Transformata Burrowsa-Wheelera uogólnionych słów Thuego-Morse'a

Paweł Papis

Wydział Matematyki Informatyki i Mechaniki Uniwersytetu Warszawskiego

13 listopada 2008

Dygresja ortograficzna

- „Nazwiska zakończone na -e nieme (tzn. niewymawiane) otrzymują polskie końcówki po apostrofie" - Słownik Ortograficzny PWN
- Wniosek: Morse'a, a nie Morsa
- „Nazwiska kończące się w wymowie na e (...) odmieniamy w liczbie pojedynczej jak przymiotniki"
- Stąd: Thuego

Transformata Burrowsa-Wheelera

- Jest podstawą algorytmu BZIP2, alternatywą dla opatentowanego w USA kodowania arytmetycznego.

BWT(barackobama)

- barackobama
- abarackobam
- mabarackoba
- amabarackob
- bamabaracko
- obamabarack
- kobamabarac
- ckobamabara
- ackobamabar
- rackobamaba
- arackobamab

BWT(barackobama)

- abarackobam
- ackobamabar
- amabarackob
- arackobamab
- bamabaracko
- barackobama
- ckobamabara
- kobamabarac
- mabarackoba
- obamabarack
- rackobamaba

BWT(barackobama)

- abarackoba **m**
- ackobamaba **r**
- amabaracko **b**
- arackobama **b**
- bamabarack **o**
- barackobam **a**
- ckobamabar **a**
- kobamabara **c**
- mabarackob **a**
- obamabarac **k**
- rackobamab **a**

BWT(barackobama)

- 0 abarackoba **m**
- 1 ackobamaba **r**
- 2 amabaracko **b**
- 3 arackobama **b**
- 4 bamabarack **o**
- 5 barackobam **a**
- 6 ckobamabar **a**
- 7 kobamabara **c**
- 8 mabarackob **a**
- 9 obamabarac **k**
- 10 rackobamab **a**

BWT(barackobama)

- $BWT(barackobama) = mrbboaacaka$
- $i_{BWT}(barackobama) = 5$

Definicja

- Słowo Thuego-Morse'a:
  0110100110010110100101100110100110010110011010010 11...
- Słowem Thuego-Morse'a nazywamy ciąg postaci $(s_n)_{n \in N}$, gdzie:

$$s_n = \begin{cases} 0 & \text{jeśli n ma parzystą liczbę jedynek w zapisie binarnym} \\ 1 & \text{jeśli n ma nieparzystą liczbę jedynek w zapisie binarnym} \end{cases}$$

Kilka prostych własności

- $s_{2i} = s_i$ oraz $s_{2i+1} = 1 - s_i$, co pociąga za sobą $(s_{2n})_{n \in N} = (1 - s_{2n+1})_{n \in N} = (s_n)_{n \in N}$
- Jeśli $s_i = s_{i+1}$, to i jest nieparzyste
- W pięciu kolejnych literach słowa Thuego-Morse'a istnieją dwa takie same następujące po sobie znaki (nie ma podsłów: 01010 oraz 10101)

Rozważane słowa

- Prefiksy $(s_n)_{n \in N}$ o długościach 2^i:
- $tm(0) = 0$
- $tm(1) = 01$
- $tm(2) = 0110$
- $tm(3) = 01101001$
- $tm(4) = 0110100110010110$
- $tm(5) = 01101001100101101001011001101001$
- ...

Rozważane słowa

- Prefiksy $(s_n)_{n \in N}$ o długościach 2^i:
- $tm(0) = 0$
- $tm(1) = 0\ 1$
- $tm(2) = 01\ 10$
- $tm(3) = 0110\ 1001$
- $tm(4) = 01101001\ 10010110$
- $tm(5) = 0110100110010110\ 1001011001101001$
- ...
- bo $s_{2^k+n} = 1 - s_n$, jeśli $n < 2^k$

