部分列数え上げデータ圧縮法とその関連法について

横尾英俊

群馬大学

2017 年 7 月 13 日

概要

1. 本スライドは，電子情報通信学会 情報理論研究会における 2017 年 7 月 13 日の講演に際し使用したスライドに講演者自身が修正を加えたものです.

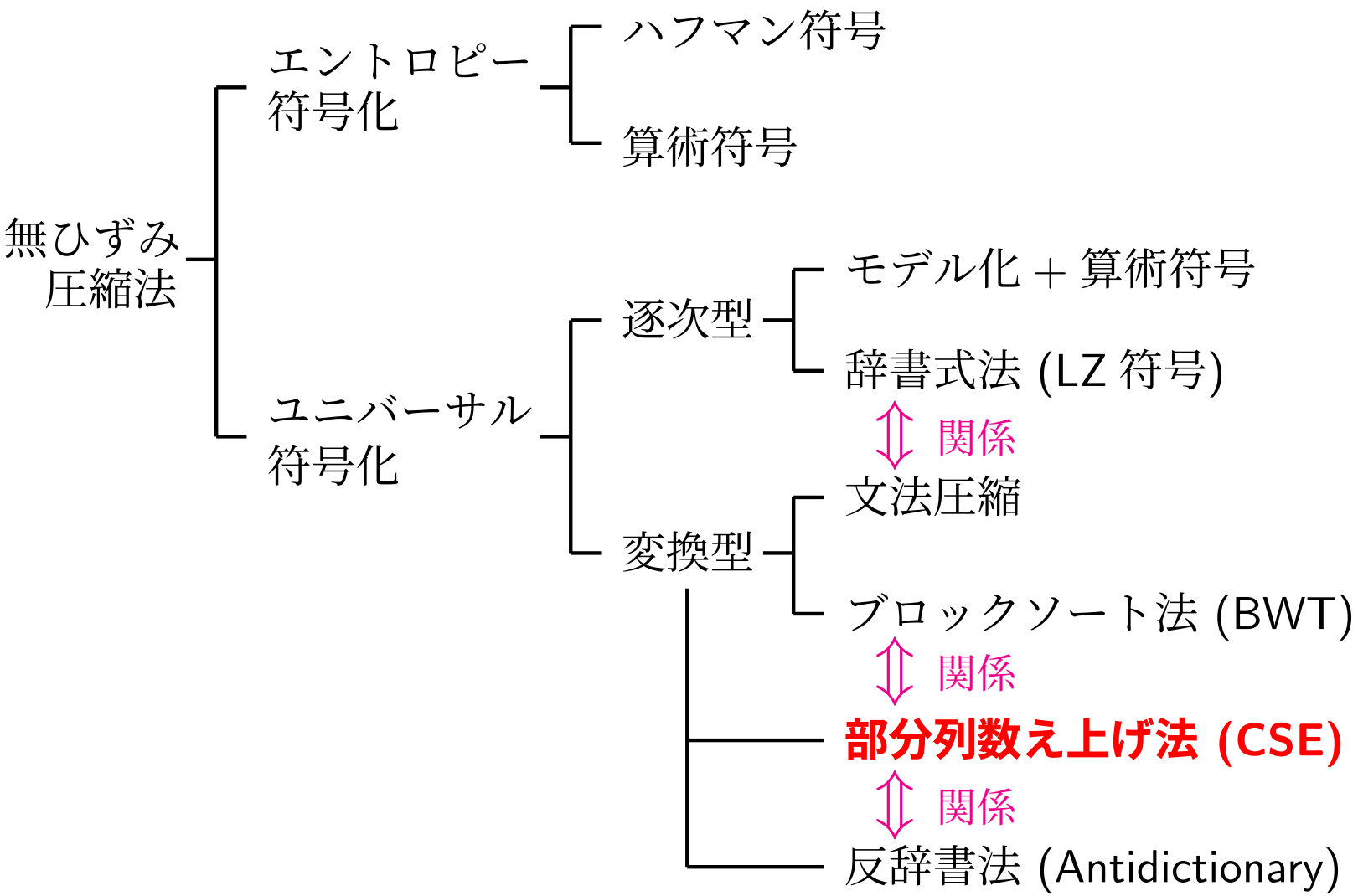
2. 以下の論文の内容を解説しています.

   横尾英俊，部分列数え上げデータ圧縮法とその関連法について，電子情報通信学会技術研究報告，Vol. 117, No. 120 (IT2017-26), pp. 55–60, 2017 年 7 月.

概要

1. 「部分列数え上げデータ圧縮法」とは
   - ➢ 符号化の基本的考え
   - ➢ 関連研究 (日本からの寄与が大きい)
2. 符号化モデルと漸近的最良性
   - ➢ 符号化モデルとは
   - ➢ 切り替えモデル
   - ➢ 漸近的最良性と冗長度
3. バリエーション
   - ➢ 多値化
   - ➢ 実装法
   - ➢ 他の手法との関連
4. むすび

無ひずみデータ圧縮法の体系



部分列数え上げデータ圧縮法

- D. Dubé and V. Beaudoin (2010) *Data Compression Conference*
- Compression {via/by} Substring Enumeration (CSE)
- ユニバーサルデータ圧縮法
- Block-based, オフライン
- 2 元データ対象: $\boldsymbol{x} \in \{0, 1\}^n$ (巡回列)
- 部分列ごとの出現回数を符号化

  (出現回数の例) $\boldsymbol{x} = 00101001$ における $w = 010$ の出現回数

$$\boldsymbol{x} = 0\overbrace{0\underbrace{1\underbrace{0}_{2}1\underbrace{0}_{2}}_{1}0}^{3}1 \qquad C_w(\boldsymbol{x})\colon\ C_{010}(00101001) = 3$$

- 任意次数のマルコフ情報源および定常エルゴード情報源に対し漸近最良

部分列数え上げデータ圧縮法

- 2 元データ対象: $\boldsymbol{x} \in \{0, 1\}^n$ (巡回列)
- 部分列ごとの出現回数を符号化

  $C_\lambda = C_\lambda(\boldsymbol{x}) = n \quad (\lambda : \text{空系列})$

  $C_0 = C_0(\boldsymbol{x}) =$ "$\boldsymbol{x}$ に含まれる 0 の個数"

  $C_{0w0} = C_{0w0}(\boldsymbol{x}) =$ "$\boldsymbol{x}$ に含まれる $0w0$ の個数" $(w \in \{0, 1\}^*)$

- **整合性条件**

  (巡回列を考えているので)

  $$\begin{aligned} C_w &= C_{w0} + C_{w1} \\ &= C_{0w} + C_{1w} \end{aligned}$$

- 例

  $\boldsymbol{x} = 00101001 \ (n = 8)$

  $C_0 = 5, \ C_{00} = 2, \ C_{000} = 0$

  これ以外は右の整合性条件から決まる：

$$\begin{aligned} C_{0w1} &= C_{0w} - C_{0w0} \\ C_{1w0} &= C_{w0} - C_{0w0} \\ C_{1w1} &= C_{w1} - C_{0w1} \end{aligned}$$

