文字列に含まれる繰り返し構造の頻度について

草野　一彦　　篠原　歩
東北大学　大学院情報科学研究科

1　はじめに

コンピュータが扱う全ての情報は文字列とみなすことができるため，文字列が持つ性質の解明はデータ圧縮や情報処理の効率化のために重要である．繰り返し構造は文字列の基礎的な構造のひとつである．とりわけ左右に延長不可能な有利数回の繰り返しである連（run, maximal repetition）について研究が進められている．長さnの文字列が含む連数は$1.024n$以下である[1]ことや，$0.9446n$個の連を含む文字列[3, 4]が知られているが，長さnの文字列が含む連の厳密な最大個数は示されていない．一方，アルファベットサイズσ長さnの文字列が含む連の平均個数[5]や平均指数（繰り返し回数）和[2]，長さnのネックレスが含む連の平均個数[6]については厳密な値が示されている．

これらの研究は全て連を繰り返し回数が2回以上と定義している．本稿ではこれを$e \geq 1$回以上としたときのアルファベットサイズσ長さnの文字列が含む連の平均個数を示す．

2　定義

Σを有限のアルファベットとし，そのサイズ$|\Sigma|$をσで表す．Σ上の長さnの文字列の集合をΣ^nとする．文字列wのi番目の文字を$w[i]$と書き，部分文字列$w[i]w[i+1]\ldots w[j]$を$w[i..j]$と書く．文字列wと整数$1 \leq p \leq |w|$について任意の位置$1 \leq i \leq |w| - p$で，$w[i] = w[i+p]$が成り立つときpをwの周期と言う．文字列wが$|w|$以外の$|w|$の約数を周期に持たないときwは素であると言う．長さnの素な文字列の集合を$Prim_{n,\sigma}$とする．文字列wと任意の整数$2 \leq i \leq |w|$について，wが$w[i..|w|]$より辞書順で小さいときwをリンドン文字列という．長さnのリンドン文字列の個数を$L(p,\sigma)$とする．メビウス関数$\mu(n)$を用いて次の等式が成り立つことが知られている．

$$|Prim_{n,\sigma}| = nL(n,\sigma) = \sum_{d|n} \mu\left(\frac{n}{d}\right)\sigma^d$$

ここで$d|n$はdがnの約数であることを表す．

文字列wの部分文字列$w[i..j]$が周期pにたいして次の条件を満たすとき，$w[i..j]$を周期pの連という．
(1) $w[i..j]$は周期pを持つ．
(2) 周期pで左右に延長不可能である．すなわち$i = 1$もしくは$w[i-1] \neq w[i+p-1]$であり，$j = |w|$もしくは$w[j+1] \neq w[j-p+1]$である．
(3) 繰り返しの根$w[i..i+p-1]$が素である．
指数が2未満の場合，同一の部分文字列が複数の周期で連となりうるが，本稿ではそのような連をそれぞれの周期で数える．同一の周期と開始位置を持つ連は一意に定まる．

3　主結果

ある位置に特定の部分文字列を含む文字列の個数（補題 1）と，文字列中に連が存在するときに文字列が含む部分文字列（補題 2）から，Σ^nが含む連の総数を求め，平均個数を導出する．

補題 1. アルファベットサイズσ長さnの全ての文字列Σ^nの中で，部分文字列$w[k..k+l-1]$が長さlの文字列の集合$F \subseteq \Sigma^l$の要素であるものの個数

$$N_{F,k}(n,\sigma) = \#\{w: w \in \Sigma^n, w[k..k+l-1] \in F\}$$

は次の値となる．

$$N_{F,k}(n,\sigma) = |F|\sigma^{n-l}$$

証明. Σ^nの要素で位置kにFの要素を含む文字列の集合を

$$C = \{w: w \in \Sigma^n, w[k..k+l-1] \in F\}$$

とする．$N_{F,k}(n,\sigma) = |C|$である．$w \in C$を$l = w[1..k-1]$，$m = w[k..k+l-1]$，$r = w[k+l..n]$と分割すると，それぞれの部分文字列は集合$\Sigma^{k-1}, F, \Sigma^{n-k-l+1}$の要素であり，全ての組み合わせが現れる．すなわち，

$$C = \{lmr: l \in \Sigma^{k-1}, m \in F, r \in \Sigma^{n-k-l+1}\}$$

である．その要素数は

$$|C| = \left|\Sigma^{k-1}\right||F|\left|\Sigma^{n-k-l+1}\right| = |F|\sigma^{n-l}$$

