden subwords is possible with complexity $|S| \cdot n$, and the question of localizing this boundary has a qualitatively different complexity.

Here is an interesting unsolved problem that was posed by the author in his early works on the completeness of the set of words. Let's introduce the function

$$f(m,n) = \max {|S_1|}/{|S_2|},$$

where $m = |A|$, and the maximum is taken for all pairs $\{S_1, S_2\}$, $S_1 \subseteq A^n$, $S_2 \subseteq A^n$ complete irreducible sets (that is, such that the deletion of any word from which leads to incompleteness). There are complete irreducible sets

whose cardinality is not minimal. For example, the set of 7 words {0000, 0001,1001,0101,0110,0111,1111} is complete and irreducible, but not minimal, since $M(4) = 6$.

**Problem.** How large can be the difference in cardinalities of complete irreducible sets? How does the function behave (for example, in n for a fixed m) $f(n,m)$ ? Accurate estimates of its growth would significantly clarify the structure of complete word sets [11-13].

Let us explain the idea of another algorithm for recognizing the completeness of a set of words presented in [14]. Despite the simplicity of the algorithm, it turns out to be useful in the following ways. Certain invariant transformations are applied to the set of words $S$ as a result of which we obtain a chain of sets "derived from while preserving their property to be a complete or incomplete set. In this case, at each step of the algorithm, the description of the resulting set of inhibitions is reduced. After a finite number of steps according to the result, we definitely have the answer "yes, complete" or "no". Thus, we obtain not only a more compact description of the original set of prohibitions, but also an efficient "encoding through prohibitions" of the set $\widehat{S}$ of words free of the prohibitions $S$.

It is natural to pose questions of the completeness of the set of words not only in relation to the set of all words, but also in relation to any infinite set of $M$ words and $\omega-$ words (words of infinite length). The set of prohibitions $S$ is complete with respect to $M$, if $P(\omega)\cap S\neq\varnothing$ for any $\omega$-word from $M$, where $P(\omega)$ - is the set of all subwords of the $\omega$-word. If $P(\omega)\subset M$ for any $\omega\in M$, then the set $M$ is called closed.This agrees with the usual definition of a closure operator by the operation of including in $M$ all subwords of each of its words. In such a case, the set of "objects" closed by joining all "subobjects" is also called hereditarily closed.

**Theorem.** For any hereditarily closed infinite $M$ and arbitrary $S$, the following is true

a) $S-$ closed set of measure 0;
b) there is $\omega-$ word, such that $P(\omega)\subset M$;
c) $S$ is complete with respect to $M$ if and only if $S\cap M-$ is a finite set;
d) in any infinite $S$, complete with respect to $M$, there exists a finite $S'\subset S$ also complete with respect to $M$;
e) there exist infinite sets $S$ and $M$, such that $S$ is complete with respect to $M$ and remains complete after removing from $S$ any of its finite subset.

Note that, generally speaking, the theorem is not true for sets $M$, that are not hereditarily closed.
As can be seen from the theorem, some properties actually repeat the completeness of prohibitions with respect to a finite set of all words in an arbitrary finite alphabet. However, in this case, it is essential to fulfill the property of hereditary closedness of the set, in relation to which the problem of completeness is considered.

Above we have already spoken about the usefulness of the "language of prohibitions" when coding sets, in particular, describing problems of embedding graphs in hypercubes and their solution. For example, these are the problems of existence and construction of Hamiltonian cycles in a hypercube with various restrictions-prohibitions on their structure, the Snake-in-the-Box problem, embeddings of integer lattices into hypercubes [15], trees, graphs of computational structures. We consider embeddings in hypercubes as encodings of objects (graphs) that preserve certain structural properties of the encoded objects in the image-code and call them structured encoding [16,17]. With this approach, the universality of the n-dimensional cube is of great importance, as a set of words of length n in a finite alphabet, which can be endowed with various types of structures of geometric, algebraic, ordinal, or metric type: a graph, a partially ordered set, an integer lattice, an abelian group, a metric vector space with Hamming metric, torus metric or other metrics. This versatility makes it possible to model the structures of nested sets in a code-image in a hypercube of various types, preserving the necessary properties of the original encoded structures in the code.

