

頻出順序木パターンを見つけるオンラインアルゴリズム

Online Algorithms for Discovering Frequent Substructures from Semi-structured Data Stream

浅井達哉
Tatsuya Asai

安部賢治
Kenji Abe

川副真治
Shinji Kawasoc

有村博紀
Hiroki Arimura

有川節夫
Setsuo Arikawa

九州大学大学院システム情報科学府・研究院
Department of Informatics, Kyushu University

1 はじめに

データマイニングとは、データベースに蓄積された大量のデータから、自明でない規則性やパターンを半自動的にとりだす方法についての科学研究であり、ビジネス分野や科学技術分野などのさまざまな対象分野で、その適用が盛んに行なわれている。また、高速なネットワークと安価な大容量記憶装置の発達によって、ウェブページやXML文書に代表される半構造データ [2, 18] がネットワーク上に蓄積されており、大規模半構造データを対象としたデータマイニング（半構造データマイニング）に関する研究が盛んに行なわれつつある [1, 5, 9, 12, 14, 19, 21]。

一方、高速ネットワークとウェブ技術の発達によって、近年、電子商取引やウェブサイト管理など、新しい形態のネットワーク上のアプリケーション普及し始めている。これらのアプリケーションでは、データは静的なレコードの集合ではなく、膨大な半構造データがネットワーク上を流れ続けるデータストリームの形態をしていることが特徴である。今後、このような大規模半構造データストリームを対象とした、効率のよいデータマイニング手法の重要性が増すものと予想される。本稿で考察する半構造データストリームは、

1. 膨大な量のデータが、
2. 時間的に変化しながら、
3. 連続して流れ続ける
4. 高速なデータストリーム

であることが特徴である。よって、このような大規模半構造データストリームを対象とするアルゴリズムは、制限された計算資源しか使わずに長時間働き続け、任意の時点で有益な近似解を提供できることが望ましい。このアルゴリズムが実現すれば、ネットワークログからの異常検出やXMLデータストリームからのトレンド検出などへの応用が期待できる。

我々は、半構造データとパターンのモデルとしてラベルつき順序木を採用し、先に開発した順序木の効率よい枚举法 [5] を用いて、半構造データストリームをラベルつき順序木の節点列として定式化する。これは、データストリーム上を、巨大なXMLデータが連続的に配送される場合をモデル化している。そして、半構

造データストリームに対する頻出パターン発見アルゴリズム StreamT を開発する [6]。このアルゴリズムでは、掃木枝上のボトム出現を逐次的に更新することにより、パターンの出現位置を漸増的に計算している。この手法と Hidber [11] による候補集合の管理戦略を組み合わせることにより、アルゴリズムは無限のデータストリームに対して有限の資源しか用いずに効率的に働く。

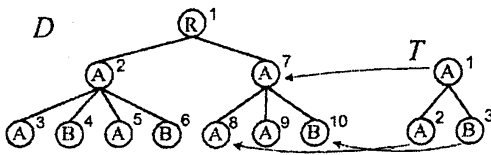
本稿の構成は次のとおりである。2節で、半構造データストリームからのオンラインマイニング問題の定式化を与え、3節では、それを解くオンラインアルゴリズム StreamT [6] について述べる。4節では実データを用いた実験を行い、忘却型アルゴリズムの有効性を評価する。最後に5節で本稿をまとめる。

1.1 関連研究

半構造データベース [2, 18] の普及にともない、半構造データマイニングの研究が盛んに行なわれるようになってきた [5, 9, 12, 14, 19, 21]。これらの多くは、Apriori [3] 風にパターンの探索を行なっている。それに対して、Zaki [21] と Asai ら [5] は、それぞれ独立に、半構造データからの効率的な順序木パターン探索法（最右拡張）を開発した。これは、Bayardo [7] による集合枚举木を用いたアイテムセット探索法の自然な拡張である。本稿で提案するアルゴリズムは、部分的に最右拡張の概念を用いるものの、上述 [5, 21] の半構造データマイニング手法とはまったく異なることに注意されたい。

最近、オンライン型データマイニングの研究が盛んになってきている [10, 13, 16]。Hidber [11] は、ランザクションデータのストリームに対するオンライン型の相関規則マイニングアルゴリズムを開発した。Parthasarathy ら [15] は、時系列パターンを発見するオンラインマイニングアルゴリズムを与えている。また、Mannila ら [13] は、エピソードと呼ばれる特殊な時系列パターンの発見問題を考察している。Yamanishi ら [20] は、効率的なオンライン異常検出システム SmartSifter を開発した。このシステムはネットワークへの侵入検出などに応用されている。

連絡先: 浅井達哉, 〒812-8581 福岡市東区箱崎 6-10-1, 九州大学大学院システム情報科学府情報理工学専攻, Phone: 092-642-2697, FAX: 092-642-2698, Mail: t-asai@i.kyushu-u.ac.jp

