圧縮表現に基づいた大規模データ処理

クップル ドミニク
東京医科歯科大学
koeppl.dsc@tmd.ac.jp

はじめに

はじめまして。東京医科歯科大学のクップルです。現在、私は M&D データ科学センターのアルゴリズム設計・解析分野で研究しています。この分野で、主に大規模データ処理のための方法を考えています。容量の大きなデータを扱うためには、アルゴリズムの速さだけではなく、解析で必要なメモリー領域にも注目しなければなりません。最も基礎的なデータのフォーマットは文字列です。文字列に関わる問題をはじめに、パターン集合・索引構造・データ圧縮・圧縮データ処理等といった問題について私の専門分野では研究しています。

研究

私の研究興味は、主に以下の専門分野です。

- 基礎のアルゴリズム・データ構造（ハッシュ・辞書構造・検索データ構造などです。）
- 文字列の中に、特徴の発見・文字列組み合わせ
- データ圧縮、または圧縮したデータの上で計算

多くの場合、データ圧縮と言えば、zip を用いて圧縮する方法が考えられています。この場合、圧縮されたデータを検索用に利用するとき、復元が必要になるとよく信じられています。しかし、復元せずに、圧縮したままのデータでも検索可能な方法があることを初めて聞いたとき、面白いと思いました。それだけではなく、復元されたデータより、圧縮表現で速く検索できる可能性が高いと考えられます。検索を速く行うため、特徴的な圧縮表現が提案されました。その方法に基づいていくつかの圧縮データ構造が提案されました。圧縮データ構造では、Lempel–Ziv 分解、文法、または Burrows–Wheeler 変換がしばしば利用されます。図 1 で、３番目の Thue–Morse 列 T_3 ($T_1 = \mathsf{a}, T_2 = \mathsf{ab}$, 任意の $k \geq 3$ に対して $T_k = T_{k-1}\overline{T_{k-1}}$、ただし $\overline{\mathsf{a}} = \mathsf{b}$、$\overline{\mathsf{b}} = \mathsf{a}$) の圧縮指標を示します。

これらは一見すると、全く関わりの無い３つの圧縮表現です。近年、いくつかの関係が発見されましたが、最適な圧縮表現はまだ見つかっていません。従来、 Kolmogorov complexity は一般的に計算不可能であり、全ての圧縮表現の下界であると知られています。例えば T_k は帰納的に提案されていますので、 Kolmogorov complexity は $O(1)$ です。最近、提案された最小の文字列のアトラクタも図 1 で示された圧縮表現の下界であることが知られていますが、計算は

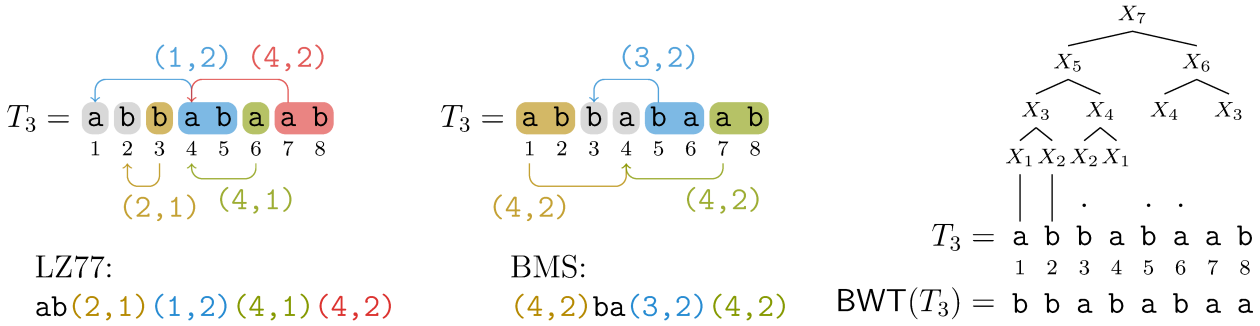

図 1: 3 番目の Thue–Morse 列 T_3 の圧縮表現。左：LZ77 の分割。中：最小の bidirectional macro scheme (BMS) の一つ。右： Burrows–Wheeler 変換 BWT(T_3)、 T_3 の最小のアトラクタ $\{3, 5, 6\}$ (丸で表現される)、 T_3 の最小の文法 $X_1 \to \mathsf{a}$, $X_2 \to \mathsf{b}$, $X_3 \to X_1X_2$, $X_4 \to X_2X_1$, $X_5 \to X_3X_4$, $X_6 \to X_4X_3$, $X_7 \to X_5X_6$。

NP 困難です。アトラクタのように、 bidirectional macro scheme (BMS) の計算は NP 困難であり、LZ77 の下界ですが、Burrows–Wheeler transform といった他の圧縮表現の下界になるとは限りません。

将来の研究。 現在、いくつかの圧縮表現の関係が知られて来ましたが、どんな圧縮表現がどのタスクに適しているのか、まだ明らかにされていません。私は圧縮表現の関係・圧縮表現の計算・圧縮表現に基づいたアルゴリズム・データ構造に興味があります。例えば、文法圧縮の圧縮率は LZ77 より弱いですが、文法のもとで索引構造を容易に提案できます。さらに、どんな圧縮表現でも、未完成があります（圧縮率・検索能力・復元の速さなど)。その未完成を改善できない場合は、複数の圧縮表現の組み合わせを検証することも面白い方法だと思います。

日本での活動経歴

博士課程に在籍していた 2015 年に初めて日本を訪れました。「Studienwerk für Deutsch–Japanischen Kulturaustausch in NRW e.V.」の奨学金を受け、２週間滞在しました。その時は、東京大学の定兼先生を訪問しました。翌年、日本学術振興会のサマー・プログラムに参加し、最初の２週間は定兼先生を訪問し、その後２ヶ月ほど、九州大学の稲永先生の下で研究しました。短期間の滞在でしたが、その間の経験により論文の内容を集めることができました。大変有意義な時間でしたので、さらに長く滞在したいと思うようになりました。博士課程を卒業するとき、日本学術振興会の外国人特別研究員のプログラム（一般）に応募して再び日本を訪れることができました。九州大学の稲永先生に受け入れて頂き、２年間のプログラムに参加しました。その後、東京医科歯科大学で勤めている坂内先生のお陰で、私は２０２０年より助教として勤めています。

私の研究分野について、日本の研究環境は充実しているように思います。まずは、アルゴリズム・データ構造に興味がある研究者はかなり多く、コンピュテーション研究会・情報処理学会アルゴリズム研究会・LA シンポジウムといったいくつかの国内学会が開催され、メール・slack

などの SNS を通じて、研究者の交流も盛んに行われています。その他に、国内の訪問であれば、交通機関が充実しているので困ることが滅多にありません。

ドイツと日本の交流。　奨学金のおかげで来日でき、同じ奨学金を受けた仲間もできました。私の経験を活かすために、校友会員になりました。校友会の目的は、日本に興味がある外国人に来日について相談する活動です。例えば、よく聞かれる質問は言語能力についてです。日本で研究したいけれど、日本語が話せない方が多いです。英語で活動できる研究機関や研究室を見つけることが難しいという悩みを耳にします。そのような来日を希望する外国人の不安を減らすために、各専門、英語で活動できる地図を作ったほうがいいという声も挙げられています。

質問など。　私の研究に関心や興味がある方は、 気楽に聞いてください。さらに詳しい情報を知りたい場合は、私のホームページをご覧ください。https://dkppl.de/