1L-03

辺重み付き系列二分決定グラフによる頻出部分列の多重集合表現とその評価

杉野 創 † 京都大学 †
川原 純 ‡ 京都大学 ‡
湊 真一 § 京都大学 §

1 概要

本研究では，Loekito らによって提案された Sequence Binary Decision Diagram (SeqBDD) [1] を拡張し，文字列の多重集合を表現する辺重み付き系列二分決定グラフ (Edge-Valued SeqBDD, EV-SeqBDD) を提案する．EV-SeqBDD は SeqBDD の枝に重みを付与することで各パスに重みを付け，各文字列に数値を対応させる．それにより，文字列集合中の各要素の出現回数を保持する．また，本研究は頻出部分列集合を含む様々な文字列多重集合データに対して EV-SeqBDD を構築し，Weighted SeqBDD (W-SeqBDD) [1] による表現と節点数を比較し評価する．

2 準備

2.1 文字列

文字列 s はアルファベットを Σ としたとき，$s \in \Sigma^*$ と表現できる．ここで，Σ 中の任意の 2 つの文字 α, β 間に $\alpha < \beta$，または $\alpha > \beta$ なる関係が定義されている．

ある文字列が複数回出現する集合を文字列多重集合と呼ぶ．文字列多重集合の要素は $s : \phi$ と表現され，文字列 s が集合中で ϕ 回出現することを表す．ただし，$\phi \in \mathbb{N}$ である．

2.2 SeqBDD

SeqBDD[1] は，文字列集合を効率的に表現する二分決定グラフである．SeqBDD の非終端節点は 1 つのラベルと 2 つの子節点を持つ．SeqBDD の非終端節点 N を，ラベルを $l \in \Sigma$，0 枝/1 枝に接続される子節点をそれぞれ N_0, N_1 を用いて $N = (l, N_0, N_1)$ と表現する．ここで，N_0 のラベルが l_0 であるとき，必ず $l < l_0$ が成り立つ．一方，N_1 のラベルとの間にはこのような制約はない．N から N_0 に向かう枝を 0 枝，N から N_1 に向かう枝を 1 枝と呼ぶ．

SeqBDD には 0-終端節点と 1-終端節点の 2 つの終端節点が存在する．これらはラベル，子節点を持たない．

根節点から終端節点までの 1 つのパスが 1 つの文字列を表現し，ある節点での 1 枝の選択はその節点のラベルの文字が文字列に出現することを意味する．1-終端節点に到達するパスに対応する文字列が SeqBDD が表現する文字列集合の要素である．

図 1 に SeqBDD の例を示す．図 1 は文字列集合 $S = \{\mathtt{aa,aab,ab,abb,b,bb}\}$ を表現する．図中の実線が 1 枝，破線が 0 枝を表す．

SeqBDD は等価な節点を共有し，冗長な節点を除去する．SeqBDD の 2 つの節点 P と Q において，ラベル，2 つの子節点が全て同じであるとき，P と Q は等価である．また，SeqBDD の節点 $N = (l, N_0, N_1)$ において，N_1 が 0-終端節点であるときに節点 N は冗長である．節点を共有，除去することを簡約化と呼ぶ．

SeqBDD は文字列集合を一意に表現でき，ZDD [2] から継承された様々な演算を適用可能である．さらに，SeqBDD の様々な性質や応用が研究されている [3, 4].

3 Edge-Valued SeqBDD

本研究は，文字列多重集合を表現する，辺重み付き系列二分決定グラフ (Edge-Valued SeqBDD, EV-SeqBDD) を提案する．EV-SeqBDD の非終端節点は，1 つのラベルと 2 つの子節点に加えて，1 枝に重みを持つ．EV-SeqBDD の節点 N を，ラベルを $l \in \Sigma$，0 枝/1 枝に接続される子節点をそれぞれ N_0, N_1，1 枝の重み $v \in \mathbb{Z}$ を用いて $N = (l, N_0, N_1, v)$ と表現する．ここで，N_0 のラベルが l_0 であるとき，必ず $l < l_0$ が成り立つ．EV-SeqBDD の 1 枝には重みが付けられる．