図 1: データ木 D とパターン木 T

2 定式化

2.1 半構造データとラベルつき順序木

本稿では、XML 文書などの半構造データ [2] をラベルつき順序木 [4] でモデル化する。

ラベルつき順序木 $\mathcal{L} = \{\ell, \ell_0, \ell_1, \dots\}$ をラベルの有限集合とする。半構造データベースとパターンの形式的なモデルとして、ラベルつき順序木を用いる。ラベルは半構造データの属性名、タグつきテキストのタグに対応する。順序木とは子供の集合に順序が付けられた木である。形式的には、 $\mathcal{L}$ 上のラベルつき順序木を次の五項組 $T = (V, E, B, L, v_1)$ で定義する。 $G = (V, E, v_1)$ は v_1 を根とする木である。 V は節点の集合を、 $E \subseteq V^2$ は親子関係を表す。 $L: V \rightarrow \mathcal{L}$ はラベル関数を、二項関係 $B \subseteq V^2$ は G における兄弟関係を表す。

ラベルつき順序木 T の大きさを、節点数 $|T| = |V|$ と定義する。節点 $v \in V$ の深さを、根節点 v_1 から v への経路に含まれる辺の数と定義する。 T の節点集合は $V_T = \{1, \dots, k\}$ であり ($k = |T|$)、これらは T の前置順巡回 (preorder traversal) で番号付けられていると仮定する。このとき、 T の根は $v_1 = 1$ であり、一番右の葉 (最右葉) $rml(T)$ は $rml(T) = k$ となることに注意されたい。 T の最右枝を $RMB(T)$ であらわす。

パターンの出現 T と D を $\mathcal{L}$ 上の順序木とする。ここで、 T をパターン木、 D をデータ木と呼ぶことにする。以下の条件を満たす関数 $\varphi: V_T \rightarrow V_D$ が存在するとき、 T が D に出現すると定義する: (i) φ は単射であり、 (ii) φ は親子関係、兄弟関係とラベルの値を保存する。ただし、パターン木 T 中で隣接する兄弟が、それらの出現先で隣接する必要はない。この φ を、 T から D へのマッピング関数と呼ぶ。

以後、パターン木 T を単にパターンと呼ぶ。

D をデータ木、 T を大きさ k のパターン、 φ を T から D へのマッピング関数とする。 φ に関する T の D への埋め込みを、 φ による T の像 $\varphi(T) \subseteq V_D$ と定義する。また、 φ に関するルート出現を $roc(\varphi) = \varphi(1)$ と、最右葉出現を $rmo(\varphi) = \varphi(k)$ と定義する。

例として、図 1 のデータ木 D とパターン T を考える。図中の矢印は、 D への埋め込みが $E_T = \{7, 8, 10\}$ となるようなマッピング関数 φ を表している。 T の D への全ルート出現は節点 2 と節点 7 の 2 種類であり、最右葉出現は節点 4, 6, 10 の 3 種類である。

2.2 ラベルつき順序木のストリーム表現

本節では、ラベルつき順序木の記号列による表現を導入する。この表現は、データストリームから XML 文書が、テキスト順に順次読み出される場合に対応する。

定義 1 $\mathcal{L}$ 上のラベルつき順序木 T の深さ・ラベル対表現とは、記号列 $s(T) = ((d_1, \ell_1), \dots, (d_k, \ell_k))$ のことである。ここに、 $(d_i, \ell_i) \in \mathbb{N} \times \mathcal{L}$ は、 T のプリオーダー順で i 番目の節点 v_i の深さとラベルの組であり、 $k = |T|$ である。

例えば、図 1 のデータ木 D に対する深さ・ラベル対表現は $s(D) = ((0, R), (1, A), (2, A), (2, B), (2, A), (2, B), (1, A), (2, A), (2, A), (2, B))$ となる。次で定義する最右拡張 [5] を用いることにより、任意のラベルつき順序木 T に対して、その深さ・ラベル対表現 $s(T)$ が一意に定まることが分かる。

定義 2 S を大きさ $k-1$ のラベルつき順序木、 $d_{max} \geq 0$ を S の最右葉 $v_{k-1} = k-1$ の深さとする。任意の自然数 $0 \leq d \leq d_{max} + 1$ と任意のラベル $\ell \in \mathcal{L}$ に対して、 S の (d, ℓ) 拡張とは、 S にラベル ℓ をもつ新たな節点 $v_k = k$ を追加して得られるラベルつき順序木 T である。ただし、 S の最右葉 v_{k-1} の先祖で深さが $d-1$ であるような節点を z とするとき、 v_k は z の一番右の子として追加する。

