Double-Occurrence Words and Word Graphs

Daniel Cruz, Nataša Jonoska, Brad Mostowski, Lukas Nabergall, Masahico Saito
University of South Florida

bmostowski@mail.usf.edu

December 1, 2017

Overview

Alphabets and Words

Definition

An **alphabet** Σ is a finite or countable set. Elements in Σ are called **symbols**.

Definition

A **word** u over Σ is a finite sequence of symbols in Σ. The set of all words over Σ is denoted Σ^*. The word u is called a **double-occurrence word**, or simply a DOW, if every symbol in Σ that appears in u appears twice. $\Sigma[u]$ is the set of all symbols used by u.

Example

456234 and 121323 are words over $\Sigma = \mathbb{N}$. In particular, 121323 is a DOW.

Factors and Subwords

Definition

Let u be a DOW. v is a **factor** of u if $u = u_1 v u_2$ for some u_1, u_2, in which case we write $v \sqsubset u$.

Example

231 is a factor of 1<u>231</u>23

Definition

Let u be a DOW. A subsequence v of u is a called a **subword**.

Example

233 is a subword of 122133

Definition

Let $u = u_1 \cdots u_n$ be a DOW. $u^R = u_n \cdots u_1$ is called the **reverse** of u.

Ascending Order

We are concerned with the structure of DOWs and hence need a notion for this structure. Ascending order gives that structure.

Definition

Let u be a DOW. u is in **ascending order** if $1, \ldots, i-1$ appear before the first instance of i. The **ascending order representation** of u is a DOW where the ith unique symbol to appear in u is rewritten as i.

Example

123123, 121323, 1234155423 are all words in ascending order. 3434 in ascending order is 1212.

Ascending Order Equivalence

Definition

Two DOWs u and v are **ascending order equivalent** if they have the same ascending order representation, in which case we write $u \sim v$.

Example

- $1212 \sim 2121$
- $456645 \sim 341134$
- $43214321 \not\sim 12344321$

Pattern Appearance

Definition

Let $u \sim 123 \cdots n$. uu (resp. uu^R) is called a **repeat pattern** (resp. **return pattern**) of size n.

Example

- 45677654 is a return pattern of size 4.
- 654321654321 is a repeat pattern of size 6.
- 11 is both a repeat and return pattern of size 1.

We are only concerned with patterns of size > 1.

Pattern Appearance (Cont.)

Definition

Let u be a DOW. The repeat (resp. return) pattern of size i **appears** in u if

$$u = u_1\alpha_1 u_2\alpha_2 u_3$$

where the u_i's are each (possibly empty) factors of u, $\alpha_1 \sim 1 \cdots i$, and $\alpha_2 = \alpha_1$ (resp. $\alpha_2 = \alpha_1^R$).

Example

The repeat pattern of size 2 appears in 11<u>2323</u> and <u>12</u>33<u>12</u>, but not in 121323.

Insertions

Consider a DOW u. We wish to insert a size i repeat/return pattern into u and still obtain a DOW. This process involves the following:

- Take a size i repeat (resp. return) pattern that doesn't use any letters in u and pick a set of indices $1 \leq j, k \leq |u| + 1$
- Insert the first half (resp. second half) of this pattern before the jth (resp. k) letter in u for some j (resp. k)
- The resulting word is denoted $u \cdot \rho_i(j, k)$ (resp. $u \cdot \tau_i(j, k)$)

Definition

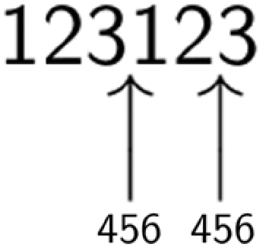
Let u be a DOW. $\rho_i(j, k)$ (resp. $\tau_i(j, k)$) denotes the insertion of a repeat (resp. return) pattern of size i in indices j, k.

Remark

The notation in practice is not an issue since we are only concerned with ascending order representation of DOWs.

Example

Let $u = 123123$. Then, $u \cdot \rho_3(4, 6) \sim 123\underline{456}12\underline{456}3$



Shifting Codes

Definition

Let p be a subword of u, and let I be an insertion. The **shifting code of p under I in u** is a sequence $c_I(p) = c_1, c_2, \ldots, c_{|p|}$ where c_ℓ is the number of letters which the ℓth letter in p is shifted after inserting I into u.

Proposition

Let $f_i(j,k), g_i(j,k)$ be insertions. If $\exists a \in \Sigma[u]$ s.t. $c_f(aa) \neq c_g(aa)$ and $c_f(a) = c_g(a)$ for either instance of a, then $u \not\sim v$.

Example

Consider $1212 \cdot \tau_2(4,5) \sim 12134243$. Then,

$$c_\tau(11) = 0, 0$$
$$c_\tau(22) = 0, 2$$

Shifting Codes (Cont.)

Example

Consider the following DOWs:

$$1212 \cdot \tau_2(4,5) \sim 12134243$$
$$1212 \cdot \rho_2(3,5) \sim 12341234$$

Then,

$$c_\tau(11) = 0, 0$$
$$c_\rho(11) = 0, 2$$

hence the two DOWs are not equivalent.