EV-SeqBDD には 0-終端節点のみが存在する．これはラベル，子節点を持たない．以降では単に終端節点と呼ぶ．

根節点から終端節点までの 1 つのパスが 1 つの文字列を表現する．さらに，そのパスの重みがパスに対応する文字列の出現回数を表現する．パスの重みとは，根節点から終端節点までに選択した枝の重みの総和に根節点に入力する枝の重み（以降 w_ϵ と表現する）を加えたものである．ただし，すべての 0 枝の重みは 0 である．終端節点に到達するパスに対応する文字列 s とそのパスの重み $\phi \in \mathbb{N}$ を用いて，$s : \phi$ は SeqBDD が表現する文字列多重集合の要素である．文字列多重集合を表現する EV-SeqBDD は根節点 N と N への入力 $w_\epsilon \in \mathbb{Z}$ のペア (N, w_ϵ) で与えられる．w_ϵ は任意の集合を表現するために必要である．例えば，$\{\mathtt{a} : 2, \mathtt{b} : 1, \epsilon : 1\}$ のような集合を表現する際，$w_\epsilon = 1$ がなくてはならない．なぜなら，空列 ϵ は 0 枝のみを選択したパスに対応する文字列であり，EV-SeqBDD は 0 枝に重みを持たないためである．そのため，根節点への入力に重みを持たせることで ϵ の出現回数を表現している．

図 2 に EV-SeqBDD の例を示す．図 2 は根節点 R とその入力 $w_\epsilon = 0$ を用いて $(R, 0)$ で与えられ，文字列多重集合 $M = \{\mathtt{aa} : 1, \mathtt{aab} : 2, \mathtt{ab} : 2, \mathtt{abb} : 3, \mathtt{b} : 2, \mathtt{bb} : 3\}$ を表現する．図中の実線が 1 枝，破線が 0 枝を表し，実線に添えられた数が 1 枝の重みである．

EV-SeqBDD は SeqBDD 同様に等価な節点を共有し，冗長な節点を除去する．EV-SeqBDD の 2 つの接点 P, Q が等価で

Edge-Valued Sequence BDDs for Representing Multisets of Frequent Subsequences and Its Evaluation

† So SUGINO, Kyoto University
‡ Jun KAWAHARA, Kyoto University
§ Shin-ichi MINATO, Kyoto University

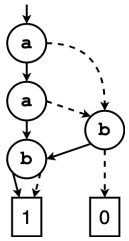

図 1　SeqBDD

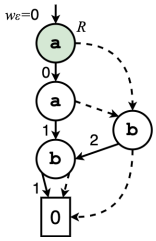

図 2　EV-SeqBDD

あるとは，それぞれのラベル，2 つの子節点，1 枝の重みが全て等しいことを言う．また，冗長な節点とは，以下の条件を満たす節点 $N = (l, N_0, N_1, v)$ である．