部分列数え上げデータ圧縮法

- 例

$\boldsymbol{x} = 00101001\ (n = 8)$

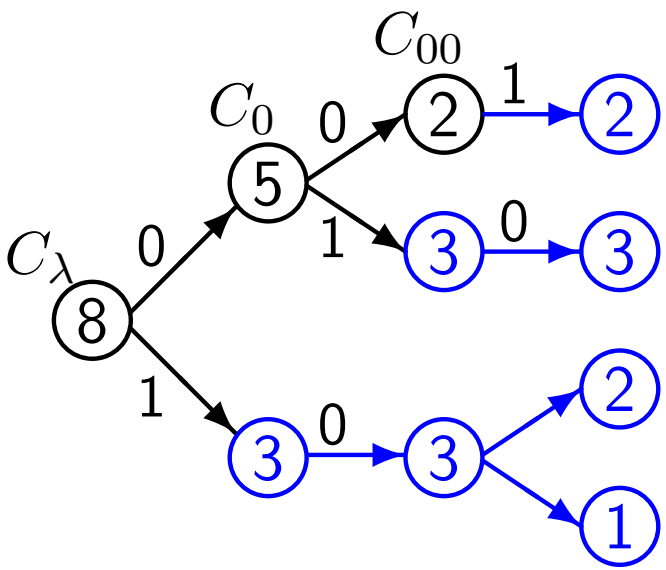
$C_0 = 5,\ C_{00} = 2,\ C_{000} = 0$

これ以外は右の整合性条件から決まる：

$$C_{0w1} = C_{0w} - C_{0w0} \quad (5)$$
$$C_{1w0} = C_{w0} - C_{0w0} \quad (6)$$
$$C_{1w1} = C_{w1} - C_{0w1} \quad (7)$$

さらに,

$$\max\{0,\ C_{0w} - C_{w1}\} \le C_{0w0} \le \min\{C_{0w},\ C_{w0}\} \quad (8)$$



部分列数え上げデータ圧縮法

- 例

$\boldsymbol{x} = 00101001\ (n = 8)$

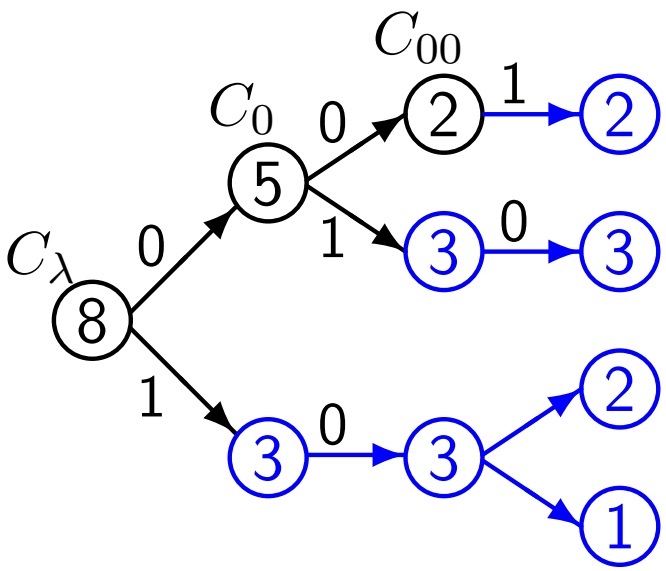
$C_0 = 5,\ C_{00} = 2,\ C_{000} = 0$

これ以外は右の整合性条件から決まる：

$$C_{0w1} = C_{0w} - C_{0w0} \quad (5)$$
$$C_{1w0} = C_{w0} - C_{0w0} \quad (6)$$
$$C_{1w1} = C_{w1} - C_{0w1} \quad (7)$$

さらに,

$$\max\{0,\ C_{0w} - C_{w1}\} \le C_{0w0} \le \min\{C_{0w},\ C_{w0}\} \quad (8)$$



$w = 1$: C_{010}?

$$\max\{0,\ 3\} \le C_{010} \le \min\{3,\ 3\}$$

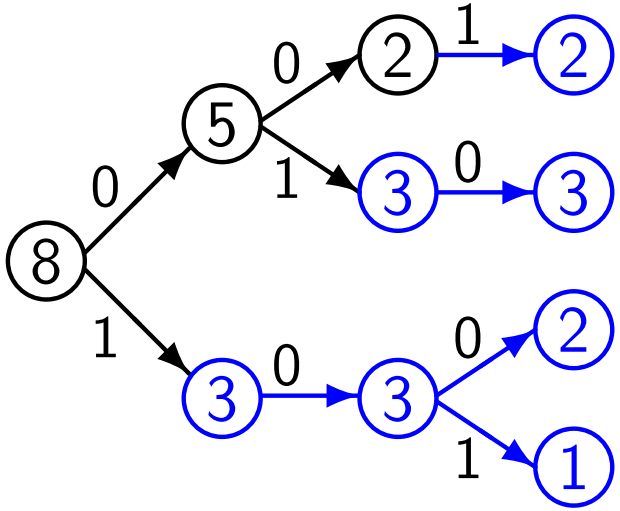
部分列数え上げデータ圧縮法

- 復号法
  - ➢ 生成規則の抽出:

    00 の後には 1 が続く

    01 の後には 0 が続く

    10 の後には (1 回の例外を除き) 0 が続く



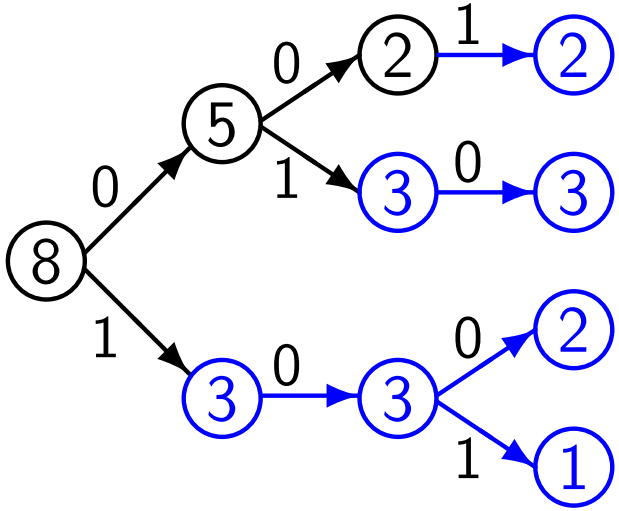
部分列数え上げデータ圧縮法

- 復号法
  - ➢ 生成規則の抽出:

    00 の後には 1 が続く

    01 の後には 0 が続く

    10 の後には (1 回の例外を除き) 0 が続く



(例外部分)

101

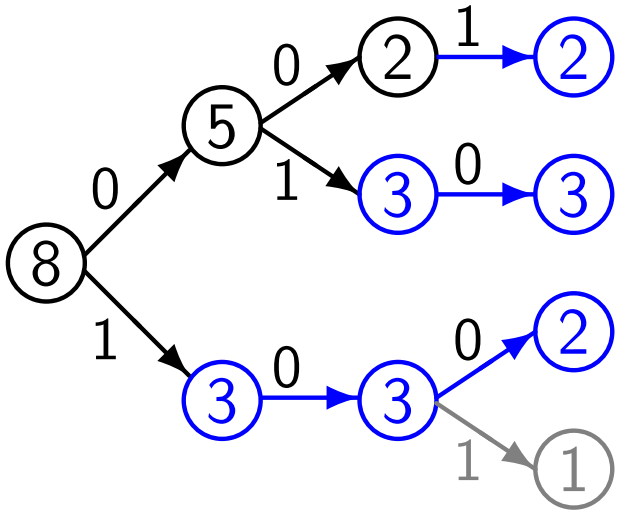
部分列数え上げデータ圧縮法

- 復号法
  - ➢ 生成規則の抽出:

    00 の後には 1 が続く

    01 の後には 0 が続く

    10 の後には (1 回の例外を除き) 0 が続く



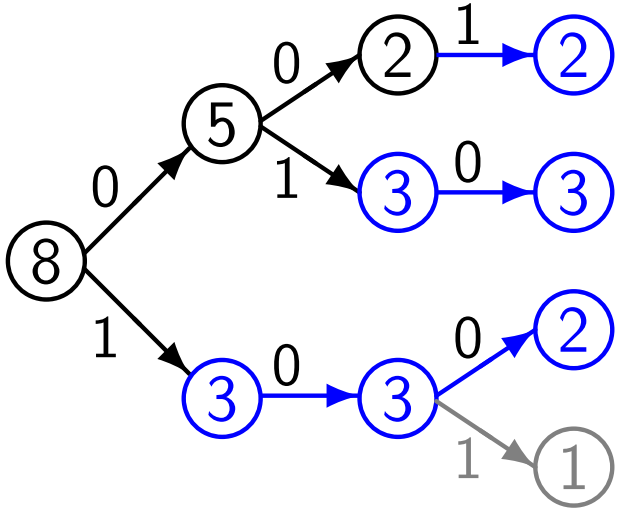
部分列数え上げデータ圧縮法

- 復号法
  - ➢ 生成規則の抽出:

