Information Processing Letters 111 (2011) 930–932

Contents lists available at ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl

On extending de Bruijn sequences

Verónica Becher *, Pablo Ariel Heiber

Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Pabellón I, Ciudad Universitaria, (1428) Buenos Aires, Argentina

ARTICLE INFO

Article history:
Received 28 December 2010
Received in revised form 21 June 2011
Accepted 24 June 2011
Available online 29 June 2011
Communicated by J. Torán

Keywords:
Combinatorial problems
Graph algorithms
De Bruijn sequences
Word problems

ABSTRACT

We give a complete proof of the following theorem:

Every de Bruijn sequence of order n in at least three symbols can be extended to a de Bruijn sequence of order $n+1$. Every de Bruijn sequence of order n in two symbols can *not* be extended to order $n+1$, but it can be extended to order $n+2$.

A (non-cyclic) *de Bruijn sequence* of order n in a k-symbol alphabet is a sequence of length k^n+n-1 such that every sequence of length n occurs exactly once as a consecutive substring [2,12]. See [3] for a fine presentation and history.

In this note we give a complete proof of the following theorem:

Theorem 1. *Every de Bruijn sequence of order n in at least three symbols can be extended to a de Bruijn sequence of order* $n+1$. *Every de Bruijn sequence of order n in two symbols can* not *be extended to order* $n+1$, *but it can be extended to order* $n+2$.

An indirect proof of part of Theorem 1 was first given by Leach in [11] with a topological and measure-theoretic argument on the set of real numbers corresponding to limits of frequency distribution sequences. In [7] Flaxman et al. use the graph-theoretical characterization of de Bruijn sequences to show that the extensions always exist in alphabets with at least three symbols. However, their proof is incomplete because it misses out part of the argument, the demonstration of the property of strong connectedness of the corresponding graph. Also, it does not cover the extension result for the two-symbol alphabet.

Nota bene. The possibility of constructing de Bruijn sequences of order $n+1$ from those of order n has been considered in several algorithms for the two-symbol alphabet, see the survey of Fredricksen [6]. However, instead of producing de Bruijn extensions, these constructions join maximal cycles obtained as an application of Lempel's homomorphism [10]. Efficient implementations of these constructions are given in [4,1].

Notation and preliminaries. Recall that a *hamiltonian* cycle of a graph is a cycle in which each vertex of the graph occurs exactly once. An *eulerian* cycle is a cycle in which each edge of the graph occurs exactly once. A graph that admits an eulerian cycle is called *eulerian*. An undirected graph is *connected* if for every pair of vertices, there is a path between them. A directed graph is *strongly connected* if for every pair of vertices there is a directed path between them. A directed graph is *regular* if each vertex has the same number of incoming and outgoing edges as all other vertices. Given a directed graph G, its *line graph* is a directed graph whose vertices are the edges of G, and whose edges correspond to the directed paths of length two of G.

* Corresponding author.
E-mail addresses: vbecher@dc.uba.ar (V. Becher), pheiber@dc.uba.ar (P. Heiber).

0020-0190/$ – see front matter
doi:10.1016/j.ipl.2011.06.013

For a sequence w we denote its length $|w|$ and we number the positions of w from 1 to $|w|$ The symbol at position i of a sequence w is denoted by $w[i]$, and segments of w are denoted by $w[i \dots j]$, for $1 \leqslant i \leqslant j \leqslant |w|$. A de Bruijn graph of order n, which we denote by G_n, is a graph whose vertices are all sequences of length n, and the edges link overlapping sequences w, v such that $w[2 \dots n] = v[1 \dots n-1]$. The edges of G_n can be labeled with sequences of length $n+1$, such that the edge (w, v) is labeled with $w[1]v = w(v[n])$. Then, each possible sequence of length $n+1$ in k symbols appears in exactly one edge of G_n. Moreover, the line graph of G_n is exactly G_{n+1}. The label of a path $v_1, \dots, v_t$ in G_n is the sequence that contains as subsequences exactly the sequences $v_1, \dots, v_t$, in that order, namely, $v_1[1]v_2[1] \dots v_{t-1}[1]v_t$. Note that the label of a path of length t is a sequence of length $t + n - 1$. If we take a path of length t in G_n and consider the set of $t-1$ traversed edges, it is easy to see that they form a path that has the same label in G_{n+1}. The label of a hamiltonian cycle in G_n is a de Bruijn sequence of order n, and the label of an eulerian cycle in G_n is a de Bruijn sequence of order $n+1$, because an eulerian cycle in G_n is a hamiltonian cycle in G_{n+1}.

We will base the proof of Theorem 1 in the characterization of eulerian directed graphs by I.J. Good [9], which states that a directed graph is eulerian if and only if it is strongly connected and the in-degree and out-degree of each vertex coincide.