- N_1 が終端節点
- 根節点から N までのパスの重みを w とし，$w + v = 0$

節点を共有，除去することを簡約化と呼ぶ．この簡約化規則は，SeqBDD[1] の簡約化規則を拡張したものである．

EV-SeqBDD は文字列多重集合を一意に表現できることが証明できるが，本稿では省略する．

4　評価実験

EV-SeqBDD と Loekito によって提案された W-SeqBDD [1] をそれぞれ用いて，様々な文字列多重集合を表現した．W-SeqBDD とは，各節点にその節点が表現する文字列多重集合のサイズを情報として保持するデータ構造である．本稿では，4 つの文字列多重集合を表現したグラフに関して，節点数を比較することで効率を評価する．文献 [1] と同様に，逐次的な和集合演算による追加によりグラフを構築した[*1]．なお，構築の時間は W-SeqBDD と EV-SeqBDD の間に違いはなく，本稿で扱うデータに対しては同等の時間で構築が可能である．データは，遺伝子配列データの例としてウイルスの遺伝子配列データである DENV1 [5]，自然言語の例として TEXT [6] を用いた．DENV1 は，1 本の配列を長さ 30 ずつで区切って使用し，A, C, G, T の 4 つをアルファベットとした．TEXT は出現する単語の集合をアルファベットとした．さらに，それらのデータで頻出な部分列の集合を求め，それぞれ Freq_DENV1, Freq_TEXT として表現した．ここで，最小支持度を Freq_DENV1 は 0.7，Freq_TEXT は 0.1 として計算した．表 1 にそれぞれの特徴をまとめる．表中の $|\Sigma|$ はアルファベットのサイズ，C, S, L がそれぞれ総文字数，集合の要素数，最大文字列長を表す．

表 2 に EV-SeqBDD と W-SeqBDD で表現したときの節点数をまとめる．DENV1，TEXT に関しては，通常の SeqBDD と同じ形になるため．節点数は同じである．対して，Freq_DENV1 では節点数が 30% 近く，Freq_TEXT では 15% ほど削減されている．特に，遺伝子配列中の頻出部分列を表現するとき，EV-SeqBDD はより効果的である．

頻出部分列の特徴として，ある部分列 p が頻出であるとき，その部分列もまた頻出である．つまり，頻出部分列集合は，集合中のある要素の全ての部分列もまた集合の要素であるという性質を持つ．EV-SeqBDD はこのような集合を表現することに適している．したがって，遺伝子配列のような長い頻出部分列が発見される集合を扱う場合に，EV-SeqBDD はより効果的である．

[*1] 本稿では素朴な手法を用いているが，構築手法については改善の余地がある．

表 1　用いたデータセット

データ名	$\|\Sigma\|$	C	S	L
DENV1	4	10710	357	30
Freq-DENV1	-	82045	12607	9
TEXT	438	1430	65	71
Freq-TEXT	-	410	196	4

表 2　節点数の比較

データ名	EV-SeqBDD	W-SeqBDD
DENV1	8105	8106
Freq-DENV1	7808	10725
TEXT	1355	1356
Freq-TEXT	186	214

5　まとめ

本研究では，同一の文字列を複数含む多重集合を表現する EV-SeqBDD を提案した．EV-SeqBDD は既存手法と比較し，集合データをより圧縮することが可能である．特に，頻出部分列集合を扱う場合により効率的に表現することができる．今後の課題として，SeqBDD に適用可能な多様な集合演算の EV-SeqBDD への適用や，頻度以外の数値と文字列の対応づけによるさらなる応用が挙げられる．

本研究の一部は，JSPS 科研費 JP20H00605, JP20H05794, JP20H05964 の助成を受けたものです．

参考文献

[1] Loekito, E., Bailey, J. and Pei, J.: A Binary Decision Diagram Based Approach for Mining Frequent Subsequences, *Knowl. Inf. Syst.*, Vol. 24, No. 2, pp. 235–268 (2010).

[2] Minato, S.: Zero-Suppressed BDDs for Set Manipulation in Combinatorial Problems, *Proceedings of the 30th International Design Automation Conference*, DAC '93, pp. 272–277 (1993).

[3] Denzumi, S.: New Algorithms for Manipulating Sequence BDDs, *Implementation and Application of Automata*, pp. 108–120 (2019).

[4] Denzumi, S., Yoshinaka, R., Arimura, H. and Minato, S.: Sequence binary decision diagram: Minimization, relationship to acyclic automata, and complexities of Boolean set operations, *Discrete applied mathematics*, Vol. 212, pp. 61–80 (2016).

[5] NCBI: Dengue virus type 1 clone 45AZ5, complete genome. `https://www.ncbi.nlm.nih.gov/nuccore/1854038`.

[6] Lit2Go: The Adventures of Huckleberry Finn by Mark Twain. `https://etc.usf.edu/lit2go/21/the-adventures-of-huckleberry-finn/`.