Combinatorics on words and its applications

Ľubomíra Balková

Department of Mathematics, FNSPE, Czech Technical University in Prague

May 4, 2012

Program

Program

Definition of Thue-Morse sequence

Definition

We denote by $\mathbf{u}_{TM} = (u_n)_{n=0}^{+\infty}$ the (Prouhet-)Thue-Morse sequence on $\{0,1\}$ defined recursively by

$$u_0 = 0, \quad u_{2n} = u_n \quad \text{and} \quad u_{2n+1} = 1 - u_n \quad \text{for } n \geq 0.$$

$$\mathbf{u}_{TM} = \begin{array}{cccccccccc} u_0 & u_1 & u_2 & u_3 & u_4 & u_5 & u_6 & u_7 & u_8 & \dots \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & \dots \end{array}$$

Binary representation

Theorem

Denote by $s_2(n)$ *the sum of digits in the binary representation of* $n \in \mathbb{N}$. *Then*

$$\mathbf{u}_{TM} = (s_2(n) \mod 2)_{n=0}^{+\infty}.$$

Substitution

Define the substitution φ on $\{0,1\}$ by $\varphi(0) = 01, \quad \varphi(1) = 10.$

Theorem

The Thue-Morse sequence is the unique fixed point of φ that begins with 0.

$$\mathbf{u}_{TM} = 0110100110010110 1001\ldots$$

Birth of Combinatorics on Words

Questions of Axel Thue:

1. Is there a binary cube-free or even overlap-free sequence?
2. Is there a ternary square-free sequence?

Remark

He answered both questions affirmatively in 1906 in an obscure Norwegian journal ⇒ long time not known, rediscovered by Morse in 1921. Thue explained he had no particular application in mind, but he thought the problem was interesting enough in itself to deserve attention. Starting point of COMBINATORICS ON WORDS.

Theorem (Thue)

The Thue-Morse sequence is overlap-free. In other words, it does not contain awawa, where $a \in \{0,1\}$ *and* $w \in \{0,1\}^*$.

Corollary

The Thue-Morse sequence is cube-free.

For $n \geq 1$ let v_n be the number of 1's between the n-th and $(n+1)$-st occurrence of 0 in the sequence $\mathbf{u}_{TM}$. Denote $\mathbf{v} = (v_n)_{n=1}^{+\infty}$.

$$\mathbf{u}_{TM} = 0110100110010110\ldots$$

$$\mathbf{v} = 21020121012\ldots$$

Corollary

The sequence $\mathbf{v}$ *is square-free.*

Program

One-way and collision-free function

Definition

$f: X \to Y$ is called **one-way** if

1. for any $x \in X$ it is easy to compute $y = f(x)$,
2. for any $y \in f(X)$ it is computationally infeasible to find a preimage, *i.e.*, $x \in X$ such that $y = f(x)$.

Definition

$f: X \to Y$ is called **collision-free** if it is computationally infeasible to find $x, x' \in X,\ x \neq x'$ such that $f(x) = f(x')$.

Definition of hash function

Definition

Let $N, n \in \mathbb{N},\ n << N$ *and* $f : \{0,1\}^N \to \{0,1\}^n$ *is called* hash function *if* f *is one-way and collision-free and behaves as a random oracle.*

$f(M)$ *is called* hash *of the message* M.

Remark

Usually $N = 2^{64} - 1,\ N = 2^{128} - 1$ *and n hundreds of bits (for MD5/SHA-1/SHA256/SHA512 it is 128/160/256/512 bits).*

Birthday paradox

- $P(365, 23) = 0,507$ and $P(365, 30) = 0,706$
- the probability that any two participants of a party of k guests celebrate their birthday the same day
  $P(365, k) = 1 - \frac{365 \cdot 364 \cdot \ldots (365-k+1)}{365^k}$

Damgard-Merkle construction - Crypto 1989

iterative hash functions based on compression functions

- message M cut into m-bits blocks $m_1, m_2, \ldots, m_k$
- Damgard-Merkle strengthening - padding M with 1, then zeros and the length of M
- compression function:

$$h_0 = IV, \quad h_i = f(h_{i-1}, m_i)$$

- output: h_k or its part
- collision-free compression function $\Rightarrow$ collision-free hash function

Program

Attack on repeating contexts

- if $h_{i-1} = h_i = f(h_{i-1}, m_i)$, then hash codes of $m_1 \dots m_{i-1} m_i m_{i+1} \dots m_k$ and $m_1 \dots m_{i-1} m_i^{\ell} m_{i+1} \dots m_k$ are the same $\Rightarrow$ the second preimage of the same length as M may be found with complexity $t \cdot 2^{n/2+1} + 2^{n-t+1}$ for messages of length 2^t close to $2^{n/2}$
- for SHA-1 the second preimage of a message of length 2^{60} may be found with complexity 2^{106} instead of 2^{160}

J. Kelsey, B. Schneier, *Second preimages on n-bit hash functions for much less than* 2^n *work*

Dithered hash functions

$h_i = f(h_{i-1}, m_i, d_i)$

1. counter: $d_i := i$
2. random sequence: $d_i := r_i$
3. alternation of 0 and 1
4. **square-free and abelian square-free sequences**
   R. Rivest, *Abelian square-free dithering for iterated hash functions*

Square-free words

square-free word does not contain *ww*

Example

abracadabra *OK*, banana *NO*

- no square-free infinite words over $\{0, 1\}$
- there exist square-free infinite words over $\{0, 1, 2\}$

Example

Thue-Morse word $\mathbf{u}_{TM} = 0110100110010110\ldots$ *is overlap-free*
$\Rightarrow \mathbf{v} = 2102012\ldots$ *is square-free*

Abelian square-free words

abelian square-free word does not contain ww', where w' is a permutation w

Example

abelianalien *NO, it contains* alien *and* elian

Example

$$\textit{magic word } S = \begin{array}{c} abcacdcbcdcadcdbdaba \\ cabadbabcbdbcbacbcdc \\ acbabdabacadcbcdcacd \\ bcbacbcdcacdcbdcdadbdcbca \end{array} \textit{ of length 85}$$

denote σ *the cyclic shift* $\sigma(abcacd) = bcdbda$, *then* **Keränen's abelian square-free word** *is a fixed point of the morphism*

$$a \to S,\ b \to \sigma(S),\ c \to \sigma^2(S),\ d \to \sigma^3(S)$$

Thank you for attention and Happy Birthday!