数理解析研究所講究録
第2051巻 2017年 143-148

Compression of Palindromes and Regularity.

Kayoko Shikishima-Tsuji
Center for Liberal Arts Education and Research
Tenri University

1 Introduction

In [1], a property of clickstream data at a view of database is discussed and it is shown that page repetitions occur for the majority as a very specific structure, namely in the form of nested palindromes. A kind of function *CFP* (*Compress First Palindrome*) required for an algorithm which extracts these structures in linear time is introduced.

In this paper, we define a rewriting system R which covers *CFP* and consider relation between R and *CFP*. We adopt the name "*wrinkled word*" instead of "nested palindrome" and it means a word which has a *non-trivial* palindrome as a factor. The set of all wrinkled words is regular, though the set of all palindromes is not regular. We also give automata on some alphabets which accept all wrinkled words.

2 Preliminaries

We assume the reader to be familiar with basic concepts as alphabet, word, language, regular expression and automaton (for more details see [2]).

Words together with the operation of concatenation form a free monoid, which is usually denoted by Σ^* for a finite *alphabet* Σ. The *length* of a finite word w is the number of not necessarily distinct symbols it consists of and is written by $|w|$. The *empty* word is denoted by λ and $|\lambda| = 0$. For a word $w = a_1a_2\cdots a_n$ for $a_1, a_2, \cdots, a_n \in \Sigma$, a *factor* of w is $a_i \cdots a_j$, where $1 \le i \le j \le n$, and the *reverse* w^R of w is $a_n \cdots a_2 a_1$. A word $p \in \Sigma^*$ is said to be *palindrome* if $p = p^R$. If a palindrome p is not in $\Sigma \cup \{\lambda\}$, the palindrome p is *non-trivial*, otherwise p is *trivial*. If a word $w \in \Sigma^*$ has at least one non-trivial palindrome as a factor, w is said to be *wrinkled*.

A *string rewriting system* R on Σ is a subset of $\Sigma^* \times \Sigma^*$. We define *reduction* relation on Σ^* that is induced by R is defined as follows: for every $u,v \in \Sigma^*$, $u \rightarrow_R v$ if and only if there exists $(l,m) \in R$ such that for some $x,y \in \Sigma^*$, $u = xly$ and $v = xmy$. By $\rightarrow_R^*$, we denote the reflexive transitive closure of $\rightarrow_R$. If $x \in \Sigma^*$ and there is no $y \in \Sigma^*$ such that $x \rightarrow_R y$, then x is *irreducible*; otherwise, x is *reducible*. The set of all irreducible words with respect to $\rightarrow_R$ is denoted by $IRR(R)$. For $x,y \in \Sigma^*$, if $x \rightarrow_R^* y$ and y is irreducible, then y is a normal form for x. If, for all $w,x,y \in \Sigma^*$, $w \rightarrow_R^* x$ and $w \rightarrow_R^* y$ imply that there exists $z \in \Sigma^*$ such that $x \rightarrow_R^* z$ and $y \rightarrow_R^* z$, we say that R is *confluent*. If for all $w,x,y \in \Sigma^*$, $w \rightarrow_R x$ and $w \rightarrow_R y$ imply that there exists $z \in \Sigma^*$ such that $x \rightarrow_R^* z$ and $y \rightarrow_R^* z$, we say that R is *locally confluent*. If there is no infinite sequence $x_1 \rightarrow_R x_2 \rightarrow_R \cdots$ where $x_1, x_2, \cdots \in \Sigma^*$, then R is said to be *noetherian*.

Two string rewriting systems R and S on Σ are called *equivalence* if $w \rightarrow_R^* z$ implies $w \rightarrow_S^* z$ and $w \rightarrow_S^* z$ implies $w \rightarrow_R^* z$ and then we denote $R \cong S$.