    00 の後には 1 が続く

    01 の後には 0 が続く

    10 の後には (1 回の例外を除き) 0 が続く



10100

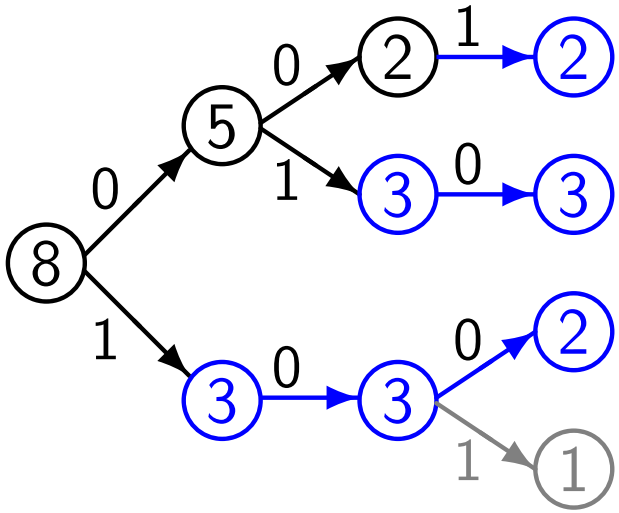
部分列数え上げデータ圧縮法

- 復号法
  - ➢ 生成規則の抽出:

    00 の後には 1 が続く

    01 の後には 0 が続く

    10 の後には (1 回の例外を除き) 0 が続く



101001

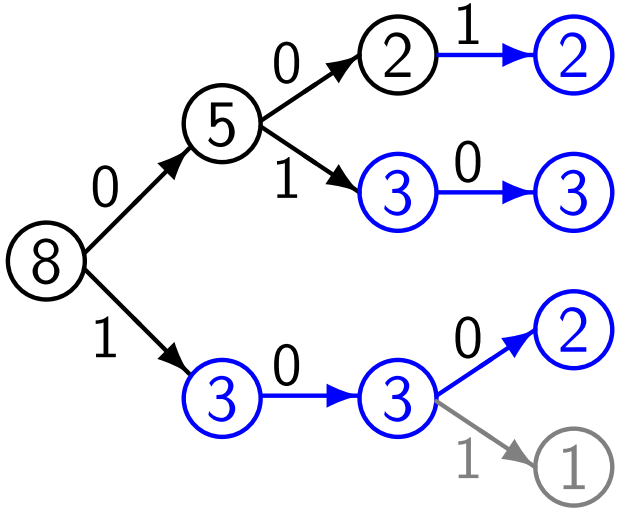
部分列数え上げデータ圧縮法

- 復号法
  - ➢ 生成規則の抽出:

    00 の後には 1 が続く

    01 の後には 0 が続く

    10 の後には (1 回の例外を除き) 0 が続く



1010010

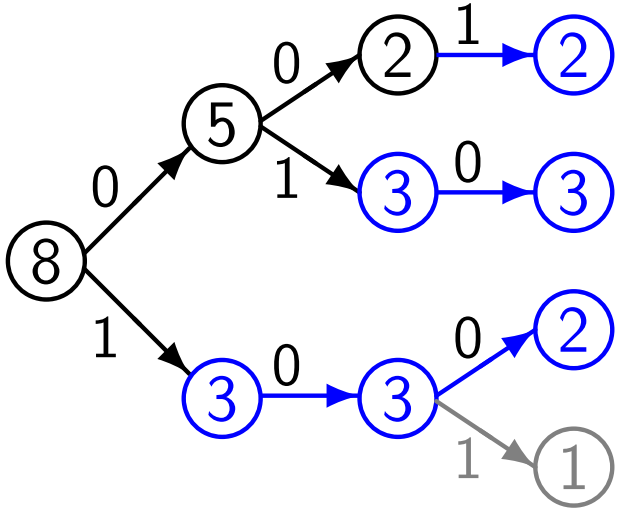
部分列数え上げデータ圧縮法

- 復号法
  - ➢ 生成規則の抽出:

    00 の後には 1 が続く

    01 の後には 0 が続く

    10 の後には (1 回の例外を除き) 0 が続く



$$
\begin{array}{c}
10100100 \\
+ \\
\mathrm{rank}(\boldsymbol{x}) \\
\| \\
\boldsymbol{x} = 00101001
\end{array}
$$

部分列数え上げデータ圧縮法

- 符号化法まとめ

CSE 符号化法 ($\boldsymbol{x} \in \{0,1\}^n$ の圧縮 with n: given)

1. **Encode** C_0;
2. **for** $l := 0$ **to** $n-2$ **do**
   for every $w \in \mathcal{X}^l \cap I(\boldsymbol{x})$ **do**
   Encode C_{0w0} in (8);
3. **Encode** rank($\boldsymbol{x}$);

より短い部分列の出現回数を利用して C_{0w0} の値を符号化

$$\max\{0,\, C_{0w} - C_{w1}\} \le C_{0w0} \le \min\{C_{0w},\, C_{w0}\} \qquad (8)$$

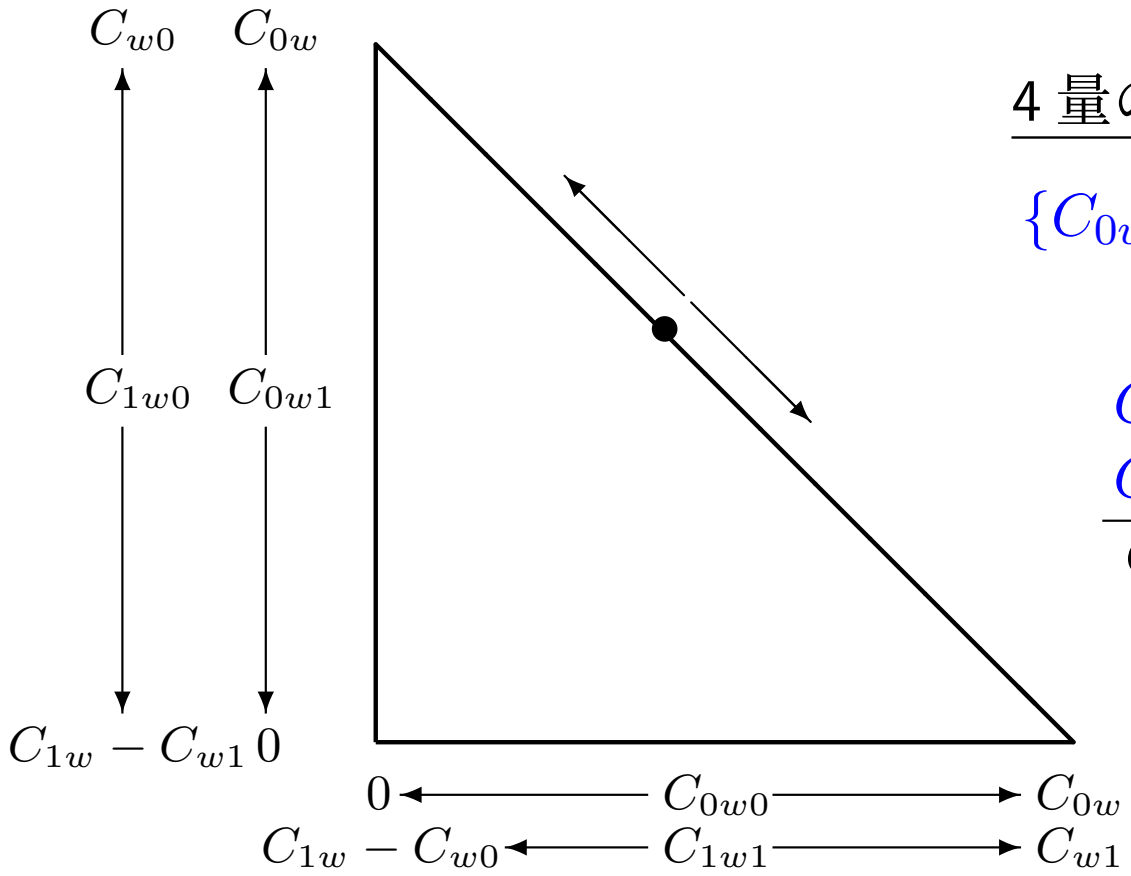
$$\text{上限} - \text{下限} = \min\{C_{0w}, C_{1w}, C_{w0}, C_{w1}\}$$

$$I(\boldsymbol{x}) = \{w \in \{0,1\}^* \mid \min\{C_{0w}, C_{1w}, C_{w0}, C_{w1}\} \ge 1\}$$