Proposition 2 *(Folklore). A directed graph G in which each vertex has its in-degree equal to its out-degree is strongly connected if and only if its underlying undirected graph is connected.*

Proof. The implication from left to right is immediate. Fix G and let $d(v)$ be the in and out-degree of vertex v in G. Let u be an arbitrary vertex of G. Let U be the set of vertices that are accessible from u, that is, the smallest set such that (1) $u \in U$, and (2) for every edge $(v, w) \in G$, if $v \in U$ then $w \in U$. Note that, by definition, there is a directed path from u to each of the vertices in U. Let G' be the subgraph of G induced by U. It is clear that each vertex v in G' has out-degree $d(v)$, because every outgoing edge in G that has its first endpoint in U is in G' by condition (2). Also, the in-degree of each vertex v in G' is less than or equal to $d(v)$ because G' is a subgraph of G. Since the sum of the in-degrees is the same as the sum of the out-degrees in G', the in-degree of vertex v in G' must also be $d(v)$. Therefore, if $w \notin G'$, then there is no edge in G, in any direction, that connects w to any vertex v in G' (because that would make the in- or out-degree of v greater than $d(v)$). Since the underlying graph of G is connected, there is no such w, which implies $G' = G$. Since every vertex in G' is accessible from u, every vertex in G is accessible from u. Since this is valid for any u, G is strongly connected. □

Lemma 3. *A hamiltonian cycle in a de Bruijn graph over an alphabet of at least three symbols can be extended to an eulerian cycle in the same graph.*

Proof. Let H be a hamiltonian path in G_n. Let I be the graph resulting of removing the edges in H from G_n. We first prove that the underlying undirected graph of I is connected. For an arbitrary pair of vertices u, v, we recursively define below a sequence of pairs u_i,v_i, for $0 \leqslant i \leqslant n$, satisfying the following properties:

(1) $u = u_0$ and $v = v_0$.
(2) For each $i < n$, there is an edge from u_i to u_{i+1} in I. Analogously for v_i.
(3) The last i symbols of u_i and v_i coincide.

Let $u_0 = u$ and $v_0 = v$. For $0 \leqslant i < n$, let a_{i+1} be such that setting $u_{i+1} = u_i[2 \dots n]a_{i+1}$ and $v_{i+1} = v_i[2 \dots n]a_{i+1}$ make the edges (u_i, u_{i+1}) and (v_i, v_{i+1}) not belong to H. Such a symbol a_{i+1} exists because each vertex has exactly one of its (at least three) outgoing edges used in H. By definition the edges (u_i, u_{i+1}) and (v_i, v_{i+1}) are in G_n but not in H, hence they are in I. If the last $i < n$ symbols of v_i and u_i coincide, so do the last $i+1$ symbols of u_{i+1} and v_{i+1}, because they are the last i symbols of u_i plus a_{i+1}. By condition (3), $v_n = u_n$, thus, u and v belong to the same connected component of the underlying undirected graph of I. Since this is valid for any pair u,v, the underlying undirected graph of I is connected.

Since G_n and H are regular, I is also regular. Then, by Proposition 2, I is strongly connected. These two properties ensure I is eulerian. Adding an eulerian cycle of I to H gives the desired extension. □

Observation 4. *Lemma* 3 *fails if the alphabet has just two symbols.*

Proof. Consider the de Bruijn graph of order 1. It has just one hamiltonian cycle. The removal of this cycle leaves the two points of the graph disconnected. As argued by Lempel in [10], the same failure occurs at every order, because removing a hamiltonian cycle from the graph leaves the self-loops (that always exist in the vertices of the form a^n) isolated in the residual graph. □

Lemma 5. *A hamiltonian cycle in a de Bruijn graph in two symbols can be extended to an eulerian cycle in the de Bruijn graph of the next order.*

Proof. Let H be a hamiltonian cycle in G_n. Since G_{n+1} is the line graph of G_n, H corresponds to a simple cycle in G_{n+1} that goes through half of the points of G_{n+1}. Let I be the graph that results from removing the edges in H from G_{n+1}. We first prove that the underlying undirected graph of I is connected. Let us call the two symbols of the alphabet 0 and 1. Note that for any sequence s of n symbols, H contains exactly one of the two vertices $s0$ and $s1$ of G_{n+1}. This is because s corresponds exactly to one vertex in G_n, and $s0$ and $s1$ correspond to its outgoing edges. Since H is a hamiltonian cycle in G_n, exactly one of these edges is used in H. This in turn implies that any vertex in G_{n+1} has exactly one successor in H and one successor not in H. For an arbitrary pair of vertices u, v in G_{n+1}, we recursively define below a sequence of pairs u_i, v_i, for $0 \leqslant i \leqslant n+1$, satisfying the following properties:

(1) For each i, at least one of u_i or v_i is not in H.
(2) There is an edge from u to u_0, and there is an edge from v to v_0.
(3) For each $i \leqslant n$, there is an edge from u_i to u_{i+1} in I. Analogously for v_i.
(4) The last i symbols of u_i and v_i coincide.

