最長共通接頭辞クエリに対する省領域の決定性データ構造とその効率的な構築法について

Deterministic Sub-Linear Space LCE Data Structures with Efficient Construction

H28海自40

派遣先　第27回組合せパターン照合に関する国際会議（Israel・Tel Aviv）
期　間　平成28年6月25日～平成28年7月2日（8日間）
申請者　九州大学　大学院システム情報科学府　情報学専攻
修士2年　谷　村　優　佳

海外における研究活動状況

研究目的

本研究は、時間領域のトレードオフがある最長共通接頭辞クエリに応答するデータ構造として、構築時間がより短いものを提案することを目的とする。本研究は大規模データに対するパターン検索や近似文字列照合などにおいて使用されることが期待される。

海外における研究活動報告

会議の動向

2016年6月27日から6月29日にかけて開催された、27th Annual Symposium on Combinatorial Pattern Matching（CPM 2016）に参加した。この会議は文字列情報処理分野の国際会議であり、文字列検索などの文字列処理アルゴリズムや文字列の組合せ論的性質、バイオインフォマティックスなどを対象とするものである。

また、6月30日にはAmiFest（A workshop on Pattern Matching, Data Structures and Compression.）というワークショップが開催され、文字列処理アルゴリズムなどにおいて著名な研究者による講演が行われた。

他の研究者の発表の中には、私と同じく最長共通接頭辞クエリに関するものも数件含まれていた。そのほか、文字列の組合せ論に関する発表もいくつかあり、応用分野は不明であるものの非常に興味深い内容であった。今後は、それらの研究内容を精読し、本研究を組み合わせることでより良いアルゴリズムを構築できないか、検討していきたいと考える。

発表の概要

文字列処理における典型的な技術の1つにデータ圧縮や文字列検索がある。データ圧縮や文字列検索の例として、LZ分解や接尾辞の整列がよく知られており、これらは最長共通接頭辞クエリを重要な処理の1つとして用いている。最長共通接頭辞クエリlcp（i, j）とは、長さnの文字列Tが与えられたときT上の位置i, jを開始位置とする接尾辞の最長共通接頭辞の長さを問うクエリである。この最長共通接頭辞クエリは、文字列中の回文や繰り返しを見つけることなどにも用いられる。最長共通接頭辞クエリは文字列処理において極めて重要で多くの場面で用いられるクエリであるため、多くの研究者が盛んに研究を行っており、様々なアプローチによるクエリ応答が提案されている。

近年ではクエリ時間とデータ構造の領域でよ

りよいトレードオフを実現するための研究が盛んに行われている。現在、最もよいトレードオフをもつアルゴリズムは、パラメータτ（1≤τ≤n）に対して、O（n/τ）領域のデータ構造でクエリの計算をO（τ）時間で行うことができる。しかし、このアルゴリズムはデータ構造の構築にO（n2+ε）時間（εは0より大きい任意の定数）かかってしまう。そのためデータ数$n=10^4$に対して構築時間が$O(10^{8+\varepsilon})$になるなど、大規模データに対して実用的ではない。

本研究では、時間領域のトレードオフがある最長共通接頭辞クエリに応答するデータ構造として、構築時間がより短いものを提案した。本研究は大規模データに対するパターン検索や近似文字列照合などにおいて使用されることが期待される。

本研究の主な結果は以下の通りである。

1. 前処理O（nτ）時間、データ構造O（n/τ）領域、クエリO（τ log τ）時間の最長共通接頭辞クエリに対するアルゴリズムを提案
2. 先行研究との組み合わせにより、前処理O（nτ）時間、データ構造O（n/τ）領域、クエリO（τlog（n/τ））時間の最長共通接頭辞クエリに対するアルゴリズムを提案

既存研究として存在するいくつかのデータ構造に比べ、本研究は前処理時間を改善できている。

また、提案手法は最長共通接頭辞クエリに対するデータ構造のうち、大規模データに対して現実的に適用可能な初めてのデータ構造である。

以上の内容を、会議の初日である6月27日に発表した。研究内容に関しては、本研究の元となった、先行研究をおこなった研究者からもコメントをいただいた。具体的な意見ではなかったが、本論文で提案したアルゴリズムのさらなる改善が望めるのではないかと考えている。引き続き、今回扱った問題のより良い解法の発見に努めていきたい。

終わりに

今回の会議参加により、自らの研究内容を対外的に発表する機会を得ることができ、それにより研究内容を進展させることができた。また、他の研究者の発表を聞いたり、あるいは議論したりすることで、最新の研究を知ることができ、さらに発表技術などを参考にすることができた。しかしながら、他者の英語を聞き取れないことも多かったため、語学の面については努力の必要性を痛感することとなった。

この会議は、私にとって非常に貴重かつ有意義なものであったと考える。最後に、本会議参加のための渡航費を助成してくださった村田学術振興財団に対し、厚くお礼申し上げます。

この派遣の研究成果等を発表した著書、論文、報告書の書名・講演題目

［著者］

Yuka Tanimura, Tomohiro I, Hideo Bannai, Shunsuke Inenaga, Simon Puglisi and Masayuki Taked

［題目］

Deterministic sub-linear space LCE data structures with efficient construction

（最長共通接頭辞クエリに対する省領域の決定性データ構造とその効率的な構築法について）

［会議名および講演概要］

Combinatorial Pattern Matching（CPM）2016, pp.1-10