部分列数え上げデータ圧縮法

$$\max\{0,\, C_{0w} - C_{w1}\} \le C_{0w0} \le \min\{C_{0w},\, C_{w0}\} \qquad (8)$$

上限 $-$ 下限 $= \min\{C_{0w}, C_{1w}, C_{w0}, C_{w1}\} = C_{0w}$ の場合



4 量の自由度は 1

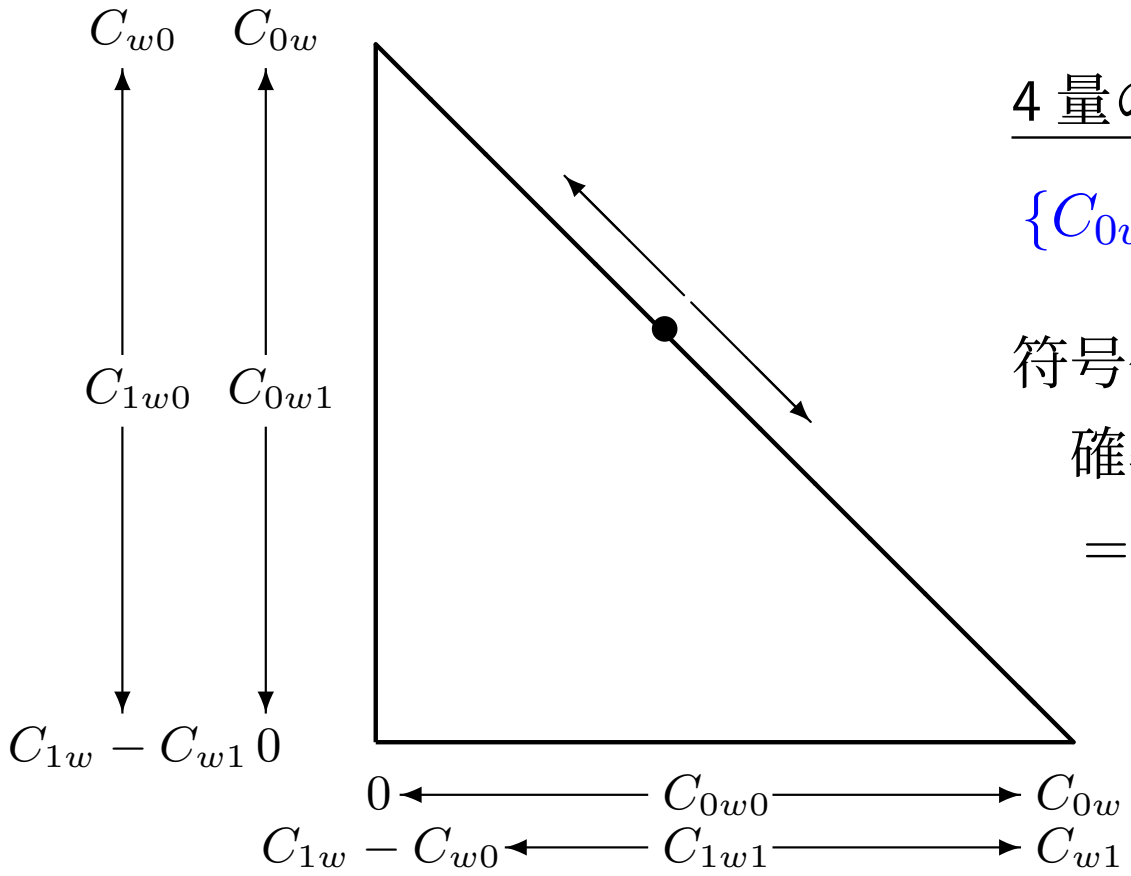
$\{C_{0w0},\, C_{0w1},\, C_{1w0},\, C_{1w1}\}$

C_{0w0}	C_{0w1}	C_{0w}
C_{1w0}	C_{1w1}	C_{1w}
C_{w0}	C_{w1}	C_w

部分列数え上げデータ圧縮法

$$\max\{0,\, C_{0w} - C_{w1}\} \le C_{0w0} \le \min\{C_{0w},\, C_{w0}\} \qquad (8)$$

上限 $-$ 下限 $= \min\{C_{0w}, C_{1w}, C_{w0}, C_{w1}\} = C_{0w}$ の場合



4 量の自由度は 1

$\{C_{0w0},\, C_{0w1},\, C_{1w0},\, C_{1w1}\}$

符号化では，この範囲に
確率分布の仮定が必要
$=$ 符号化モデル

関連研究 (例)

- 改良と最初の理論解析
  嶋–岩田–有村 (IT 研 2011, ICIS 2011)
- 定常エルゴード情報源に対する漸近的最良性
  横尾 (CCP 2011)
- 超幾何分布の導入とマルコフ情報源に対する漸近的最良性
  Dubé–横尾 (ISIT 2011)
- (個別) 冗長度解析
  岩田–有村–嶋 (ISITA 2012, IEICE Trans. Fundumentals 2014)
- 反辞書法との関係
  太田–森田 (ISIT 2013, ISITA 2014)
- 多値化
  太田–森田 (ISITA 2014)
  岩田–有村 (IEICE Trans. Fundumentals 2016)
- 実装法
  金井–横尾–山崎–金安 (IEICE Trans. Fundamentals 2016)

符号化モデルと漸近的最良性

1. 「部分列数え上げデータ圧縮法」とは
   - ➢ 符号化の基本的考え
   - ➢ 関連研究 (日本からの寄与が大きい)
2. 符号化モデルと漸近的最良性
   - ➢ 符号化モデルとは
   - ➢ 切り替えモデル
   - ➢ 漸近的最良性と冗長度
3. バリエーション
   - ➢ 多値化
   - ➢ 実装法
   - ➢ 他の手法との関連
4. むすび

符号化モデルと漸近的最良性

$$\max\{0,\, C_{0w} - C_{w1}\} \le C_{0w0} \le \min\{C_{0w},\, C_{w0}\} \tag{8}$$

- 符号化では，この範囲に確率分布の仮定が必要
  = 符号化モデル
- 一様分布モデル
- 組合せモデル (超幾何分布モデル)

$$p_c(C_{0w0}) = p_c(C_{0w0} \mid C_{0w}, C_{w0}, C_{w1}) = \frac{\binom{C_{0w}}{C_{0w0}}\binom{C_{1w}}{C_{1w0}}}{\binom{C_w}{C_{w0}}} \tag{10}$$

- 切り替えモデル (整数パラメータ k に対し)
  - $|w| < k$ ならば，C_{0w0} を一様分布モデルで符号化;
  - $|w| \ge k$ ならば，C_{0w0} を組合せモデルで符号化

