修士学位論文要約（令和 4 年 3 月）

パラメタ化コンパクト接尾辞グラフのオンライン構築手法

市川 慎太朗
指導教員：篠原 歩，学位論文指導教員：吉仲 亮

Online Construction of Parameterized Compact Directed Acyclic Word Graphs

Shintaro ICHIKAWA
Supervisor: Ayumi SHINOHARA, Research Advisor: Ryo YOSHINAKA

Two parameterized strings of equal length are said to match if there is a bijection on the symbols such that the two parameterizd strings become identical. The parameterized pattern matching problem is to look for substrings in a given text that parameterizd match a given pattern. Parameterized Compact Directed Acyclic Word Graph (PCDAWG) is a space efficient indexing structure for parameterized pattern matching. We propose the first online algorithm that directly constructs PCDAWG.

1. はじめに

長さ n のテキスト文字列 T から，長さ m のパターン文字列 P の出現位置を見つける文字列照合問題は，古くから盛んに研究が行われている．文字列照合問題を効率的に解くための索引構造が数多く提案され，それらに対してオンライン構築アルゴリズム [3] が研究されている．

それに対して，厳密文字列照合問題のさまざまな変種が提案されている．文字列の構造を捉えた照合を行う事ができるパターン照合として，パラメタ化パターン照合 [1] が存在する．パラメタ化パターン照合では，定数アルファベット Σ，変数アルファベット Π としたときに，$(\Sigma \cup \Pi)^*$ 上の文字列であるパラメタ化文字列を考える．2 つのパラメタ化文字列 x, y がパラメタ化一致するとは，x 中の変数文字を全単射 $f : \Pi \to \Pi$ によって置き換えることによって y に一致させることができる f が存在することである．パラメタ化パターン照合問題は，様々な応用先が提案されており，ソフトウェアのメンテナンス [1]，RNA の構造解析 [6] などがある．また，通常の文字列照合問題における索引構造をパラメタ化パターン照合問題に拡張した様々な索引構造 [1] [4] が提案され，それらのオンライン構築アルゴリズム [6] [4] の研究がなされている．しかし，パラメタ化コンパクト接尾辞グラフについては，テキストから直接構築する手法は提案されておらず，パラメタ化コンパクト接尾辞グラフよりもサイズの大きい別の索引構造を変形することによってしか得ることができていない．つまり，構築の際には，目標とする索引構造よりも多くのスペースを要する手法しか知られていない．

2. パラメタ化コンパクト接尾辞グラフ

パラメタ化コンパクト接尾辞グラフ *(Parameterized Compact Directed Acyclic Word Graph, PCDAWG)* [5] は，Nakashima らによって提案された索引構造であり，パラメタ化接尾辞グラフ [4] 中の源点を除く出次数が 1 の頂点を削除し，残った頂点同士を結んでできるグラフである．また，パラメタ化接尾辞木 [1] を変形することによっても得られる．これらの索引構造同士の関係を図 1 に示す．パラメタ化コンパクト接尾辞グラフでは，図 1 のように 1 つの頂点から，同じ開始ラベルを持つ辺が複数存在する場合がある．これらの辺は行き先が異なっており，そのままだと正しい遷移辺を選択することができない．そこで，今までたどった文字数の区間を条件として持たせることで，正しい遷移辺を選択できる．区間で条件付けられた辺を条件付き辺といい，区間 $(i : j)$ を持つような条件付き辺は，その頂点にたどり着くまでに，i 文字以上 j 文字以下たどってきた後にのみたどることができる．

3. 本研究の成果

本研究では，パラメタ化パターン照合における索引構造の 1 つである，パラメタ化コンパクト接尾辞グラフをオンライン直接構築するアルゴリズムを提案する．本研究で提案するアルゴリズムは，通常の文字列に対するコンパクト接尾辞グラフのオンライン構築アルゴリズム [3] をベースとしている．パラメタ化コンパクト接尾辞グラフでは，条件付き辺を持つため，コンパクト接尾辞グラフのオンライン構築アルゴリズムをそのまま適用することはできない．