Pattern Destruction

From the previous example, it is possible that a pattern that appears in a DOW u may not appear in $u \cdot f_i(j,k)$.

Definition

Let p be a repeat/return pattern. Suppose p appears in u and $u \cdot f_i(j,k) \sim v$. We say $f_i(j,k)$ **destroys** p if p is not a repeat/return pattern in v and **completely destroys** p if p contains no instance of a nontrivial pattern in v.

Example

$1212 \cdot \rho_1(4,4) \sim 121332$
$\rho_1(4,4)$ completely destroys 1212.

$123321 \cdot \rho_1(3,7) \sim 12433214$
$\rho_1(3,7)$ destroys 123321. 1221 still appears in 123321 in the new word.

Reverse Proposition

Proposition

Let u and v be two double occurrence words. Then,
$(u \cdot f_i(j,k))^R \sim u^R \cdot f_i(|u|+2-k, |u|+2-j)$

Example

Let $u = 121332$. Then,

$$u \cdot \rho_2(2,5) \sim 1452134532$$
$$u^R \cdot \rho_2(3,6) \sim 2345312451$$
$$(u^R \cdot \rho_2(3,6))^R \sim 1542135432$$

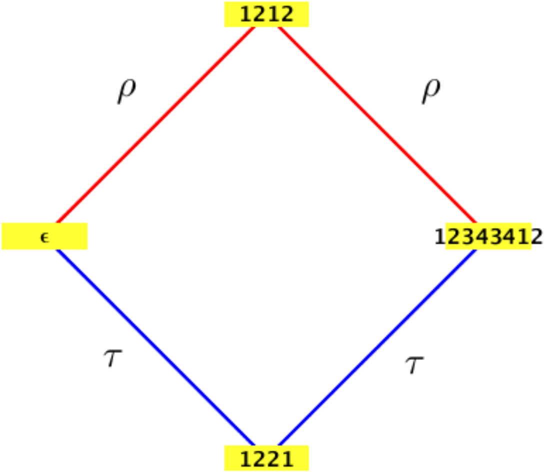
Word Graphs

Based on the notion of insertions on double-occurrence words, we can construct graphs on these DOWs, drawing edges between words when there is an insertion that takes one DOW to another.

Definition

A word graph $G = (V, E)$ is a graph with the vertices DOWs and the edges representing repeat/return insertions. For $v_1, v_2 \in V$. (v_1, f, v_2) is an edge in E if there is an insertion $f_i(j, k)$ such that $v_1 \cdot f_i(j, k) \sim v_2$.

An Example of a Word Graph



Classification of Word Graphs

- We would like to figure the structure of the word graph on the set of all double-occurrence words.
- Where do we start? Let's try to classify cycles on this graph. The simplest cycles? Digons.
- Under what situations do digons appear? In particular, given a word graph $G = (V, E)$ on double-occurrence words, does there exist $u_1, u_2 \in V$ such that $(u_1, \rho, u_2), (u_1, \tau, u_2) \in E$?
- Another way of stating the above: Does there exist a situation where $u \cdot \rho_i(j, k) \sim u \cdot \tau_i(m, n)$?

Classification of Digons

Naturally we suspect such a situation does not exist for $i > 1$. By counting the number of patterns that appear in $u \cdot \rho_i(j,k)$ and $u \cdot \tau_i(m,n)$, the following conclusion can be reached:

Lemma

Let u be a double-occurence word and $i > 1$. Then, $u \cdot \rho_i(j,k) \sim u \cdot \tau_i(m,n)$ implies the following statements hold:

1. *u contains an instance of a size i repeat pattern, p, and a size i return pattern, q, with $2 \leq i \leq 3$.*
2. *$\rho_i(j,k)$ completely destroys p and $\tau_i(m,n)$ completely destroys q.*

Classification of Digons (Cont.)

Theorem

For $i > 2$, there exists no pair of insertions $\rho_i(j,k)$ and $\tau_i(m,n)$ such that $u \cdot \rho_i(j,k) \sim u \cdot \tau_i(m,n)$.

Proof.

Suppose otherwise. Then, u has a size 3 return pattern, say $p = 123321$, which is completely destroyed by τ. Then, $c_\tau(11)$ is $0,6$. But this means $c_\rho(11) = 0,6$ or $3,3$, whence $c_\rho(p) = 0,0,0,6,6,6$ or $3,3,3,3,3,3,3$. Then $c_\tau(p)$ is either:

$$0,0,0,6,6,6$$
$$0,0,3,3,6,6$$
$$0,3,3,3,3,6$$

Then, τ does not completely destroy p, a contradiction. □

Classification of Digons (Continued)

The previous statement holds for $i \geq 2$, but the proof requires consideration of 24 different cases!

Corollary

For $i \geq 2$, there exists no pair of insertions $\rho_i(j, k)$ and $\tau_i(m, n)$ such that $u \cdot \rho_i(j, k) \sim u \cdot \tau_i(m, n)$.

Acknowledgements

Special thanks to the USF Math/Bio Research Group.