文字列情報処理に関する研究

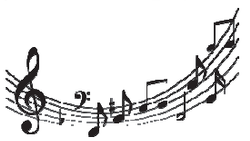
情報学部門 数理情報講座 教授 竹田正幸 准教授 坂内英夫 准教授 稲永俊介

文字列情報学とは

■ 計算機可読なほとんどのデータは
文字列（記号の列）とみなすことができる.

テキストデータ

生物学的データ



音楽データ

➔ **大規模文字列データ処理技術が必要！**

【文字列情報学】

文字列の**組み合わせ的性質**を用いて文字列処理問題の本質を解き明かし，**アルゴリズムとデータ構造技術**によって，高速・省領域な文字列処理手法を実現する.

圧縮文字列処理アルゴリズム

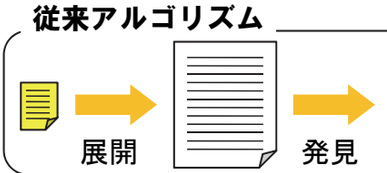
■ 大規模な文字列データを圧縮しておくことで，
記憶領域削減と通信高速化を実現できる.

■ 圧縮の例）$ababababbbaaa = (ab)^5\,b^1\,a^3$

■ しかし，データの中身にアクセスしたいときに
逐一データを展開していたのでは，せっかっく
データを圧縮しておいた利点が失われてしまう.

■ 【本研究】圧縮データを陽に展開することなく，
圧縮したまま高速処理するアルゴリズムを開発.

従来アルゴリズム



提案アルゴリズム

従来手法に対して**劇的な高速化**を達成！
➔ 従来は圧縮サイズ n の指数時間かかって
いたものを，n の低次多項式時間に削減した.

ところで，なぜ圧縮したまま
処理すると高速化できるのだろう？

「ラーメン早食い競争」に例えると分かりやすい

展開してから法

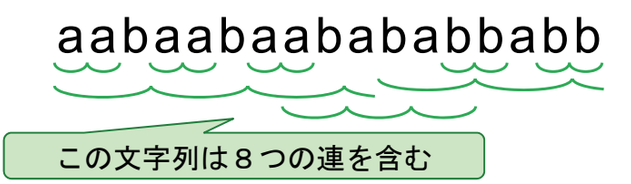
展開しながら法

圧縮したまま法

文字列中の繰り返し構造

■ 文字列中に現れる繰り返し構造は，
文字列の特徴抽出，頻出パターン発見，
データ圧縮などと深い関係がある.

■ 文字列中の極大な繰り返し構造を**連**と呼ぶ.



■ 長さ N の任意の文字列が含む連の最大個数を
$\rho(N)$ と表す.

■ 1999年，Kolpakov と Kucherov は $\rho(N) = O(N)$
を示し，さらに $\rho(N) < N$ と予想した（連予想）.

$\rho(N) < cN$	Kolpakov と Kucherov，1999
$\rho(N) < 5N$	Rytter，2006
$\rho(N) < 3.48N$	Puglisi ら，2006
$\rho(N) < 1.6N$	Crochemore と Ilie，2008
$\rho(N) < 1.029N$	Crochemore ら，2011
$\rho(N) < N$	**我々の研究グループ，2014**

熾烈な国際競争を制し，$\rho(N) < N$ の証明に成功！
➔ **「連定理」**の誕生

その他の研究テーマ

■ パターン照合
- 厳密一致照合，近似照合，順序同型照合，パラメタ化照合

■ 文字列データ構造
- 接尾辞木，接尾辞配列，DAWG，圧縮索引，LCE

■ 文字列組み合わせ論
- 回文構造，ギャップ付き反復，アーベル連

■ データ圧縮
- Lempel-Ziv 圧縮，文法圧縮，連長圧縮

■ 文字列比較
- 編集距離，最長共通部分列 (LCS)，アラインメント