となる．□

整数pと実数$e \geq 1$について文字列の集合$Rn1_{p,e,\sigma}$と$Rn2_{p,e,\sigma}$を次のように定義する．

On the frequency of the repetitions in strings
Kazuhiko KUSANO and Ayumi SHINOHARA
Graduate School of Information Sciences, Tohoku University

$$Rn1_{p,e,\sigma} = \left\{ l m^{\lceil ep \rceil / p} : l \neq m[p], m \in Prim_{p,\sigma} \right\}$$
$$Rn2_{p,e,\sigma} = \left\{ m^{\lceil ep \rceil / p} : m \in Prim_{p,\sigma} \right\}$$

要素の長さはそれぞれ$\lceil ep \rceil + 1$と$\lceil ep \rceil$で，要素数はそれぞれ$(\sigma-1)pL(p,\sigma)$と$pL(p,\sigma)$である．

連が位置kから開始するとき，繰り返しはkより左に延長できないことから次の補題が導ける．

補題 2. 文字列wが位置kに周期pで指数$e \geq 1$以上の連を含むときかつそのときに限りwは次の部分文字列を含む．

$w[k-1..k+\lceil ep \rceil - 1] \in Rn1_{p,e,\sigma}$　$k > 1$ のとき

$w[1..\lceil ep \rceil] \in Rn2_{p,e,\sigma}$　$k = 1$ のとき

証明. 紙面の制約により省略．

定理 1. 任意の整数nとσ，任意の実数$e \geq 1$について，アルファベットサイズσ長さnの文字列が含む指数e以上の連の平均個数は次の値となる．

$$R(n,e,\sigma) = \sum_{p=1}^{\frac{n}{e}} pL(p,\sigma)((n-\lceil ep \rceil + 1)\sigma^{-\lceil ep \rceil} - (n-\lceil ep \rceil)\sigma^{-\lceil ep \rceil - 1})$$

証明. 補題1と補題2から，

$$R(n,e,\sigma) = \frac{1}{|\Sigma^n|}\sum_{w\in\Sigma^n} run(w,e)$$
$$= \frac{1}{\sigma^n}\sum_{w\in\Sigma^n}\sum_{p=1}^{\frac{n}{e}}\sum_{k=1}^{n-\lceil ep \rceil+1} [w\text{の位置}k\text{に周期}p\text{の連がある}]$$
$$= \frac{1}{\sigma^n}\sum_{p=1}^{\frac{n}{e}}\sum_{k=1}^{n-\lceil ep \rceil+1}\sum_{w\in\Sigma^n} [w\text{の位置}k\text{に周期}p\text{の連がある}]$$
$$= \frac{1}{\sigma^n}\sum_{p=1}^{\frac{n}{e}}\left(\sum_{w\in\Sigma^n} [w[1..\lceil ep \rceil] \in Rn2_{p,e,\sigma}] + \sum_{k=2}^{n-\lceil ep \rceil+1}\sum_{w\in\Sigma^n} w[k-1..k+\lceil ep \rceil - 1] \in Rn1_{p,e,\sigma}\right)$$
$$= \frac{1}{\sigma^n}\sum_{p=1}^{\frac{n}{e}}\left(|Rn2_{p,e,\sigma}|\sigma^{n-\lceil ep \rceil} + \sum_{k=2}^{n-\lceil ep \rceil+1} |Rn1_{p,e,\sigma}|\sigma^{n-\lceil ep \rceil-1}\right)$$
$$= \sum_{p=1}^{\frac{n}{e}} pL(p,\sigma)((n-\lceil ep \rceil + 1)\sigma^{-\lceil ep \rceil} - (n-\lceil ep \rceil)\sigma^{-\lceil ep \rceil - 1})$$