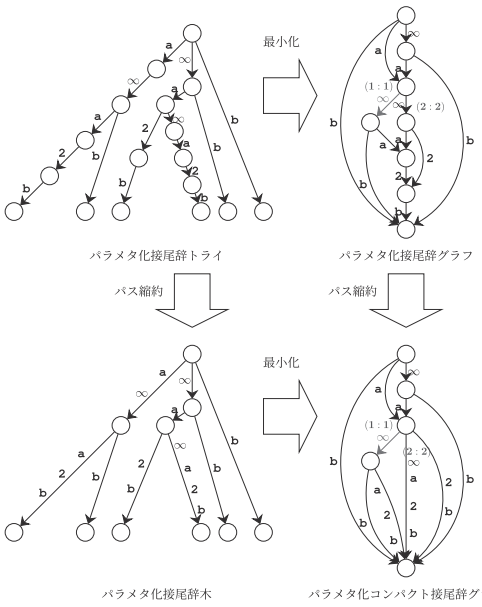

図 1. $\Sigma = \{\mathtt{a}, \mathtt{b}\}$, $\Pi = \{\mathtt{X}, \mathtt{Y}\}$, $w = \mathtt{XaYaYb}$ に対する索引構造同士の関係．丸は頂点，実線は辺を表す．

そこで，パラメタ化接尾辞木，パラメタ化接尾辞グラフのオンライン構築アルゴリズム [6) 2)] のアイデアを使用することで，パラメタ化コンパクト接尾辞グラフをオンライン構築する．構築時間はテキスト長 n に対して $O(n|\Pi|^2 \log(|\Sigma| + |\Pi|))$ 時間である．アルゴリズムの動作例を図 2 に示す．

4. まとめと今後の課題

本研究では，パラメタ化パターン照合における索引構造の 1 つである，パラメタ化コンパクト接尾辞グラフをオンライン直接構築するアルゴリズムを提案した．パラメタ化接尾辞木は乱択アルゴリズムによって $O(n)$ 時間，パラメタ化接尾辞グラフは $O(n|\Pi| \log(|\Sigma| + |\Pi|))$ 時間でオンライン構築できるため，パラメタ化コンパクト接尾辞グラフのより高速な構築アルゴリズムの開発が課題となっている．

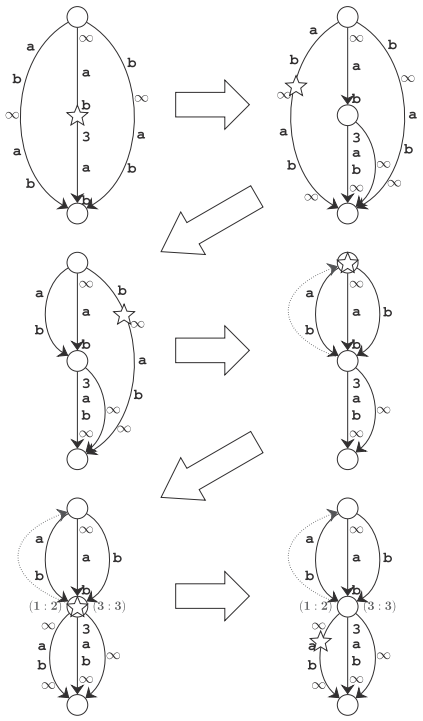

図 2. $\Sigma = \{\mathtt{a}, \mathtt{b}\}$, $\Pi = \{\mathtt{X}, \mathtt{Y}\}$, $w = \mathtt{XabXab}$ のとき，新たに Y を読んだときの更新の様子．

文献

1) Brenda S. Baker. A theory of parameterized pattern matching: algorithms and applications. In *Proc. 25th annual ACM symposium on Theory of computing*, pp. 71–80, 1993.
2) Anselm Blumer, Janet Blumer, David Haussler, Andrzej Ehrenfeucht, Mu-Tian Chen, and Joel Seiferas. The smallest automaton recognizing the subwords of a text. *Theoretical Computer Science*, Vol. 40, pp. 31–55, 1985.
3) Shunsuke Inenaga, Hiromasa Hoshino, Ayumi Shinohara, Masayuki Takeda, Setsuo Arikawa, Giancarlo Mauri, and Giulio Pavesi. On-line construction of compact directed acyclic word graphs. *Discrete Applied Mathematics*, Vol. 146, No. 2, pp. 156–179, 2005.
4) Katsuhito Nakashima, Noriki Fujisato, Diptarama Hendrian, Yuto Nakashima, Ryo Yoshinaka, Shunsuke Inenaga, Hideo Bannai, Ayumi Shinohara, and Masayuki Takeda. DAWGs for Parameterized Matching: Online Construction and Related Indexing Structures. In *31st Annual Symposium on Combinatorial Pattern Matching, CPM 2020*, 2020.
5) Katsuhito Nakashima, Noriki Fujisato, Diptarama Hendrian, Yuto Nakashima, Ryo Yoshinaka, Shunsuke Inenaga, Hideo Bannai, Ayumi Shinohara, and Masayuki Takeda. DAWGs for parameterized matching: online construction and related indexing structures. *arXiv:2002.06786*, 2020.
6) Tetsuo Shibuya. Generalization of a Suffix Tree for RNA Structural Pattern Matching. In *SWAT 2000*, pp. 393–406, 2000.