Morfizm deiniujący słowa Thuego-Morse'a

- Przypomnienie faktu sprzed paru chwil: $s_{2i} = s_i$ oraz $s_{2i+1} = 1 - s_i$, a skoro tak, zdefiniujmy następujący morfizm:

$$\phi : \begin{cases} 0 \to 01 \\ 1 \to 10 \end{cases}$$

- Wówczas $(s_n)_{n \in N} = \phi^k(0)$ przy $k \to \infty$, dodatkowo: $tm(k) = \phi^k(0)$

BWT słów Thuego-Morse'a

- $tm(0) = 0$, $BWT(tm(0)) = 0$
- $tm(1) = 01$, $BWT(tm(1)) = 10$
- $tm(2) = 0110$, $BWT(tm(2)) = 1010$
- $tm(3) = 01101001$, $BWT(tm(3)) = 1^2 01010^2$
- $tm(4) = 0110100110010110$,
  $BWT(tm(4)) = 1^4 0^2 10101^2 0^4$
- $tm(5) = 0110100110010110100101100110100 1$,
  $BWT(tm(5)) = 1^8 0^4 1^2 01010^2 1^4 0^8$
- $BWT(tm(6)) = 1^{16} 0^8 1^4 0^2 10101^2 0^4 1^8 0^{16}$
- ...

BWT słów Thuego-Morse'a

- $BWT(tm(0)) = 0$
- $BWT(tm(1)) = 10$
- $BWT(tm(2)) = 1010$
- $BWT(tm(3)) = 1^2 01010^2$
- $BWT(tm(4)) = 1^4 0^2 10101^2 0^4$
- $BWT(tm(5)) = 1^8 0^4 1^2 01010^2 1^4 0^8$
- $BWT(tm(6)) = 1^{16} 0^8 1^4 0^2 10101^2 0^4 1^8 0^{16}$
- ...

BWT słów Thuego-Morse'a

- $BWT(tm(k)) = v(k)\bar{v}(k)^R$, gdzie:
- $v(k) = 1^{2^{k-2}}0^{2^{k-3}} \ldots 1^4 0^2 10$, gdy k parzyste
- $v(k) = 1^{2^{k-2}}0^{2^{k-3}} \ldots 0^4 1^2 01$, gdy k nieparzyste
- Dlaczego BWT słów Thuego-Morse'a jest postaci $v\bar{v}^R$?

Twierdzenie

Twierdzenie: Jeśli $w \in \{0,1\}^*$, to $BWT(w\bar{w}) = v\bar{v}^R$.
Trochę definicji:

- $cshift(w_0 w_1 \dots w_{n-1}, i) = w_i w_{i+1} \dots w_{n-1} w_0 w_1 \dots w_{i-1}$
- $CShift(w_0 w_1 \dots w_{n-1}) =$
  $= \bigcup_{i=0,\dots,n-1} \{cshift(w_0 w_1 \dots w_{n-1}, i)\}$

Twierdzenie - dowód

Twierdzenie: Jeśli $w \in \{0,1\}^*$, to $BWT(w\bar{w}) = v\bar{v}^R$.

- Niech w' będzie cyklicznym przesunięcia słowa $w\bar{w}$:

$$w\bar{w} = w_0 w_1 \dots w_{n-1}(1-w_0)(1-w_1)\dots(1-w_{n-1}),$$

wtedy:

- dla $0 \leqslant i \leqslant n-1$:

$$w' = w'_0 w'_1 \dots w'_{2n-1} = cshift(w\bar{w}, i) =$$

$$= w_i \dots w_{n-1}(1-w_0)(1-w_1)\dots(1-w_{n-1})w_0 \dots w_{i-1},$$

- dla $n \leqslant i \leqslant 2n-1$:

$$w' = w'_0 w'_1 \dots w'_{2n-1} = cshift(w\bar{w}, i) =$$

$$= (1-w_{i-n})(1-w_{i-n+1})\dots(1-w_{n-1})w_0 w_1 \dots w_{n-1}(1-w_0)\dots$$
$$\dots(1-w_{i-n-1}).$$

Twierdzenie - dowód

Twierdzenie: Jeśli $w \in \{0,1\}^*$, to $BWT(w\bar{w}) = v\bar{v}^R$.