Compress First Palindrome CFP : $\Sigma^* \rightarrow \Sigma^*$ is defined as follows (see [1]) : if $w \in \Sigma^*$ is a wrinkled word, for the left most aba of w where $a \in \Sigma$, $b \in \Sigma \cup \{\lambda\}$, $a \neq b$, there are $u,v \in \Sigma^*$ such that $w = uabav$ and we define $CFP(w) = uv$ and if $w \in \Sigma^*$ is a non-wrinkled word, we define $CFP(w) = w$. Since a wrinkled word w has only finite non-trivial palindromes as factor and $|w| > |CFP(w)|$, then we can define $CFP^\infty(w) = CFP^n(w)$ where n is a large enough number such that $CFP^n(w) = CFP^{n+1}(w)\,(= CFP(CFP^n(w)))$.

3 Regularity of the set of all wrinkled words.

Proposition 1. Let R and S be string rewriting systems $R = \{apa \rightarrow_R a \mid a \in \Sigma,$ p is a palindrome$\}$ and $S = \left\{aba \rightarrow_S a \mid a \in \Sigma, b \in \Sigma \cup \{\lambda\}\right\}$. Then R and S are equivalent.

Proof) Since $R \supseteq S$, it is clear that if $w \rightarrow_S z$ for some $w, z \in \Sigma^*$, then we have $w \rightarrow_R z$.

On the other side, let $w \rightarrow_R z$ for some $w, z \in \Sigma^*$, then there exist $u,v \in \Sigma^*$, $a \in \Sigma$ and a palindrome $p \in \Sigma^*$ such that $w = uapav$ and $z = uav$. If $p \in \Sigma \cup \{\lambda\}$,

then $w \rightarrow_S z$. If $p \notin \Sigma \cup \{\lambda\}$, then p is written $xbcbx^R$ by $c \in \Sigma \cup \{\lambda\}$, $b \in \Sigma$ and $x \in \Sigma^*$. Since $w = uaxbcbx^R av \rightarrow_S uaxbx^R av$ and so on, we have
$w = uaxbcbx^R av \rightarrow_S^* uav = z$.

q.e.d.

The following lemma is well-known (see [3]).

Lemma 1. If a string rewriting system R is noetharian and locally confluent, then R is confluent.

Proposition 2. Let R and S be string rewriting systems $R = \{apa \rightarrow_R a \mid a \in \Sigma,$ p is a palindrome$\}$ and $S = \{aba \rightarrow_S a \mid a \in \Sigma, b \in \Sigma \cup \{\lambda\}\}$. Then R and S are confluent.

Proof) By Proposition 1 and Lemma 1, it is enough to prove that S is locally confluent.

(a) If $w = u_1 v_1 u_2 v_2 u_3$ where $u_1, v_1, u_2, v_2, u_3 \in \Sigma^*$ and $v_1 \rightarrow_S a, v_2 \rightarrow_S b$ for some $a, b \in \Sigma$, then $w \rightarrow_S u_1 a u_2 v_2 u_3$, $w \rightarrow_S u_1 v_1 u_2 b u_3$ and $u_1 a u_2 v_2 u_3 \rightarrow_S u_1 a u_2 b u_3$, $u_1 v_1 u_2 b u_3 \rightarrow_S u_1 a u_2 b u_3$.

(b) If $w = u_1 v u_3$ and $v \in \{aaa, abaa, aaba, abab\}$ where $u_1, u_3 \in \Sigma^*$, $a, b \in \Sigma$. Since $aaa \rightarrow_S^* a$, $abaa \rightarrow_S^* a$, $aaba \rightarrow_S^* a$, $abab \rightarrow_S^* ab$, we have $w \rightarrow_S^* u_1 v' u_3$ where $v' = a$ when $v \in \{aaa, abaa, aaba\}$ and $v' = ab$ when $v = aaba$.

q.e.d.

Proposition 3. Let R' be one of string rewriting systems of Proposition 2. If, for $w, z \in \Sigma^*$ and a natural number n, $CFP^n(w) = z$, then $w \rightarrow_{R'}^* z$. On the other hand, if $w \rightarrow_{R'}^* z$ for $w, z \in \Sigma^*$, then there exist a natural number n and $z' \in \Sigma^*$ such that $CFP^n(w) = z'$ and $w \rightarrow_{R'}^* z'$.