Let u_0 be any successor of u and v_0 be the successor of v not in H. For $0 \leqslant i \leqslant n$, if $v_i \notin H$, let a_{i+1} be such that $u_i[2 \dots n+1]a_{i+1}$ is not in H. Otherwise, let a_{i+1} be such that $v_i[2 \dots n+1]a_{i+1}$ is not in H. Set $u_{i+1} = u_i[2 \dots n+1]a_{i+1}$ and $v_{i+1} = u_i[2 \dots n+1]a_{i+1}$. By definition, at least one of the endpoints of both (u_i, u_{i+1}) and (v_i, v_{i+1}) is not in H, so both edges are in I. Moreover, at least one of u_{i+1} and v_{i+1} is not in H. If the last $i < n$ symbols of v_i and u_i coincide, so do the last $i+1$ symbols of u_{i+1} and v_{i+1}, because they are the last i symbols of u_i plus a_{i+1}. By condition (4), $u_{n+1} = v_{n+1}$, thus, u and v belong to the same connected component of the underlying undirected graph of I. Since this is valid for any pair u,v, the underlying undirected graph of I is connected.

Every vertex in I has its in-degree equal to its out-degree, because it is a cycle subtracted from a regular graph. So, by Proposition 2, I is strongly connected. These two properties ensure I is eulerian. Adding an eulerian cycle in I to H gives the desired extension. □

We are now ready to give the proof of the already stated Theorem 1:

Theorem 1. *Every de Bruijn sequence of order n in at least three symbols can be extended to a de Bruijn sequence of order $n+1$. Every de Bruijn sequence of order n in two symbols can* not *be extended to order $n+1$, but it can be extended to order $n+2$.*

Proof. De Bruijn sequences of order n correspond exactly to the hamiltonian cycles in de Bruijn graphs G_n. In turn, the hamiltonian cycles in G_{n+1} are exactly the eulerian cycles in G_n. Now the first assertion follows from Lemma 3 and the second from Observation 4 and Lemma 5. □

Definition 6. An infinite sequence in an alphabet of at least three symbols is an *infinite de Bruijn sequence* if it is the inductive limit of extending de Bruijn sequences of order n, for each n. In case of a two-symbol alphabet, an infinite de Bruijn sequence is the limit of extending de Bruijn sequences of order $2n$, for each n.

Corollary 7. *Infinite de Bruijn sequences exist over any alphabet.*

Question 1. Theorem 1 raises naturally the question of giving efficient algorithms to construct extensions of de Bruijn sequences and infinite de Bruijn sequences. According to the proof of Lemmas 3 and 5, the extension problem is just to construct an eulerian cycle in the remaining graph.

Any of the known algorithms for eulerian cycles is usable, even the ancient algorithm of Fleury [8]. However, this may not be the most efficient way of proceeding.

Question 2. Let $occ(w, s) = \#\{j\colon\ s[j \dots j + |w| - 1] = w\}$ be the number of occurrences of a sequence w in a sequence s. For each order n and each length $\ell < n$, what are the maximum and minimum values attained by the frequencies $occ(w, s[1 \dots i])/i$, for all sequences w of length ℓ, extended de Bruijn sequences s of order n, and all $i \leqslant n$?

In [5] Cooper and Heitsch address the problem for $\ell = 1$ in the lexicographically least classical de Bruijn sequence. The range of frequencies in extended de Bruijn sequences may be tighter.

References

[1] F.S. Annexstein, Generating de Bruijn sequences: An efficient implementation, IEEE Transactions on Computers 46 (2) (1997) 198–200.
[2] Nicolaas Gover de Bruijn, A combinatorial problem, Koninklijke Nederlandse Akademie v. Wetenschappen 49 (1946) 758–764, Indagationes Mathematicae 8 (1946) 461–467.
[3] Jean Berstel, Dominique Perrin, The origins of combinatorics on words, European Journal of Combinatorics 28 (2007) 996–1022.
[4] Taejoo Chang, Bongjoo Park, Yun Hee Kim, Iickho Song, An efficient implementation of the D-homomorphism for generation of de Bruijn sequences, IEEE Transactions on Information Theory 45 (4) (1999) 1280–1283.
[5] Joshua Cooper, Christine Heitsch, The discrepancy of the lex-least de Bruijn sequence, Discrete Mathematics 310 (2010) 1152–1159.
[6] Harold Fredricksen, A survey of full length nonlinear shift register cycle algorithms, SIAM Review 24 (2) (1982) 195–221.
[7] Abraham Flaxman, Aram Harrow, Gregory Sorkin, Strings with maximally many distinct subsequences and substrings, Electronic Journal of Combinatorics 11 (2004), #R8.
[8] M. Fleury, Deux problèmes de géométrie de situation, Journal de Mathématiques Élémentaires (1883) 257–261.
[9] I.J. Good, Normal recurring decimals, Journal of the London Mathematical Society 21 (3) (1946) 167–169.
[10] Abraham Lempel, On a homomorphism of the de Bruijn graph and its applications to the design of feedback shift registers, IEEE Transactions on Computers C-19 (12) (1970) 1204–1209.
[11] E.B. Leach, Regular sequences and frequency distributions, Proceedings of American Mathematical Society 11 (1960) 566–574.
[12] Camille Flye Sainte-Marie, Question 48, L'intermédiaire des mathématiciens 1 (1894) 107–110.