- Zatem dla wszystkich przesunięć cyklicznych słowa $w\bar{w}$ jest spełnione:
$$w'_i = 1 - w'_{i+n}.$$
- Z czego wynika, że dla każdego $w' \in CShift(w\bar{w})$ istnieje słowo $\bar{w}' \in CShift(w\bar{w})$:
$$\bar{w}' = cshift(w', n).$$
- Ponieważ $\#_0(w\bar{w}) = \#_1(w\bar{w})$, to dokładnie n słów będących cyklicznymi przesunięciami $w\bar{w}$ zaczyna się od 1, tyle samo zaczyna się od 0.

Twierdzenie - dowód

Twierdzenie: Jeśli $w \in \{0,1\}^*$, to $BWT(w\bar{w}) = v\bar{v}^R$.

- Niech u będzie i-tym w porządku $\preceq$ cyklicznym przesunięciem $w\bar{w}$, zaczynającym się od 0.
- Wówczas $\bar{u} = cshift(u, n)$ (rozpoczynające się 1) będzie i - tym w porządku $\succeq$ cyklicznym przesunięciem, czyli $2n - i + 1$-szym w porządku $\preceq$.

Twierdzenie - dowód

Twierdzenie: Jeśli $w \in \{0,1\}^*$, to $BWT(w\bar{w}) = v\bar{v}^R$.

- Niech $v(i)$ będzie ostatnim znakiem i-tego w porządku $\preceq$ przesunięcia cyklicznego $w\bar{w}$.
- Wówczas $v(i) = 1 - v(2n - i + 1)$.
- Wynika z tego:

$$BWT(w\bar{w}) = v(1)v(2)\dots v(2n) =$$

$$v(1)v(2)\dots v(n)(1-v(n))(1-v(n-1))\dots(1-v(2))(1-v(1)) = v\bar{v}^R$$

Fakt

Jeśli $x, y \in \{0,1\}^*$ i $x \preceq y$, to $\phi(x) \preceq \phi(y)$.

Przykład

- $\phi(0011) = 01011010$
- $\phi(0110) = 01101001$
- $\phi(1001) = 10010110$
- $\phi(1100) = 10100101$
- Jeśli x kończy się na $0/1$, $\phi(x)$ kończy się na $1/0$

Przykład

- $\phi(0011) = 01011010$
- 10110100
- $\phi(0110) = 01101001$
- 11010010
- $\phi(1001) = 10010110$
- 00101101
- $\phi(1100) = 10100101$
- 01001011

00 . . . i 0100 . . .

- Ile jest cyklicznych przesunięć $tm(k)$ zaczynających się od 00 / 0100?
- Czy zawsze kończą się 1?

Indeksy BWT słów Thuego-Morse'a

- (numerowane od 0)
- $i_{BWT}(tm(0)) = 0$
- $i_{BWT}(tm(1)) = 0$
- $i_{BWT}(tm(2)) = 1$
- $i_{BWT}(tm(3)) = 3$
- $i_{BWT}(tm(4)) = 7$
- $i_{BWT}(tm(5)) = 15$
- ...
- $i_{BWT}(tm(k)) = 2^{k-1} - 1$ (dla $k \geqslant 1$)
- Wniosek: $tm(k)$ jest ostatnim w porządku leksykograficznym elementem $CShift(tm(k))$ zaczynającym się od 0.

Definicja

- $s_b(n)$ – suma cyfr zapisu liczby n w systemie liczbowym o podstawie b
- Uogólnione słowa Thuego-Morse'a:

$$(utm_{b,m}(n))_{n\geqslant 0} = (s_b(n) \ mod \ m)_{n\geqslant 0}$$

- Słowa Thuego-Morse'a są szczególnym przypadkiem: (przy $b = 2$ oraz $m = 2$)

Definicja - morfizm

$$\phi_{b,m} : \begin{cases} 0 \to (0 \bmod m)(1 \bmod m)\ldots((b-1) \bmod m) \\ 1 \to (1 \bmod m)(2 \bmod m)\ldots(b \bmod m) \\ 2 \to (2 \bmod m)(3 \bmod m)\ldots((b+1) \bmod m) \\ \ldots \\ m-1 \to ((m-1) \bmod m)\ldots((m+b-2)) \bmod m) \end{cases}$$