Note that structured coding, assuming that the display preserves different types of properties, allows you to recognize and correct "errors" in the image by checking and analyzing the implementation or non-fulfillment of the properties stored by the display (generalized concept of error-correcting coding). The problems of coding data while preserving their structural properties also arise in connection with the implementation of a hypercube architecture in computing systems, in which the information interaction of microprocessors is determined by their connection into

the structure of an n-dimensional cube. For such networks, questions arise of information exchange, parallelization of computations, and a number of other problems of mapping data structures to the structure of a computer network. For example, full or partial preservation in codes of the metric structure of the original set during its isometric or locally isometric embedding into a hypercube. An important applied aspect of structured coding is that it allows you to conveniently and quickly work with the elements of the original set already in their machine codes using operations interconnected with the structure of a computer network. This speeds up data processing and increases the computation speed.

Various classes of mappings defining embeddings of discrete metric spaces were considered in [16,17]. These are isometric and locally isometric mappings, embeddings "with stretching" of distances, a parametric family of mappings of bounded distortion, a discrete analogue of homeomorphic embeddings. The latter presupposes the introduction of an analog of the continuity of the mapping for the discrete case, when the mapping “does not break close points of a metric space, and does not transfer distant points to close ones”. In [17], definitions are given in various versions: parametric maps of bounded distortion, definitions in topological terms of discrete neighborhoods, in particular, a discrete analogue of homeomorphic embeddings. By varying the parameters and type of mappings, we obtain different classes of nestings that determine the type of structured coding that preserves, in a strong or weak form, the necessary structural properties of the encoded set.

The considered directions of research and the results obtained were reported at international conferences, seminars and schools in Russia and foreign countries in different periods, starting from the end of the 70s.

This work was carried out with the financial support of the program for increasing the competitiveness of the Novosibirsk State University.

1. Evdokimov A.A., Krainev V.A., Problems on the completeness of word systems, XXII Obl. sci. - tech. conf. Novosibirsk. 1979. P.105 -107;

2. Evdokimov A. A. Strongly asymmetric sequences generated by a finite number of symbols // Soviet Math. Dokl. 1968. V. 9. P. 536-539.

3. Evdokimov A.A. The existence of a basis that generates 7-digit non-repetitive sequences // Coll. "Discrete Analysis". 1971. Iss. 18. P. 25-30.

4. Krainev V. A. Words that do not contain consecutive subwords of equal frequency composition. // Methods of discrete analysis in solving combinatorial problems; Coll. sci. op. Novosibirsk: Institute of Mathematics, SB of the USSR AS, 1980. Iss. 34. P. 27-37.

5. Turing machines and recursive functions. – M.: Mir. 1972.

6. Discrete mathematics and mathematical problems of cybernetics. The science. The main editorial office of physical and mathematical literature. Moscow. 1974.

7. A.A. Evdokimov. On the maximal chain length of an unit n-dimensional cube // Math. Notes. 1970. V. 6. P. 642-648.

8. Berstel J., Karhum¨aki J. Combinatorics on words – A tutorial // Bull. EATCS 79 (2003), 178–229.

9. Lothaire M. Applied Combinatorics on Words. Cambridge University Press, Cambridge: 2005

10. Salomaa A. Pearls of the theory of formal languages – M.: Mir. 1986.

11. Evdokimov A.A. Complete sets of words and their numerical characteristics // Methods of discrete analysis in the study of extremal structures: Coll. sci. op. Novosibirsk: Institute of Mathematics, SB of the USSR AS, 1983. Iss. 39. P. 7 -19.

12. Evdokimov A.A. Investigation of the completeness of sets of words and languages with prohibitions // Bulletin of the Tomsk State University. Application. 2004. № 9 (1). P. 8-12.

13. Evdokimov A. A. Kitaev S. V. Crucial words and the complexity of some extremal problems for sets of prohibited words // J. Comb. Theory. Ser. A. 2004. V. 105. P. 273-289.

14. Evdokimov A.A. Algorithm for recognizing the completeness of a set of words and dynamics of prohibitions // ADM. Annex. 2016. № 9. P. 10–12.

15. Evdokimov A.A. Coding of a finite integer lattice in the class of bounded distortion mappings. // Applied discrete mathematics. Application. Abstracts of reports.2011, № 4, P. 8 – 9.

16. Evdokimov A.A. Metric Attachment Properties and Distance Preserving Codes // Models and Optimization Methods. Novosibirsk: Science, 1988. Tr. / USSR AS Siberian branch. Institute of Mathematics; V. 10. P. 116-132.

17. Evdokimov A.A. Encoding structured information and nesting discrete spaces. // Discrete analysis and operations research. Series 1. 2000. V.7, N 4. P. 48–58.