学位論文審査の要旨

博士の専攻分野の名称　博士 (情報科学)　　氏名　髙木　拓也

審査担当者　主　査　教　授　有村 博紀
副　査　教　授　湊 真一
副　査　教　授　Zeugmann Thomas
副　査　准教授　喜田 拓也

学位論文題名

Studies on Efficient Index Construction for Multiple and Repetitive Texts
（複数テキストと繰り返しテキストに対する効率の良い索引構築の研究）

ウェブや SNS に代表される情報通信技術の発展によって, 多様で膨大な非構造データが日々生み出されている. これらの非構造データの中でも最も基本的なデータが, 記号を連結した列として表されるテキストデータである. テキスト処理は, 情報科学における最も基本的な問題の一つであり, 同時に, 現在のウェブ情報処理や遺伝子情報処理等, その応用は多岐にわたる.

現在のテキスト処理において最も重要なデータ構造の一つが, 接尾辞木に代表される全文テキスト索引 (text indexing) である. 従来型の逐次テキスト処理では, 入力文字列を先頭から順に読みながら逐次処理するため, 読み込みだけで膨大な時間を要する. これに対して, 1960 年代から, テキストをあらかじめ前処理しておき, 後続して多数回行われる処理を高速化するというテキスト索引方式が提案された. 任意の部分を検索可能な, いわゆる「全文テキスト索引」の開発は長年の目標であったが, 1970 年代から 1980 年代にかけて, 接尾辞木 (suffix tree) と有向非巡回語グラフ (DAWG) の二つの画期的な全文テキスト索引が相次いで提案された.

これらの古典的な全文索引は, 一本のテキストを入力として線形領域と線形構築時間で実現でき, 文字列照合や, エラーを許す近似照合, テキスト圧縮等の様々なアルゴリズムの鍵として広く使用されている. しかし, 2000 年代以降の情報通信環境での現代的テキスト処理応用においては, 多数のデータ生成源から, 複数のテキストを受けとりつつ, 急速に増大する膨大なテキストデータに対して効率よく索引構築を行いつつ, 問合せを高速に処理可能なことが要求される. これは接尾辞木や DAWG のような古典的な全文テキスト索引では対応が難しく, 新しい全文テキスト索引構築技術の確立が不可欠である.

本論文では, そのような現代的テキスト処理応用のための全文テキスト索引構築技術について述べている. 特に課題として, (1) 高速なクエリ処理, および, (2) 複数のデータ発生源への対応, (3) テキスト索引の圧縮の三つの課題を考慮した索引構築が必要である. これらに対して, 本論文では,

(1) テキスト長より小さな準線形時間しか要しない高速な索引構築, および,

(2) 複数の成長するテキストに対するオンライン索引構築,

(3) テキストの本質的なデータ複雑さ程度まで圧縮可能な高圧縮率のテキスト索引構築

を, それぞれ提案している. さらに理論解析により, 提案技術が従来手法の性能を大幅に改善することを示している.

第 3 章では, 高速なクエリ処理を実現可能な索引構築に関して, $\Theta(\log n)$ ビット幅のレジスタ上の

演算が定数時間で実行可能な計算モデル上で, 接尾辞木の基となるデータ構造であるコンパクトトライ (パス圧縮探索木) の文字列照合と, 挿入, 削除の演算をレジスタ幅程度に高速化し, テキスト処理と索引構築を準線形時間で実行可能なことを示している.

第 4 章では, 新しい記号を末尾に非同期に追加しつつ, 複数のテキストが制限なく伸長していくという完全オンライン設定の索引構築問題を提案し, 一般化接尾辞木と呼ばれるテキスト索引を, 入力長の線形時間で完全オンライン構築する世界初のアルゴリズムを与えた. これは, 接尾辞木と, その逆接尾辞木, 有向非巡回語グラフの三者間の非自明な対応付けに基づいて設計されている.

第 5 章では, DNA 配列や Web データのように繰返しの多いテキストに対する圧縮テキスト索引の構築を考察する. 従来のエントロピー圧縮を用いた圧縮索引の限界を打破するために, 接尾辞木から同型な部分木同士をまとめて得られるコンパクト化有向非巡回語グラフ (CDAWG) を改良し, グラフ構造を用いた新しい圧縮手法を開発することで, 繰返しの多いテキストを効率よく圧縮し, 各種の演算を高速に実行可能な L-CDAWG と呼ばれる全文圧縮テキスト索引を与えた.

第 6 章は, 本論文の結論と今後の課題について述べている.

これを要するに, 著者は, 現代的なテキストデータ処理のための全文テキスト索引構築を考察し, その鍵となる高速なクエリ処理と, 複数テキストのオンライン処理, 圧縮索引構築の 3 つの観点から, 効率良い索引構築技術を提案するとともに, 従来研究を上回る性能を実現可能なことを理論的解析によって示しており, 情報科学, 特にデータ工学とアルゴリズム理論の研究の発展に関して貢献するところ大なるものがある. よって著者は, 北海道大学博士 (情報科学) の学位を授与される資格あるものと認める.