Wówczas $(utm_{b,m}(n))_{n\in N} = \phi^k_{b,m}(0)$ przy $k \to \infty$

Definicja - morfizm, przykład

- Dla $b = 4$ i $m = 3$:

$$\phi_{b,m} : \begin{cases} 0 \longrightarrow 0120 \\ 1 \longrightarrow 1201 \\ 2 \longrightarrow 2012 \end{cases}$$

- $(utm_{4,3}(n))_{n \in N} = 0120120120120120120120120120 \ldots$

Rozważane słowa

- Prefiksy uogólnionych słów Thuego-Morse'a długości b^i.
- Niech $t_{b,m}(i)$ oznacza prefiks słowa $(utm_{b,m}(n))_{n\in N} = \phi^k_{b,m}(0)$ długości b^i.
- Gdy $b = 4$ i $m = 3$:
  - $t_{3,4}(1) = 0120$, $BWT(t_{3,4}(1) = 20^21$
  - $t_{3,4}(2) = (012)^50$, $BWT(t_{3,4}(2) = 2^50^61^5$
  - $t_{3,4}(3) = (012)^{21}0$, $BWT(t_{3,4}(3) = 2^{21}0^{22}1^{21}$
  - Podobna sytuacja ma miejsce zawsze, gdy $b \equiv 1 \pmod{m}$.
  - Jaki jest i_{BWT} w takim przypadku?

$BWT(t_{4,2})$

- $BWT(t_{4,2}(1)) = 1^2 0^2$
- $BWT(t_{4,2}(2)) = 1^4 0^2 1^2 0^2 1^2 0^4$
- $BWT(t_{4,2}(3)) = 1^{16} 0^2 1^2 0^2 1^2 0^2 1^6 0^6 1^2 0^2 1^2 0^2 1^2 0^{16}$
- $BWT(t_{4,2}(4)) = 1^{64} 0^6 1^6 0^{10} 1^2 0^2 1^2 0^2 1^2 0^6 1^{26} 0^{26} 1^6 0^2 1^2 0^2 1^2 0^2 1^{10} 0^6 1^6 0^{64}$
- $BWT(t_{4,2}(5)) = 1^{256} 0^{26} 1^{26} 0^{38} 1^6 0^6 1^{10} 0^2 1^2 0^2 1^2 0^2 1^6 0^{26} 1^{102} 0^{102} 1^{26} 0^6 1^2 0^2 1^2 0^2 1^2 0^{10} 1^6 0^6 1^{38} 0^{26} 1^{26} 0^{256}$

$BWT(t_{6,2})$

- $BWT(t_{6,2}(1)) = 1^3 0^3$
- $BWT(t_{6,2}(2)) = 1^9 0^3 1^6 0^6 1^3 0^9$
- $BWT(t_{6,2}(3)) = 1^{54} 0^3 1^6 0^6 1^3 0^6 1^{30} 0^{30} 1^6 0^3 1^6 0^6 1^3 0^{54}$
- $BWT(t_{6,2}(4)) =$
  $1^{324} 0^{15} 1^{30} 0^{39} 1^3 0^6 1^6 0^3 1^6 0^{30} 1^{186} 0^{186} 1^{30} 0^6 1^3 0^6 1^6 0^3 1^{39} 0^{30} 1^{15} 0^{324}$
- $BWT(t_{6,2}(5)) =$
  $1^{1944} 0^{93} 1^{186} 0^{231} 1^{15} 0^{30} 1^{39} 0^3 1^6 0^6 1^3 0^6 1^{30} 0^{186} 1^{1110}$
  $0^{1110} 1^{186} 0^{30} 1^6 0^3 1^6 0^6 1^3 0^{39} 1^{30} 0^{15} 1^{231} 0^{186} 1^{93} 0^{1944}$