符号化モデルと漸近的最良性

- 組合せモデル (超幾何分布モデル)

$$
\left.\begin{array}{l}
C_{0w}\left\{\begin{array}{lll} 0 & \boxed{w} & b_1 \\ 0 & \boxed{w} & b_2 \\ \vdots & \vdots & \vdots \\ 0 & \boxed{w} & b_{C_{0w}} \end{array}\right. \quad \begin{array}{l} C_{0w0} \text{ 個の} \\ \\ b_i = 0 \end{array} \\
C_{1w}\left\{\begin{array}{lll} 1 & \boxed{w} & b_{C_{0w}+1} \\ \vdots & \vdots & \vdots \\ 1 & \boxed{w} & b_{C_w} \end{array}\right. \quad \begin{array}{l} C_{1w0} \text{ 個の} \\ \\ b_i = 0 \end{array}
\end{array}\right\} \begin{array}{l} C_{w0} \text{ 個の} \\ \\ b_i = 0 \end{array}
$$

$$
p_c(C_{0w0} \mid C_{0w}, C_{w0}, C_{w1}) = \frac{\binom{C_{0w}}{C_{0w0}}\binom{C_{1w}}{C_{1w0}}}{\binom{C_w}{C_{w0}}} \tag{10}
$$

符号化モデルと漸近的最良性

- 組合せモデル (超幾何分布モデル)

$$p_c(C_{0w0}) = \frac{\binom{C_{0w}}{C_{0w0}}\binom{C_{1w}}{C_{1w0}}}{\binom{C_w}{C_{w0}}}. \tag{10}$$

$w = v,\ 0v,\ 1v,\ 00v,\ 01v, \ldots$

$$\begin{aligned}
&-\log p_c(C_{0v0}) - \log p_c(C_{00v0}) - \log p_c(C_{01v0}) - \cdots \\
&= \log\binom{C_v}{C_{v0}} - \log\binom{C_{0v}}{C_{0v0}} - \log\binom{C_{1v}}{C_{1v0}} \\
&\quad + \log\binom{C_{0v}}{C_{0v0}} - \log\binom{C_{00v}}{C_{00v0}} - \log\binom{C_{10v}}{C_{10v0}} - \cdots \\
&= \log\binom{C_v}{C_{v0}} \qquad \text{(望遠鏡級数)}
\end{aligned}$$

(参考) Wikipedia (ドイツ語)
Teleskopsumme

符号化モデルと漸近的最良性

$$\max\{0,\, C_{0w} - C_{w1}\} \le C_{0w0} \le \min\{C_{0w},\, C_{w0}\} \tag{8}$$

- 切り替えモデル (整数パラメータ k に対し)

• $\lvert w\rvert < k$ ならば，C_{0w0} を一様分布モデルで符号化;
• $\lvert w\rvert \ge k$ ならば，C_{0w0} を組合せモデルで符号化

- 符号語長 (固定した k に対し)
  - 第 1 成分 $=$ 定数 $\times \log n$
  - 第 2 成分 $\to nH(X_k \mid X_0^{k-1})$，なぜなら

$$\log \binom{C_v}{C_{v0}} \to C_v \mathscr{H}\left(\frac{C_{v0}}{C_v}\right) \qquad (C_v \to \infty)$$

ここで，$\mathscr{H}(p) = -p\log p - (1-p)\log(1-p)$

(2 元エントロピー関数)

漸近的最良性

- 切り替えモデル (整数パラメータ k に対し)

• $\lvert w\rvert < k$ ならば，C_{0w0} を一様分布モデルで符号化;
• $\lvert w\rvert \geq k$ ならば，C_{0w0} を組合せモデルで符号化

- 任意の 2 元 m 次マルコフ情報源に対し，
  $k \geq m$ と選んで $n \to \infty$ とするとき，圧縮比は情報源のエントロピーレートに漸近する.
- 任意の 2 元定常エルゴード情報源に対し，
  $k = \log\log n$ と選んで $n \to \infty$ とするとき，圧縮比は情報源のエントロピーレートに漸近する.
- 多値情報源 ($\mathcal{X}$: 有限アルファベット) に拡張可能

符号化モデルと漸近的最良性

最大冗長度 (Iwata–Arimura–Shima 2012–2014)

- $R_n^{(m)}(\varphi_n)$: 一意復号可能な符号 $\varphi_n : \{0,1\}^n \to \{0,1\}^*$ の m 次マルコフ情報源 $\{P_{X_1^n}\}$ に対する最大冗長度

$$R_n^{(m)}(\varphi_n) = \sup_{P_{X_1^n}} \max_{\boldsymbol{x}\in\{0,1\}^n} \{L(\varphi_n(\boldsymbol{x})) + \log P_{X_1^n}(\boldsymbol{x})\}$$

$$R_n^{(m)} = \min_{\varphi_n} \sup_{P_{X_1^n}} \max_{\boldsymbol{x}\in\{0,1\}^n} \{L(\varphi_n(\boldsymbol{x})) + \log P_{X_1^n}(\boldsymbol{x})\}$$

- 定理 (一般の有限アルファベット $\mathcal{X}$ の場合)
  - 切り替えモデルのパラメータを $k = m$ にとると,

$$R_n^{(m)}(\mathrm{CSE}_n) \le R_n^{(m)} + \left(\frac{|\mathcal{X}|^m(|\mathcal{X}|-1)}{2} + 1\right)\log n + O(1)$$

$$R_n^{(m)} = \frac{|\mathcal{X}|^m(|\mathcal{X}|-1)}{2}\log n + O(1) \qquad \text{(Krichevsky–Trofimov'81, CS'04)}$$

バリエーション

1. 「部分列数え上げデータ圧縮法」とは
   - ➢ 符号化の基本的考え
   - ➢ 関連研究 (日本からの寄与が大きい)
2. 符号化モデルと漸近的最良性
   - ➢ 符号化モデルとは
   - ➢ 切り替えモデル
   - ➢ 漸近的最良性と冗長度
3. **バリエーション**
   - ➢ 多値化
   - ➢ 実装法
   - ➢ 他の手法との関連
4. むすび

バリエーション

CSE 法の従来研究の基本的枠組 (理論–実現)

情報源アルファベット	2 元 $\{0,1\}$	多値 $\{0,1,\ldots,J-1\}$
符号化モデル	一様分布	
	超幾何分布	タイプの大きさの比
切り替えモデル	固定次数切り替え	可変次数切り替え
漸近的最良性	m 次マルコフ情報源	定常エルゴード情報源
実装法	Compacted Substring Tree (CST)	BWT 行列
アルファベット拡大	k-フェーズ法	多値化

多値化 (Ota–Morita 2014, Iwata–Arimura 2016)

$\mathcal{X} = \{0, 1\}$

C_{0w0}	C_{0w1}	C_{0w}
C_{1w0}	C_{1w1}	C_{1w}
C_{w0}	C_{w1}	C_{w}

$\mathcal{X} = \{0, 1, 2, \ldots, J'\}$

C_{0w0}	C_{0w1}	C_{0w2}	$\cdots$	$C_{0wJ'}$	C_{0w}
C_{1w0}	C_{1w1}	C_{1w2}	$\cdots$	$C_{1wJ'}$	C_{1w}
C_{2w0}	C_{2w1}	C_{2w2}	$\cdots$	$C_{2wJ'}$	C_{2w}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$C_{J'w0}$	$C_{J'w1}$	$C_{J'w2}$	$\cdots$	$C_{J'wJ'}$	$C_{J'w}$
C_{w0}	C_{w1}	C_{w2}	$\cdots$	$C_{wJ'}$	C_{w}

多値化 (Ota–Morita 2014, Iwata–Arimura 2016)