ここで，$run(w,e)$はwが含む指数e以上の連の個数である．□

図1に$R(n,e,2)$の実際の値を示す．$e=1$以外では連の平均個数は文字列長nに対して線形となっている．

定理 2. $e > 1$のとき$R(n,e,\sigma)$は$O(n)$である．

証明. $\sigma = 1$のとき長さnの文字列はa^nのみであり，含まれる連はa^nが1つだけである．

$\sigma \geq 2$のとき$\mu(p) \leq 1$と$\lceil epd \rceil \geq epd$が成り立つことから，

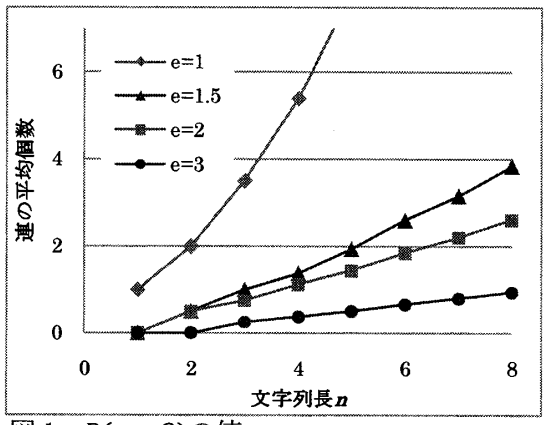

図1：$R(n,e,2)$の値

$$R(n,e,\sigma) = \sum_{p=1}^{\frac{n}{e}} pL(p,\sigma)((n-\lceil ep \rceil + 1)\sigma^{-\lceil ep \rceil} - (n-\lceil ep \rceil)\sigma^{-\lceil ep \rceil - 1})$$
$$= \sum_{p=1}^{\frac{n}{e}}\sum_{d=1}^{\frac{n}{ep}} \mu(p)\sigma^{d-\lceil epd \rceil}((n-\lceil epd \rceil)(1-\sigma^{-1}) + 1)$$
$$< \sum_{p=1}^{\frac{n}{e}}\sum_{d=1}^{\infty} \sigma^{d-epd}((n-epd)(1-\sigma^{-1}) + 1)$$
$$< \sum_{p=1}^{\frac{n}{e}}\sum_{d=1}^{\infty} \sigma^{d-epd} n = n\sum_{p=1}^{\frac{n}{e}} \frac{1}{\sigma^{ep-1}-1} < n\sum_{p=1}^{\infty} \frac{1}{\sigma^{ep-1}-1}$$
$$< n\left(\frac{n}{a}\sum_{p=1}^{t} \frac{1}{\sigma^{ep-1}-1} + \sum_{p=t+1}^{\infty} \frac{1}{\sigma^{ep-2}}\right) = n\left(\frac{n}{a}\sum_{p=1}^{t} \frac{1}{\sigma^{ep-1}-1} + \frac{\sigma^{2-et}}{\sigma^e - 1}\right)$$

ここで，$t = \left\lfloor \frac{2}{e} \right\rfloor$である．括弧の中は$n$の値によらず定数となるため，$R(n,e,\sigma)$は$O(n)$である．□

まとめ

本稿ではアルファベットサイズσ長さnの文字列が含む指数e以上の連の平均個数を表す厳密な式を示した．また$e > 1$のとき平均個数が$O(n)$であることを示した．

参考文献

[1] M. Crochemore, L. Ilie and L. Tinta. Towards a solution to the "runs" conjecture. In *Proceedings of the 19th Annual Symposium on Combinatorial Pattern Matching (CPM 2008)*, Vol. 5029 of LNCS, pp. 290-302, 2008.

[2] Kazuhiko Kusano, Wataru Matsubara, Akira Ishino and Ayumi Shinohara. AVERAGE VALUE OF SUM OF EXPONENTS OF RUNS IN A STRING, *International Journal of Foundations of Computer Science (special issue for Prague Stringology Conference)*, Vol. 20, Issue 6, pp. 1135-1146, 2009.

[3] W. Matsubara, K. Kusano, H. Bannai, and A. Shinohara. A series of run-rich strings. In *Proceedings of the 3rd International Conference on Language and Automata Theory and Applications*, pp. 578-587. 2009.

[4] J. Simpson. Modified Padovan words and the maximum number of runs in a word. *Australasian Journal of Combinatorics (to appear)*.

[5] S. J. Puglisi and J. Simpson. The expected number of runs in a word. Australasian Journal of Combinatorics, 42:45-54, 2008.

[6] 草野 一彦，篠原 歩．ネックレス文字列中の繰り返し構造について，2009年度夏のLAシンポジウム(LA 2009), S9, pp. 1-8, 2009.