Proof) If $CFP(w) = z$, it is obvious that $w \rightarrow_{R'} z$.
If w has no non-trivial palindrome as a factor, then we have $w = z$ and $CFP(w) = w$. We may assume that w is wrinkled. There exists a finite sequence: $w = w_0$, $CFP(w) = w_1, \cdots, CFP^n(w) = w_n$ such that w_{n-1} is wrinkled and w_n is not wrinkled. Then we have a sequence $w \rightarrow_S w_1 \rightarrow_S \cdots \rightarrow_S w_n$. Since w_n is not wrinkled, w_n has no non-trivial palindrome as a factor and $w_n \in IRR(R')$. By Proposition 2, the rewriting system R' is confluent and then we have $w \rightarrow_{R'}^* w_n$.

The following corollary is clear by Proposition 3.

Corollary 1. Let R be the string rewriting system $R=\{apa \to_R a \mid a \in \Sigma,$ p is a palindrome$\}$. Then we have $CFP^{\infty}(\Sigma^*) = IRR(R)$.

By Corollary 1, we have $CFP^{\infty}(\Sigma^*)=$ {the set of all non-wrinkled words}
The following lemma is well-known (see [3]).

Lemma 2. A string rewriting system R is finite, then $IRR(R)$ is a regular set.

By Proposition 3 and Lemma 2, we have the following theorem.

Theorem 1. The language $\mathcal{NW}$ = {the set of all non-wrinkled words} and the language $\mathcal{W}$ = {the set of all wrinkled words} are both regular.

Proof) The language $\mathcal{NW} = CFP^{\infty}(\Sigma^*)$ is regular and then $\mathcal{W} = (\mathcal{NW})^c$ is regular.
q.e.d.

The set of all palindromes are contest-free but not regular. $\mathcal{W}$ = {the set of all wrinkled words}= $\{w \mid w$ has at least one non-trivial palindrome as a factor$\}$ is regular.

4 Examples.

For $|\Sigma| \le 4$, we give an automaton on Σ which accepts all wrinkled words on Σ.

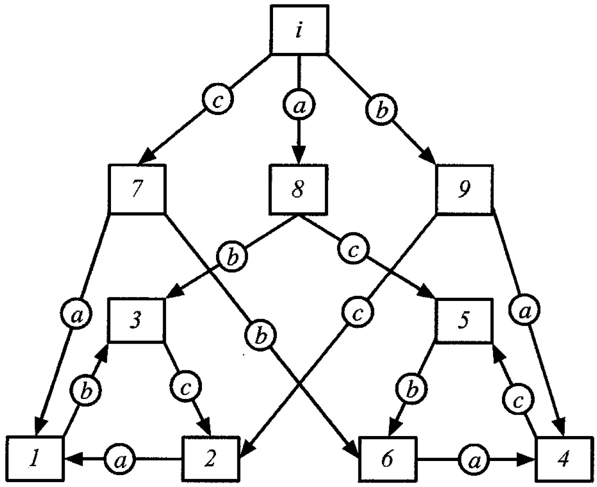
Example 1. If $\Sigma = \{a\}$, then $\mathcal{NW} = \{ w \in \Sigma^* \mid w$ is non-wrinkled word$\}$ is the empty set.

Example 2. If $\Sigma = \{a,b\}$, then $\mathcal{NW}$ is the set $\{ab, ba\}$

Example 3. If $\Sigma = \{a,b,c\}$, then $\mathcal{NW}$ is the set $\{u \in \Sigma^* \mid u$ is a factor of $(abc)^* \cup (acb)^*$ such that $|u| > 1\}$.

By the following automaton $\mathcal{A} = (Q, \Sigma, \delta, i, F)$ (Figure 1), $\mathcal{NW}$ is accepted, where $Q = \{i, 1, 2, \cdots, 9\}$, $i \in Q$ is the initial state and $F = Q$ is a set of final states.

