博 士（情報科学） 上 村 卓 史

学 位 論 文 題 名

Efficient Construction of Constrained Suffix Trees

（制約付き接尾辞木の効率良い構築）

学位論文内容の要旨

インターネットの急速な発展は世界を変え続けている。これを理論の面から支えているのは数多くの効率的なアルゴリズムであり、中でもテキスト処理アルゴリズムは、検索エンジンをはじめとする応用に直結するだけでなく、他のあらゆるアルゴリズムの基礎であり続けるきわめて重要な分野である。

1973 年に Weiner が提案した接尾辞木は、与えられた文字列に現れるすべての部分文字列を格納する全文索引であり、文字列処理における最も重要なデータ構造の一つである。長さ n のテキスト T に対して、接尾辞木は $O(n \log |\Sigma|)$ 時間、$O(n \log |\Sigma|)$ 領域で構築できることが知られている。ここに、T は有限アルファベット Σ 上の文字列であり、$|\Sigma|$ は Σ の要素数である。この接尾辞木を用いて、長さ m の任意の文字列 P に対し、$O(m \log |\Sigma| + occ)$ 時間で T 中で P が現れるすべての位置を計算できる。ここに、occ は T 中の P の出現回数である。1996 年には Ukkonen が、接尾辞木をオンライン構築する線形時間アルゴリズムを与えている。

本研究では接尾辞木を拡張し、テキストに対し索引を作る際、すべての部分文字列を含めるのではなく、様々な制約のクラスに対して、制約付き接尾辞木の効率よい構築について考察する。例えば、「テキスト中に 10 回以上現れる任意の文字列」のような制約を与え、その制約を満たすすべての部分文字列を格納する索引構造を構築するというのが、本研究で扱う問題である。

初めに、第 3 章では、単語幅限定接尾辞木 (word-based truncated suffix tree) を提案し、その効率良い構築について考察する。ここでは、入力とする文字列として、英語のような自然言語テキストを仮定し、各単語の間は空白のような文字で区切られていると仮定する。このとき、与えられた整数 k に対して、その単語幅限定接尾辞木は、テキスト中に現れる長さが k 単語以下のすべての部分文字列を格納する索引構造である。本章では、Ukkonen による接尾辞木のオンライン線形時間構築アルゴリズムを拡張し、任意の整数 k と長さ n.の任意のテキストに対し、その単語幅限定接尾辞木を $O(n \log |\Sigma|)$ 時間でオンライン構築するアルゴリズムを示す。

次に、第 4 章では、Amir らが 2006 年に提案したプロパティ接尾辞木の効率良い構築について考察する。プロパティとは互いに重なりを許したテキスト上の区間の集合である。プロパティ接尾辞木とは、あるテキスト T とプロパティ Π が与えられたとき、Π 中のいずれかの区間内に含まれるすべての T の部分文字列を格納する索引構造である。この問題に対して、Amir らはプロパティ接尾辞木の $O(n \log |\Sigma| + n \log \log n)$ 時間構築アルゴリズムを与えている。これに対して、本章では、プロパティに含まれる区間の内側と外側を分ける境界 (border) を計算する境界計算問題に注目する。この境界計算問題に対して、接尾辞リンクの巡回を用いた効率良いアルゴリズムを与え、プロ

パティ接尾辞木が $O(n \log |\Sigma|)$ 時間で構築できることを示す。

第 5 章では、単語幅限定接尾辞木を拡張したデータ構造である単語 N グラム木 (word N-gram tree) とその構築アルゴリズムを提案する。第 3 章で提案した単語幅限定接尾辞木では、単語の先頭から始まる文字列だけでなく、それらの任意の部分文字列も索引化するが、これは応用によっては好ましくない。一方で、単語の先頭から始まる文字列のみを扱う索引構造として、Anderson らが 1999 年に提案し、Inenaga らが 2006 年にオンライン構築アルゴリズムを与えた単語接尾辞木 (word suffix tree) があるが、これは部分文字列の開始点のみを制約し、終了点は制約しない。これらに対して、単語 N グラム木は、部分文字列の開始点と終了点の両方を単語区切り位置で制約したものであり、単語接尾辞木と単語幅限定接尾辞木の両方のアイディアを用いた拡張である。また、本章では、単語 N グラム木の応用として、ウェブ閲覧におけるキーワード抽出を考察する。ウェブページからの情報収集では、俗語や新語の他、本や映画の題名など、単一の名詞だけでなく、ある程度長い単語列が一つのキーワードとして抽出され方が望ましい場合が考えられる。本章では、与えられた N に対し、任意の N 単語以下の連接をキーワードの候補として考え、単語 N グラム木を用いたキーワード抽出アルゴリズムを与える。計算機実験では、小説のほかブログ記事を対象としたキーワード抽出を行い、その有用性を考察した。