1. $\mathcal{P} := \emptyset$;
2. **for** $a \in \mathcal{X}$ **do**
   $\mathcal{S} := \emptyset$;
   for $b \in \mathcal{X}$ **do**
   Encode C_{awb};
   $\mathcal{S} := \mathcal{S} \cup \{b\}$;
   $\mathcal{P} := \mathcal{P} \cup \{a\}$;

$$\begin{array}{ccccc|c}
C_{0w0} & C_{0w1} & C_{0w2} & \cdots & C_{0wJ'} & C_{0w} \\
C_{1w0} & C_{1w1} & C_{1w2} & \cdots & C_{1wJ'} & C_{1w} \\
C_{2w0} & C_{2w1} & C_{2w2} & \cdots & C_{2wJ'} & C_{2w} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
C_{J'w0} & C_{J'w1} & C_{J'w2} & \cdots & C_{J'wJ'} & C_{J'w} \\
\hline
C_{w0} & C_{w1} & C_{w2} & \cdots & C_{wJ'} & C_{w}
\end{array}$$

$$\begin{aligned}
&\max\Big\{0,\ C_{aw} - \sum_{c\in\mathcal{S}} C_{awc} - \sum_{c\notin\mathcal{S}\cup\{b\}} C_{wc}, \\
&\qquad C_{wb} - \sum_{c\in\mathcal{P}} C_{cwb} - \sum_{c\notin\mathcal{P}\cup\{a\}} C_{cw}\Big\} \\
&\le C_{awb} \le \min\Big\{C_{aw} - \sum_{c\in\mathcal{S}} C_{awc},\ C_{wb} - \sum_{c\in\mathcal{P}} C_{cwb}\Big\}
\end{aligned}$$

多値化 (組合せモデル Iwata–Arimura 2016)

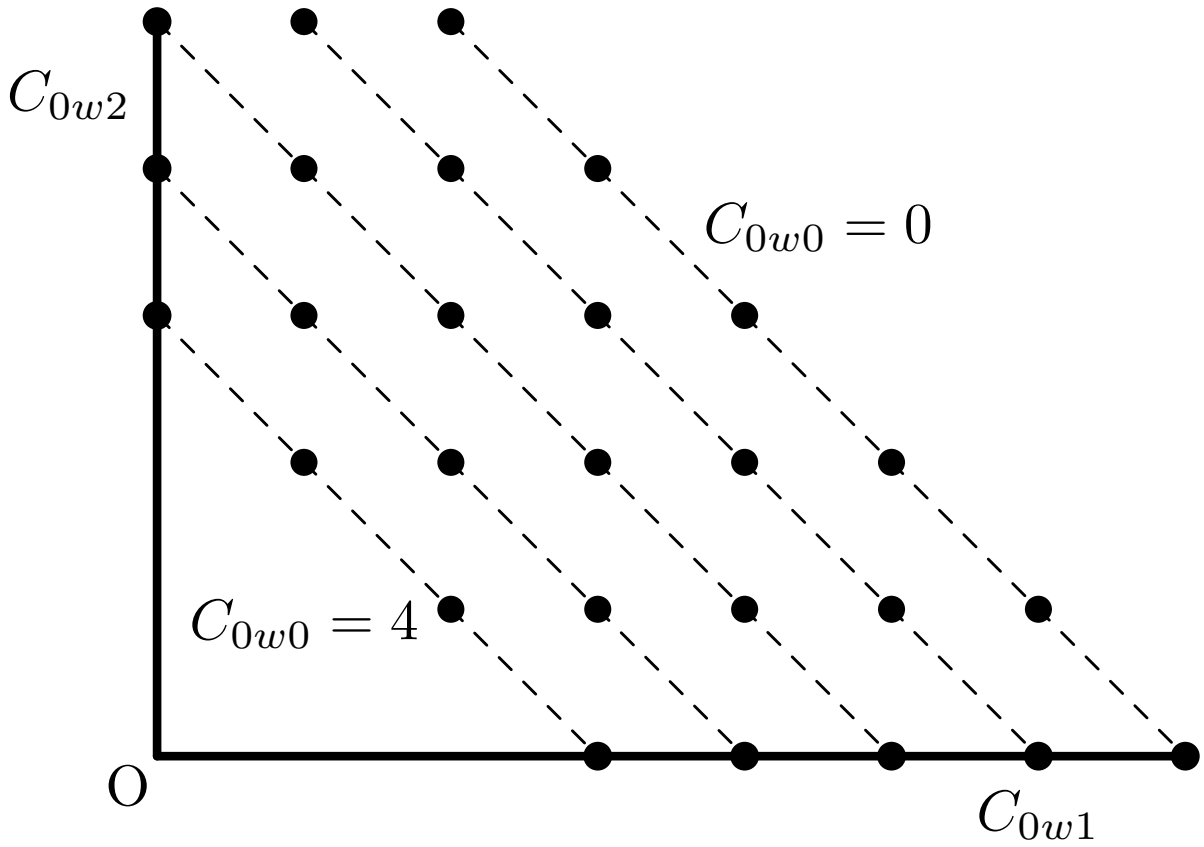
- 二項係数を多項係数に一般化

$$P_c(\{C_{awb} : a, b \in \mathcal{X}\})$$
$$= \frac{\binom{C_{0w}}{C_{0w0}, \ldots, C_{0wJ'}} \cdots \binom{C_{J'w}}{C_{J'w0}, \ldots, C_{J'wJ'}}}{\binom{C_w}{C_{w0}, \ldots, C_{wJ'}}}$$

- (理論的性能) 2 元の場合とほぼパラレルな議論が可能
- 注意 (2 元の場合には区別不要だった概念):
  - 出現回数 C_{0w0} の確率分布 $\to$ 集合 $\{C_{awb} : a, b \in \mathcal{X}\}$ 上の確率分布
  - 一様分布の場合，「C_{awb} の可能な値の上の一様分布の集合」と「集合 $\{C_{awb} : a, b \in \mathcal{X}\}$ 上の一様分布」は，一般には，一致しない.

多値化

- 両者が一致しないことのイメージ：
  - C_{awb} の可能な値の上の一様分布の集合
  - 集合 $\{C_{awb} : a, b \in \{0, 1, 2\}\}$ 上の一様分布



実装法

- 理論的な性能を実際に検証してみたい

  ↓

  実装が必要

  必要な $\{C_{0w0}\}$ を効率的に数え上げる方法?

- 最初の提案 (DV '10) では，アルゴリズムと実装とが混然としていた
  - Compacted Substring Tree (CST)
  - しかし，提案者ですら実際には利用せず
  - より原始的な方法で実性能を検証 (確かに，高圧縮性能)
- CST を実装しつつ，実性能向上のための改良の試み?
- (金井–横尾 '14) Burrows–Wheeler 変換 (BWT) の利用
- CST の意義—後述
  - 反辞書法の漸近的最良性 (Ota–Morita '13, '14)

Burrows–Wheeler 変換 (BWT) の利用

- $\max\{0,\, C_{0w} - C_{w1}\} \le C_{0w0} \le \min\{C_{0w},\, C_{w0}\}$ (8)

  C_{0w0}, C_{0w}, C_{w1}, C_{w0} が同時に必要

- (例) $\boldsymbol{x} = 00001101$

  $w[i]$ = 第 i 行とすぐ上の行の共通語頭 $(i = 0, 1, \ldots, n-1)$
  = $S[i]$ 行から $E[i]$ 行の長さ $\mathrm{LCP}[i]$ の共通する語頭

右以外の w が式(8)の候補になることはない

i	w	LCP	S	E
0	–	−1	–	–
1	000	3	0	1
2	00	2	0	2
3	0	1	0	4
4	01	2	3	4
5	λ	0	0	7
6	10	2	5	6
7	1	1	5	7

BWT 行列							BWT($\boldsymbol{x}$)	R
0	0	0	0	1	1	0	**1**	0
0	0	0	1	1	0	1	**0**	1
0	0	1	1	0	1	0	**0**	2
0	1	0	0	0	0	1	**1**	2
0	1	1	0	1	0	0	**0**	3
1	0	0	0	0	1	1	**0**	4
1	0	1	0	0	0	0	**1**	4
1	1	0	1	0	0	0	**0**	5