$BWT(t_{8,2})$

- $BWT(t_{8,2}(1)) = 1^4 0^4$
- $BWT(t_{8,2}(2)) = 1^{16} 0^4 1^{12} 0^{12} 1^4 0^{16}$
- $BWT(t_{8,2}(3)) = 1^{128} 0^4 1^{12} 0^{12} 1^4 0^{12} 1^{84} 0^{84} 1^{12} 0^4 1^{12} 0^{12} 1^4 0^{128}$
- $BWT(t_{8,2}(4)) = 1^{1024} 0^{28} 1^{84} 0^{100} 1^4 0^{12} 1^{12} 0^4 1^{12} 0^{84} 1^{684} 0^{684} 1^{84} 0^{12} 1^4 0^{12} 1^{12} 0^4 1^{100} 0^{84} 1^{28} 0^{1024}$
- $BWT(t_{8,2}(5)) = 1^{8192} 0^{228} 1^{684} 0^{796} 1^{28} 0^{84} 1^{100} 0^4 1^{12} 0^{12} 1^4 0^{12} 1^{84} 0^{684} 1^{5460} 0^{5460} 1^{684} 0^{84} 1^{12} 0^4 1^{12} 0^{12} 1^4 0^{100} 1^{84} 0^{28} 1^{796} 0^{684} 1^{228} 0^{8192}$

Hipoteza dla $m = 2$

- Jeśli $b = 4k + 2$ dla $k \geqslant 0$, to istnieje w takie, że $t_{b,2}(i) = w\bar{w}$.
- Jeśli $b = 4k$, to $t_{b,2}(i) = ww$.
- $BWT(t_{b,2}(i))$ dla $b \geqslant 4$ parzystych ma następującą postać:

$$BWT(t_{b,2}(i)) = v_{b,2}(i)\bar{v}_{b,2}(i))^R,$$

- Gdzie: $v_{b,2}(i) = 1^{k_{i,1}}0^{k_{i,2}}1^{k_{i,3}}\ldots 0^{k_{i,4i-6}}1^{k_{i,4i-5}}$ (składa się z $4i-5$ spójnych bloków 0 lub 1 dla $i \geqslant 2$ i z jednego bloku dla $i = 1$)

Hipoteza dla $m = 2$

- Długość pierwszego bloku:

$$k_{1,1} = \frac{b}{2}, \ k_{i,1} = \frac{b^i}{4} \ dla \ i \geqslant 2.$$

- Długość drugiego bloku:

$$k_{2,2} = \frac{b}{2}, \ k_{i,2} = \frac{2}{b-2} \cdot k_{i,3} \ dla \ i \geqslant 2.$$

- Długość trzeciego bloku:

$$k_{2,3} = \frac{b^2 - 2b}{4}, \ k_{i,3} = k_{i-1,4i-9} \ dla \ i \geqslant 3.$$

- Długość czwartego bloku:

$$k_{i,4} = \begin{cases} k_{2,3} & \text{dla } i = 3, \\ k_{i-1,4} \cdot b + \frac{b}{2} & \text{dla } i \geqslant 4 \text{ parzystych}, \\ k_{i-1,4} \cdot b - \frac{b}{2} & \text{dla } i \geqslant 5 \text{ nieparzystych}. \end{cases}$$

Hipoteza dla $m = 2$

- Długość bloków od piątego do przedostatniego:

$$k_{i,j} = k_{i-1,j-3} \text{ dla } 5 \leqslant j \leqslant 4i - 6, \ i \geqslant 3.$$

- Długość ostatniego bloku:

$$k_{i,4i-5} = \begin{cases} \frac{b^2-2b}{4} & \text{dla } i = 2, \\ k_{i-1,4i-9} \cdot b - \frac{b^2-2b}{4} & \text{dla } i \geqslant 3 \text{ nieparzystych}, \\ k_{i-1,4i-9} \cdot b + \frac{b^2-2b}{4} & \text{dla } i \geqslant 4 \text{ parzystych}. \end{cases}$$

Cele pracy

- Postawienie jak największej liczby hipotez dotyczących postaci transformaty Burrowsa-Wheelera uogólnionych słów Thuego-Morse'a.
- Podjęcie próby udowodnienia postawionych hipotez.