

どんな研究	データ圧縮では数値や文字などを短い0/1の系列に変換するルールが必要です。多くの場合、そのルールは符号木というyes/noチャートのような構造で表します。本研究では、符号木を複数組み合わせ合わせた符号森でより圧縮効率の高いルールを記述できる汎用的な手法を提案しました。
どこが凄い	符号森の設計は一般的には全ての木を同時に最適化する必要があるため難しく、従来法で設計可能な符号森は得意なデータが限定的でした。提案法では、理論解析を基に予め符号木の概形を決め打ちして問題を分解することで、幅広いデータで高効率なルールが設計可能になりました。
めざす未来	データ圧縮は情報通信の根拠を支える重要技術の一つです。効率よく圧縮できると、同じデータならより少ない容量で、同じ容量ならより多くのデータを扱えます。提案法の知見を活かし、あらゆるデータ圧縮効率の底上げを通じて、より快適なコミュニケーションの実現をめざします。

可逆な圧縮符号化

- 画像・音楽・テキストなど、種類を問わずデータは0/1の系列の形で扱われています。
- 可逆な圧縮符号化はデータを短く正確に0/1で表すことを目的とした基盤技術です。
- 圧縮では出現頻度の高いパターンほど短く表せるようなルールを設計することが大事です。

データ (例: テキスト)

abacabbabadaaac

符号①: 000100100001010001001100000010

符号②: 0100110010100100111000110

符号化のルール①

 $a \leftrightarrow 00$ $b \leftrightarrow 01$ $c \leftrightarrow 10$ $d \leftrightarrow 11$

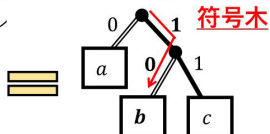
符号化のルール②

 $a \leftrightarrow 0$ $b \leftrightarrow 10$ $c \leftrightarrow 110$ $d \leftrightarrow 111$ ルール②を使った方が
符号を短くできる😊

符号「木」と符号「森」

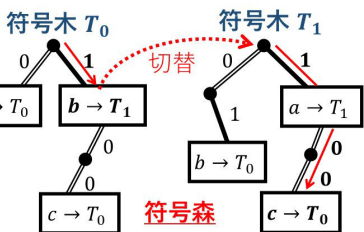
- 符号木で符号化のルールを表現できます。

符号化のルール

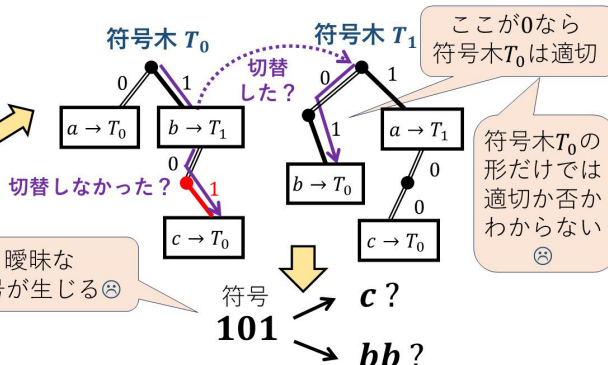
 $a \leftrightarrow 0$
 $b \leftrightarrow 10$
 $c \leftrightarrow 11$


- 複数の符号木からなる符号森でより高効率なルールが作れます。

符号化のルール

 $a \leftrightarrow 0$
 $ba \leftrightarrow 11$
 $bb \leftrightarrow 101$
 $bc \leftrightarrow 1100$
 $\vdots$


- 符号森は全体を適切に設計しないと曖昧な符号になってしまいます。
- 各符号木の適切さは他の木に依ってしまうことが設計を難しくしています。

きちんと
作らないと...

提案法: 木の概形を先に決めて森を作る

- 全ての符号木の「概形」を制約として先に決めておく手法を提案しました。

符号木 T_0
の概形符号木 T_1
の概形符号木 T_2
の概形

...

概形を使って...

- 概形の決め打ちにより、他の木の細かい形に依らずに適切な符号木を設計できるようになりました。
- 理論解析で、使用すべき概形を明らかにできました。

符号木 T_1 $a \rightarrow T_2$

1

1

 $b \rightarrow T_0$ 展開した符号木 T_1 切替先の
概形を展開

展開

展開した木が
概形に収まってい
ればOK!展開は他の木の
細かい形に
依存しない😊

関連文献

- [1] R. Sugiura, Y. Kamamoto, T. Moriya, "General form of almost instantaneous fixed-to-variable-length codes," *IEEE Transactions on Information Theory*, Vol. 69, No. 12, pp. 7672-7690, 2023.
- [2] R. Sugiura, M. Nishino, N. Yasuda, Y. Kamamoto, T. Moriya, "Optimal construction of N-bit-delay almost instantaneous fixed-to-variable-length codes," under review.

連絡先

杉浦 亮介 (Ryosuke Sugiura) 協創情報研究部 言語知能研究グループ