Burrows–Wheeler 変換 (BWT) の利用

i	w	LCP	S	E
0	–	−1	–	–
1	000	3	0	1
2	00	2	0	2
3	0	1	0	4
4	01	2	3	4
5	λ	0	0	7
6	10	2	5	6
7	1	1	5	7

BWT 行列

0	0	0	0	1	1	0	**1**
0	0	0	1	1	0	1	**0**
0	0	1	1	0	1	0	**0**
0	1	0	0	0	0	1	**1**
0	1	1	0	1	0	0	**0**
1	0	0	0	0	1	1	**0**
1	0	1	0	0	0	0	**1**
1	1	0	1	0	0	0	**0**

BWT($\boldsymbol{x}$)

R
0
1
2
2
3
4
4
5

BWT($\boldsymbol{x}$) での 0 の累積個数

$w = w[i]$ に対して,

$$
\begin{aligned}
C_w &= E[i] - S[i] + 1 \\
C_{w0} &= i - S[i] \\
C_{0w} &= R[E[i]] - R[S[i] - 1] \\
C_{0w0} &= R[i - 1] - R[S[i] - 1]
\end{aligned}
$$

Burrows–Wheeler 変換 (BWT) の利用

- 高速かつ省メモリの実装が実現 (長大な入力にも対応可能)
- 圧縮性能例 (bits per byte)
  (2 元，一様分布のみ)

File [KB]	n [bit]	gzip-b	bzip2-9	ppmD5	CSE
alice29 [149]	1,216,712	2.85	2.27	2.20	2.24
asyoulik [122]	1,001,432	3.12	2.53	2.49	2.56
lcet10 [417]	3,414,032	2.71	2.02	1.95	1.99
plrabn12 [471]	3,854,888	3.23	2.42	2.36	2.41
E.coli [4530]	37,109,520	2.24	2.16	1.99	2.31
bible [3953]	32,379,136	2.33	1.67	1.58	1.53
world [2415]	19,787,200	2.33	1.58	1.52	1.33

- マルコフ情報源でのシミュレーション
  切り替えモデルの理論予測どおりの結果

k-フェーズ CSE 法 (多値データへの対応)

- 多値版では実性能は向上しない
- k-フェーズ CSE 法 (Béliveau–Dubé '14)
  - 2 値データのままでの処理
  - ただし，フェーズを導入する
  - ($k = 4$ の例) 下つき添字がフェーズ p

$$\boldsymbol{x} = 0_0 0_1 0_2 1_3 1_0 1_1 0_2 1_3 0_0 1_1 0_2 1_3 0_0 0_1 1_2 1_3$$

- C_w^p: フェーズ p を開始位置とする w の出現回数 ($\sum_p C_w^p = C_w$)

  整合性条件 ($p \oplus \delta = (p + \delta) \bmod k$)

$$\begin{aligned} & C_{0w}^p + C_{1w}^p \\ = & C_{w0}^{p\oplus 1} + C_{w1}^{p\oplus 1} = C_w^{p\oplus 1} \end{aligned}$$

- $\max\{0,\, C_{0w}^p - C_{w1}^{p\oplus 1}\} \le C_{0w0}^p \le \min\{C_{0w}^p,\, C_{w0}^{p\oplus 1}\}$

k-フェーズ CSE 法 (多値データへの対応)

- フェーズごとの BWT 行列 (佐久間–成澤–篠原 '15)
- (例) $\boldsymbol{x} = 001\,101\,010,\ \ k = 3$

フェーズ 0

i	w	LCP	S	E
0	–	-1	–	–
1	0	1	0	1
2	λ	0	0	2

BWT 行列

									R
0	0	1	1	0	1	0	1	0	1
0	1	0	0	0	1	1	0	1	1
1	0	1	0	1	0	0	0	1	1

フェーズ 1

i	w	LCP	S	E
0	–	-1	–	–
1	01	2	0	1
2	λ	0	0	2

									R
0	1	0	1	0	0	0	1	1	0
0	1	1	0	1	0	1	0	0	1
1	0	0	0	1	1	0	1	0	2

フェーズ 2

i	w	LCP	S	E
0	–	-1	–	–
1	λ	0	0	2
2	1	1	1	2

									R
0	0	0	1	1	0	1	0	1	0
1	0	1	0	0	0	1	1	0	1
1	1	0	1	0	1	0	0	0	2

k-フェーズ CSE 法 (多値データへの対応)

- フェーズごとの BWT 行列 (佐久間–成澤–篠原 '15)
- 圧縮性能例 (bits per byte)
  (一様分布のみ)

		フェーズ				改善率
File [KB]	n [bit]	$k=1$	$k=2$	$k=4$	$k=8$	$k=8$ vs. 1
bib [109]	890,088	1.97	1.92	1.90	1.87	5.1%
book1 [751]	6,150,168	2.41	2.40	2.39	2.39	0.8%
book2 [597]	4,886,848	2.03	2.01	2.00	1.99	2.0%
geo [100]	819,200	5.80	5.50	5.28	5.03	13.3%
news [368]	3,016,872	2.55	2.49	2.47	2.45	3.9%
obj1 [21]	172,032	5.09	4.75	4.44	4.15	18.5%
obj2 [241]	1,974,512	2.71	2.62	2.53	2.44	10.0%
progc [39]	316,888	2.59	2.50	2.45	2.40	7.3%
progp [48]	395,032	1.79	1.72	1.68	1.65	7.8%
trans [91]	749,560	1.68	1.60	1.55	1.50	10.7%

他の手法との関連

- 完全に CSE 法に帰着可能な符号化法が存在する例：
- 反辞書法
  - DCA (Data Compression Using Antidictionaries)
    M. Crochemore, F. Mignosi, A. Restivo, S. Salemi (2000), *Proc. IEEE*
  - CSE 法に帰着した漸近的最良性
    T. Ota and H. Morita (2013), *ISIT 2013*, Istanbul
    T. Ota and H. Morita (2014), *ISITA 2014*, Melbourne
  - 反辞書と CST (Compacted Substring Tree) の同型性に着目
- ブロックソート法
  - M. Burrows and D. Wheeler (1994), *SRC Research Report*
    BW 変換 (BWT)
  - BWT+ Second-Step Encoder
    Existing Second-Step Encoders: Move-To-Front, Run-Length, etc.
  - BWT+Second-Step Encoder = CSE 法
    となる Second-Step Encoder が存在

反辞書法

- 反辞書
  - 最小禁止語 (MFW: Minimum Forbidden Word) の集合
  - 最小禁止語
    awb が $\boldsymbol{x}$ の最小禁止語 $(a, b \in \mathcal{X},\ w \in \mathcal{X}^*,\ \boldsymbol{x} \in \mathcal{X}^n)$
    $\Leftrightarrow C_{awb}(\boldsymbol{x}) = 0,\ C_{aw}(\boldsymbol{x}) > 0,\ C_{wb}(\boldsymbol{x}) > 0$
    (例)
  - $\boldsymbol{x} = 002101$ over $\mathcal{X} = \{0, 1, 2\}$

    $$\boldsymbol{x} \text{ の反辞書} = \{11, 12, 20, 22, 000, 001, 102, 0101, 2100\}$$

    $\boldsymbol{x} = 00 \cdots \rightarrow$

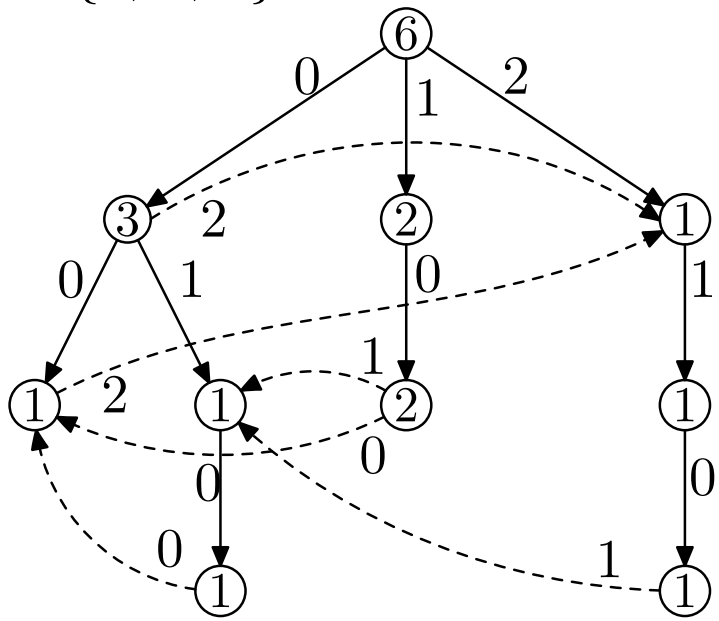
- 反辞書をいかにして符号化するか?
  - T. Ota and H. Morita (2013), *ISIT 2013*, Istanbul
    T. Ota and H. Morita (2014), *ISITA 2014*, Melbourne