第 6 章では、接尾辞木を用いた教師無しスパム文書検出アルゴリズムを提案する。スパム文書とは広告を主な目的として不特定多数に送信される文書であり、ここでは、事前に訓練データなどによる予備知識を与えることなく、与えられた文書集合だけを用いて、その中のスパム文書を検出する問題を考察する。ここでは、与えられた文書集合 D から推定される確率モデル上を構築し、文書集合中の各文書 d に対して、その推定された生成確率 $P(d|D)$ が他の文書に比べて著しく小さいときに、スパム文書と判定する方法を提案する。これは、スパム作成者はコピーなどにより低いコストで大量にスパムを生成しているため、スパム文書の生成確率はこのモデル上で極端に高い値を示すという性質を利用している。本章では、文書集合 D の接尾辞木に基づく確率モデルを提案し、これを用いて、すべての文書 d に対する生成確率 $P(d|D)$ を全文書の合計長の線形時間で計算するアルゴリズムを与える。計算機実験では、提案手法を Narisawa らが 2007 年に提案した接尾辞木を用いる手法と比較し、従来手法では検出が困難とされているワードサラダと呼ばれているスパム文書の検出における提案手法の優位性を示した。

第 7 章では、接尾辞木の応用である STVF 符号と呼ばれる圧縮手法に対して、圧縮率を改善する手法を提案する。VF(variable-length-to-fixed-length) 符号は圧縮手法の一つであり、可変長の文字列の集合に固定長の符号を与えた辞書を構築し、テキストを辞書中の要素の列に変換する手法である。このため VF 符号では、より長く、かつテキストに頻出する文字列を選択することが重要である。STVF 符号は接尾辞木を用いることで、テキスト中の任意の部分文字列とその出現頻度を考慮して辞書を構築する。本章では、一度構築した辞書を用いて圧縮を行い、辞書中のどの文字列がテキスト中でよく使われるかを評価し、この評価に基づいて辞書を作り直す手法を提案する。この操作を繰り返し適用することにより、元の辞書を用いた圧縮よりさらに圧縮率を向上させる。計算機実験では、提案手法を用いて辞書を訓練することで、よく知られた圧縮手法である gzip と bzip2 の中間程度の圧縮率を得られた。

以上のように、本論文では、テキストに対する制約を考慮した接尾辞木の構築という問題について考察し、単語幅限定接尾辞木、プロパティ接尾辞木の線形時間構築アルゴリズムを与えた。また、キーワード抽出および、スパム検出、データ圧縮への応用を行った。

学位論文審査の要旨

主　査　教　授　有　村　博　紀
副　査　教　授　原　口　　誠
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

Efficient Construction of Constrained Suffix Trees

(制約付き接尾辞木の効率良い構築)

文字列処理 (string processing) は, 理論情報科学とデータ工学の副分野の一つであり, 記号系列に対する格納や, 索引付け, 検索, 比較, 発見, 圧縮などの各種の問題の効率良い解法を研究する. 最近では, ネットワーク上の大規模データに対する高速な線形時間アルゴリズムやオンラインアルゴリズムの設計・開発の研究に興味が持たれている. この文字列処理の方式には, おおまかに分類して, 直接入力データ (テキスト) を走査してパターンの情報を計算するパターン照合方式と, テキストを前処理して適切なデータ構造構築しておき, データ構造を用いて問い合わせ (パターン) に答える索引方式の二つの方式がある.