Figure 1



Example 4. If $\Sigma = \{a,b,c,d\}$, then $\mathcal{NW}$ is the language which is accepted by the following automaton $\mathcal{A} = (Q, \Sigma, \delta, i, F)$ (Figure 2), where $Q = \{i, 1, 2, \cdots, 16\}$, $i \in Q$ is the initial state and $F = Q$ is a set of final states. In Figure 2, states $i, 13, 14, 15, 16$ and the following transition functions which start from these states are omitted for simplicity: $a: i \to 13$, $b: i \to 14$, $c: i \to 14$, $d: i \to 15$, $b: 13 \to 2$, $c: 13 \to 8$, $d: 13 \to 10$, $a: 14 \to 11$, $c: 14 \to 7$, $d: 14 \to 1$, $a: 15 \to 5$, $b: 15 \to 6$, $d: 15 \to 9$, $a: 16 \to 3$, $b: 16 \to 12$, $a: 16 \to 4$ (see Figure 3).

Figure 2

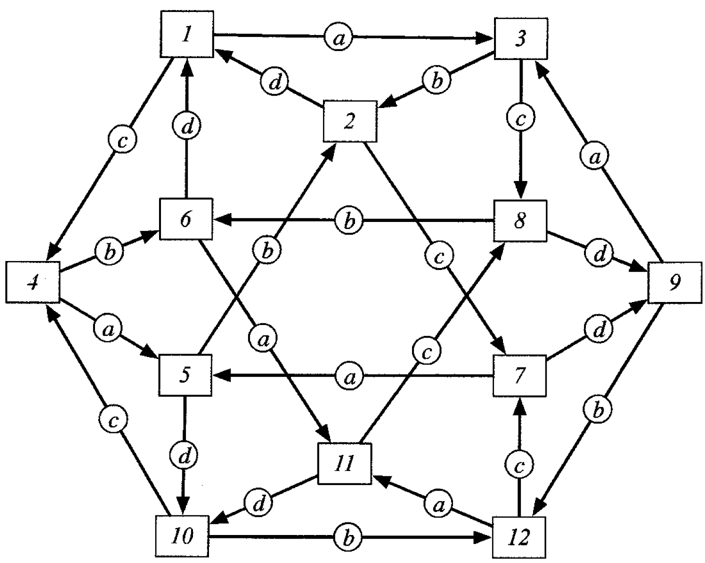
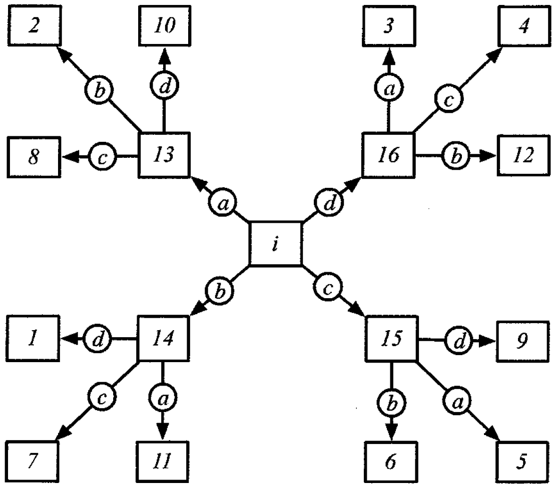


Figure 3



References.

[1] Michel Speiser, Gianluca Antonini, Abderrahim Labbi, Juliana Sutanto: On nested palindromes in clickstream data. The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012. ACM 2012: 1460-1468

[2] Peter Linz, An Introduction to Formal Language and Automata, Jones and Bartlett Publishers, Inc. , USA 2006, ISBN, 0763737984

[3] Ronald V. Book, Friedrich Otto: String-Rewriting Systems. Texts and Monographs in Computer Science, Springer 1993, ISBN 978-3-540-97965-4

Center for Liberal Arts Education and Research
Tenri University
1050 Somanouchi, Tenri,
Nara 632-8510, Japan
E-mail address: tsuji sta.tenri-u.ac.jp

天理大学・総合教育研究センター　辻佳代子