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 $\Sigma^*$ for a finite alphabet $\Sigma$. 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 $\lambda$ and $|\lambda|=0$. For a word $w = a_1a_2\cdots a_n$ for $a_1,a_2,\ldots,a_n \in \Sigma$, a factor of $w$ is $a_i\cdots a_j$, where $1 \leq i \leq j \leq n$, and the reverse $w^R$ of $w$ is $a_n\cdots a_2a_1$. 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 $\Sigma$ is a subset of $\Sigma^* \times \Sigma^*$. We define reduction relation on $\Sigma^*$ 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 $\text{IRR}(R)$.

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 $\text{IRR}(R)$.

Two string rewriting systems $R$ and $S$ on $\Sigma$ 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 $R \equiv S$.

Compress First Palindrome $\text{CFP} : \Sigma^* \rightarrow \Sigma^*$ 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 $\text{CFP}(w) = uv$ and if $w \in \Sigma^*$ is a non-wrinkled word, we define $\text{CFP}(w) = w$. Since a wrinkled word $w$ has only finite non-trivial palindromes as factor and $|w| > |\text{CFP}(w)|$, then we can define $\text{CFP}^{\infty}(w) = \text{CFP}(w)$ where $n$ is a large enough number such that $\text{CFP}^n(w) = \text{CFP}^{n+1}(w) = \text{CFP}(\text{CFP}(w))$.

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 $S = \{aba \rightarrow_S a \mid a \in \Sigma, b \in \Sigma \cup \{\lambda\}\}$. 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 \( xbcx^{\circ} \) by \( c \in \Sigma \cup \{ \lambda \} \), \( x \in \Sigma^{*} \). Since \( w = uaxbcx^{\circ}av \rightarrow_{S} uaxbx^{\circ}av \) and so on, we have \( w = uaxbcx^{\circ}av \rightarrow_{S}^{*} uav = z \).

\[ \text{q.e.d.} \]

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 \( R = \{ apa \rightarrow_{R} a \mid a \in \Sigma, p \text{ is a palindrome} \} \) 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 \( w = u_{1}v_{2}u_{3}v_{2}u_{4} \) where \( u_{1}, v_{2}, u_{3} \in \Sigma^{*} \) and \( v_{1} \rightarrow_{S} a, v_{2} \rightarrow_{S} b \) for some \( a, b \in \Sigma \), then \( w \rightarrow_{S} u_{1}u_{2}u_{3}v_{2}u_{4} \), \( w \rightarrow_{S} u_{1}v_{1}u_{2}u_{3}u_{4} \), and \( u_{1}u_{2}v_{1}u_{3}u_{4} \rightarrow_{S} u_{1}u_{2}u_{3}u_{4} \).

(b) If \( w = u_{1}v_{3} \) 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 \( w \rightarrow_{S}^{*} u_{1}v_{3} \) where \( v' = a \) when \( v \in \{ aaa, abaa, aaba \} \) and \( v' = ab \) when \( v = aaba \).

\[ \text{q.e.d.} \]

**Proposition 3.** Let \( R' \) be one of string rewriting systems of Proposition 2. If, for \( w, z \in \Sigma^{*} \) and a natural number \( n \), \( CFP^{\circ}(w) = z \), 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: \( w = w_{0}, CFP(w) = w_{1}, \ldots, CFP^{n}(w) = w_{n} \) 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 \( R = \{ apa \rightarrow_R a | a \in \Sigma, \) \( p \) is a palindrome\}. Then we have \( CFP^w(\Sigma^*) = IRR(R) \).

By Corollary 1, we have \( CFP^w(\Sigma^*) = \{ \text{the set of all non-wrinkled words} \}\)

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 \( \mathcal{N} \mathcal{W} = \{ \text{the set of all non-wrinkled words} \} \) and the language \( \mathcal{W} = \{ \text{the set of all wrinkled words} \} \) are both regular.

Proof) The language \( \mathcal{N} \mathcal{W} = CFP^w(\Sigma^*) \) is regular and then \( \mathcal{W} = (\mathcal{N} \mathcal{W})^c \) is regular.

q.e.d.

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

4 Examples.

For \( |\Sigma| \leq 4 \), we give an automaton on \( \Sigma \) which accepts all wrinkled words on \( \Sigma \).

Example 1. If \( \Sigma = \{a\} \), then \( \mathcal{N} \mathcal{W} = \{ w \in \Sigma^* | w \text{ is non-wrinkled word} \} \) is the empty set.

Example 2. If \( \Sigma = \{a,b\} \), then \( \mathcal{N} \mathcal{W} \) is the set \( \{ab, ba\} \)

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

By the following automaton \( \mathcal{A} = (Q, \Sigma, \delta, i, F) \) (Figure 1), \( \mathcal{N} \mathcal{W} \) 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.
Example 4. If $\Sigma = \{a,b,c,d\}$, then $\mathcal{N}$ 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: $a:i \rightarrow 13$, $b:i \rightarrow 14$, $c:i \rightarrow 14$, $d:i \rightarrow 15$, $b:13 \rightarrow 2$, $c:13 \rightarrow 8$, $d:13 \rightarrow 10$, $a:14 \rightarrow 11$, $c:14 \rightarrow 7$, $d:14 \rightarrow 1$, $a:15 \rightarrow 5$, $b:15 \rightarrow 6$, $d:15 \rightarrow 9$, $a:16 \rightarrow 3$, $b:16 \rightarrow 12$, $a:16 \rightarrow 4$ (see Figure 3).
References.