    CSE 符号化法の CST (Compacted Substring Tree) として符号化

反辞書法

- CST (Compacted Substring Tree)
  - 巡回列に対するトライ (Trie)
  - 頂点の重み $= C_w$

    $C_{awb} = C_{wb}$ のとき，道 awb の b に対応する辺が後退辺になる．
  - $\boldsymbol{x} = 002101$ over $\mathcal{X} = \{0, 1, 2\}$



  - 最小禁止語 (awb) の同定

    道 aw はある，道 awb は**ない**(= 出次数 $|\mathcal{X}|$ 未満の頂点), 道 wb はある．

ブロックソート法の CSE 法への帰着

- CSE 法
  - **整合性条件**

    (巡回列を考えているので)

    $$\begin{aligned} C_w &= C_{w0} + C_{w1} \\ &= C_{0w} + C_{1w} \end{aligned}$$

- ところが，「巡回性」と「整合性条件」を切り離すことができる．
  $\boldsymbol{x} \in \mathcal{X}^n$, $a \in \mathcal{X}$, $v \in \mathcal{X}^*$ に対し，

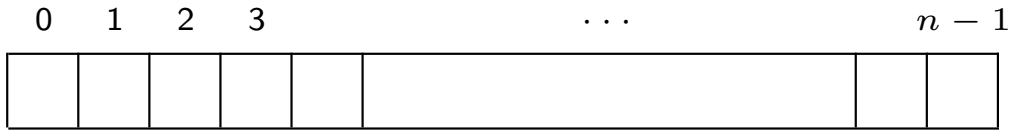
  $$A_\lambda(\boldsymbol{x}) = n,$$
  $$A_{av}(\boldsymbol{x}) = \text{“}\boldsymbol{x}\Big[\sum_{u \prec v} A_u(\boldsymbol{x}),\ \sum_{u \preceq v} A_u(\boldsymbol{x}) - 1\Big] \text{ に含まれる } a \text{ の個数”}$$

  を定義すると，次の “整合性条件” が成立：

  $$\sum_{c \in \mathcal{X}} A_{cw}(\boldsymbol{x}) = \sum_{c \in \mathcal{X}} A_{wc}(\boldsymbol{x}) = A_w(\boldsymbol{x})$$

ブロックソート法の CSE 法への帰着

$$\boldsymbol{x} = \boldsymbol{x}[0, n-1] =$$



- $A_{av}(\boldsymbol{x}) =$ "$\boldsymbol{x}\big[\sum_{u \prec v} A_u(\boldsymbol{x}),\ \sum_{u \preceq v} A_u(\boldsymbol{x}) - 1\big]$ に含まれる a の個数"

$A_{awb}(\boldsymbol{x}) =$ "wb で決まる区間に含まれる a の個数"

$$\begin{aligned}\sum_{a \in \mathcal{X}} A_{awb}(\boldsymbol{x}) &= \text{“}wb\text{ で決まる区間に含まれる記号の総数”} \\ &= \text{“}wb\text{ で決まる区間の長さ”} = A_{wb}\end{aligned}$$

$$\begin{aligned}\sum_{b \in \mathcal{X}} A_{awb}(\boldsymbol{x}) &= \bigcup_{b \in \mathcal{X}} \text{“}wb\text{ で決まる区間” に含まれる } a\text{ の総数} \\ &= A_{aw}\end{aligned}$$

ブロックソート法の CSE 法への帰着

- $\boldsymbol{x}' = \mathsf{BWT}(\boldsymbol{x})$

$$\boldsymbol{x} = \begin{array}{|cccccc|c|} \hline 0 & 0 & 2 & 1 & 0 & 1 & \\ \hline 0 & 1 & 0 & 0 & 2 & 1 & \\ 0 & 2 & 1 & 0 & 1 & 0 & \\ 1 & 0 & 0 & 2 & 1 & 0 & \\ 1 & 0 & 1 & 0 & 0 & 2 & \\ 2 & 1 & 0 & 1 & 0 & 0 & \\ \hline \end{array} \quad \begin{array}{c} \boldsymbol{x}' \\ \| \\ \end{array}$$

- $\boxed{A_w(\boldsymbol{x}') = C_w(\boldsymbol{x})}$

  w の長さについての帰納法で証明可能

$$A_\lambda(\boldsymbol{x}') = |\boldsymbol{x}'| = 6$$
$$A_0(\boldsymbol{x}') = 3$$
$$A_1(\boldsymbol{x}') = 2$$
$$A_2(\boldsymbol{x}') = 1$$
$$A_{00}(\boldsymbol{x}') = \text{“}\boldsymbol{x}'[0,\,2] \text{ での } 0 \text{ の個数”} = 1$$
$$A_{10}(\boldsymbol{x}') = \text{“}\boldsymbol{x}'[0,\,2] \text{ での } 1 \text{ の個数”} = 2$$
$$A_{20}(\boldsymbol{x}') = \text{“}\boldsymbol{x}'[0,\,2] \text{ での } 2 \text{ の個数”} = 0$$
$$A_{01}(\boldsymbol{x}') = \text{“}\boldsymbol{x}'[3,\,4] \text{ での } 0 \text{ の個数”} = 1$$
$$A_{11}(\boldsymbol{x}') = \text{“}\boldsymbol{x}'[3,\,4] \text{ での } 1 \text{ の個数”} = 0$$
$$A_{21}(\boldsymbol{x}') = \text{“}\boldsymbol{x}'[3,\,4] \text{ での } 2 \text{ の個数”} = 1$$
$$A_{02}(\boldsymbol{x}') = \text{“}\boldsymbol{x}'[5,\,5] \text{ での } 0 \text{ の個数”} = 1$$
$$A_{12}(\boldsymbol{x}') = \text{“}\boldsymbol{x}'[5,\,5] \text{ での } 1 \text{ の個数”} = 0$$
$$A_{22}(\boldsymbol{x}') = \text{“}\boldsymbol{x}'[5,\,5] \text{ での } 2 \text{ の個数”} = 0$$
$$\vdots$$

ブロックソート法の CSE 法への帰着

- $\boldsymbol{x} \xrightarrow{\text{BW 変換}} \boldsymbol{x}' = \mathrm{BWT}(\boldsymbol{x})$

  $\downarrow$

  $\{A_w(\boldsymbol{x}')\}$ の符号化

  $\max\{0,\, A_{0w} - A_{w1}\} \le A_{0w0} \le \min\{A_{0w},\, A_{w0}\}$

  (多値化も可能)

- CSE 法と等価であり，CSE 法に対して成り立つ性質がそのまま成立
- CSE 法の復号器がそのまま使える (BW 逆変換を必要としない)

むすび

- 部分列数え上げデータ圧縮法とその関連法を紹介した

  「部分列の個数を数える」という，ただそれだけのことなのに . . .

  分野横断的な多種多様な観点を誘導．しかし，. . .

- これを学ぶ意義

  pedagogical?

謝辞

情報理論研究専門委員会
委員長 大橋正良先生
委員の先生方に
感謝申し上げます.