接尾辞木 (suffix trees) は, 1973 年に Weiner が提案し, 1976 年に McCreight が改良した索引方式のデータ構造であり, 前処理において, 与えられたテキストのすべての部分文字列を線形記憶領域でコンパクトに格納し, 問い合わせ時には, 文字単位の効率良い検索を可能にする. 接尾辞木は多数の応用をもち, 文字列処理における最も重要なデータ構造の一つとなっている. 最近, 接尾辞木の研究において, メタ情報や区切り構造等の制約を付加した「制約付きテキスト」に対する構築の研究が盛んになっているが, 接尾辞木の複雑さから線形時間の高速なアルゴリズムの設計は難しく, 理論的な成果はわずかしか得られていなかった.

そこで, 本研究は, このような背景から, 制約付きテキストに対する索引として, 「制約付き接尾辞木」(constrained suffix trees) という新しい概念を提案し, 各種の制約付き接尾辞木構築問題に対して, Ukkonen のアルゴリズムを拡張して, 高速な構築アルゴリズムを与えている. さらに, 応用として, テキストマイニングと, スパム検出, テキスト圧縮の三つの応用分野に, 提案した手法を適用し, その有用性を示している. 特に, アルゴリズム設計の観点からは, この種の拡張された接尾辞木構築問題に対して, 基本アルゴリズムである Ukkonen の構築アルゴリズムが自然に拡張できるかどうかは長年の未解決問題であった. これに対して, 本研究では, その基本機構である「接尾辞リンク」が各種の制約付き接尾辞木に拡張可能であることを示し, Ukkonen の線形時間構築アルゴリズムが自然に一般化できることを示している. このことは工学的にも重要であり, 簡潔かつ効率良い構築アル

ゴリズムの開発に貢献している.

第一の研究では, 単語区切りテキストに対する k 単語限定接尾辞木を提案し, その線形時間構築アルゴリズムを与えている. これは, 従来研究の文字単位長さ限定接尾辞木 (Na et al., Theor. Comp. Sci., Vol.304, 2003) の拡張になっており, 大規模自然言語処理に有用である. 第二の研究では, テキスト上に区間設定の構造をもつプロパティ接尾辞木の構築問題を考察し, その線形時間アルゴリズムを与えている. これは, 2008 年の Amir 他 (Theor. Comp. Sci., Vol.395) による先行研究の $O(n \log \log n)$ 時間アルゴリズムの計算量を改善し, 世界で最初の最適な線形時間アルゴリズムを与えた点で意義がある. 第三の研究では, 単語 k グラム木を提案し, テキストマイニングにおけるキーワード抽出問題における効率的なスコア計算に適用している. これは, 第一の研究の自然な一般化であり, 後に著者らにより, オンライン線形時間構築アルゴリズムが開発されている. 第四の研究では, 接尾辞木をスパム検出問題に適用し, MO(maximal overlap method) と呼ばれる未知文字列の出現確率推定法の線形時間アルゴリズムを与えている. 第五の研究では, 接尾辞木をテキスト圧縮に適用し, 互いにオーバーラップしないという大域的制約を考慮した部分文字列発見を, 接尾辞木上で行う効率良い手法を与えている.

本論文の成果は, 以下にまとめられる.

1. 大規模テキスト処理における制約付きテキスト索引について, 「制約付き接尾辞木」という新しい概念を提案し, その統一的扱いの有効性を示した.

2. 種々の「制約付き接尾辞木」構築問題に対して, 基本アルゴリズムの一つである Ukkonen のアルゴリズムを一般化して, 効率良い線形時間アルゴリズムを開発した. これにより, 理論計算機科学におけるいくつかの未解決問題を肯定的に解決した.

3. 実際の大規模テキスト処理問題に提案手法を適用し, その有効性を示した.

これを要するに, 著者は, 制約付きテキストに対する文字列索引構造の高度化という課題について, 制約付き接尾辞木という新しい概念を提案し, 種々の索引構築問題に対して, 効率良い構築アルゴリズムが構築可能であるという新知見を得たものであり, 理論計算機科学とデータ工学におい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.