T が S の (d, ℓ) 拡張であるとする。このとき、 T を S の前者といい、 S を T の後者という。

2.3 データマイニング問題

集合 A に対して、 A^∞ は A 上のすべての無限列をあらわす。

D をデータ木とする。 D の深さは有限であると仮定する (幅は無限であってもよい)。 D に対する半構造データストリームとは、無限列 $S = s(D) = (v_1, v_2, \dots, v_i, \dots) \in (\mathbb{N} \times \mathcal{L})^\infty$ のことである。ここに、 $v_i = (d_i, \ell_i)$ は D の i 番目の節点 $v_i = i$ の深さ・ラベル対表現である。このとき、 v_i を S の i 番目の節点と呼ぶことにする。我々は節点 $v_i = i$ と時刻 i を同一視する。

我々のデータマイニング問題を以下に記す。

半構造データストリームからの

オンライン型頻出パターン発見問題

入力: 半構造データストリーム $S = (v_1, v_2, \dots, v_i, \dots)$, 実数 $0 < \sigma \leq 1$.

問題: 時刻 i で v_i が入力されると仮定する ($i \geq 1$)。このとき、出現頻度が σ 以上となるような頻出パターン T を、各時刻 i において出力せよ。

オンライン型マイニングアルゴリズムの目標は、無限のデータストリームに対して有限の資源しか用いずに半永久的に働き、任意の時点で利用者の質問に迅速に応えることである。

パターンの出現頻度として、我々は以下の 2 つのモデルを定義する。ただし、 $1 \leq j \leq i$ に対して、 $hit_j^{(i)}$

Algorithm StreamT($\mathcal{L}, S, \sigma$)

入力: ラベル集合 $\mathcal{L}$, データ木 D に対応する半構造データストリーム $S = s(D) = (v_1, v_2, \dots, v_i, \dots)$, 最小サポート $0 < \sigma \leq 1$.

出力: パターン集合の列 $(\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_i, \dots)$ ($i \geq 1$). ただし, 任意の $j \geq 1$ に対して $\mathcal{F}_j \subseteq T$.

変数: 候補集合 $C \subseteq T$, 掃木枝スタック $B = (B[0], \dots, B[Top])$.

手法:

1. C を大きさ 1 の全パターンの集合とする;
 $B := \emptyset, Top := -1, i := 1$;
2. 次の節点 $v_i = (d, \ell)$ が存在するかぎり, 以下を行なう:
 - 2-(a) 掃木枝スタックのバケット $B[0], \dots, B[d-1]$ を更新する:
 $(B, EXP) := \text{UpdateB}(B, C, (d, \ell), i)$;
 - 2-(b) 候補集合 C とバケット $B[d]$ を更新する:
 $(B, C) := \text{UpdateC}(EXP, B, C, (d, \ell), i)$;
 - 2-(c) $\mathcal{F}_i = \{T \in C \mid \text{freq}_i(T) \geq \sigma\}$ を出力する;
 $i = i + 1$;

図 2: オンライン型半構造データマイニングアルゴリズム

は次のように定義される関数である. もし $v_j \in V_D$ が T の D_i へのルート出現ならば $\text{hit}_j^{(i)}(T) = 1$ であり, それ以外のときは $\text{hit}_j^{(i)}(T) = 0$.

定義 3 T を任意のパターン, S を半構造データストリームとする. このとき, 時刻 i における T の頻度 $\text{freq}_i(T)$ を, 以下のように定義する.

基本型 このモデルでは, 時刻 i における T の D_i への異なるルート出現数 $\text{count}_i(T)$ を数える.

$$\text{freq}_i(T) = \frac{1}{i} \text{count}_i(T) = \frac{1}{i} \sum_{j=1}^i \text{hit}_j^{(i)}(T). \quad (1)$$

忘却型 これは, 過去のデータの影響力を, 過ぎ去った時間に対して指数的に減衰させるモデルである [20]. $0 < \gamma < 1$ を実数 (忘却定数) とし, $Z_i = \sum_{j=1}^i \gamma^{i-j}$ とおく.

$$\text{freq}_{\gamma, i}^{\text{fg}}(T) = \frac{1}{Z_i} \sum_{j=1}^i \gamma^{i-j} \text{hit}_j^{(i)}(T). \quad (2)$$

3 オンラインマイニング手法

本節では, 基本型モデルに対するオンライン半構造データマイニングアルゴリズム StreamT を与える.

3.1 アルゴリズムの概要

我々の提案するアルゴリズム StreamT を, 図 2 に示す. このアルゴリズムは, 各ステージ $i = 1, 2, \dots$ において, 半構造データストリーム S から節点 $v_i = (d, \ell_i)$ を受け取り, 候補パターンの頻度を計算する. アルゴリズムは, 利用者からの要請に応じて, ステージ i で

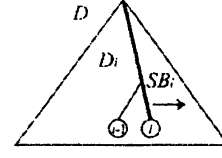


図 3: 時刻 i における掃木枝 SB_i

σ 以上の頻度をもつ頻出パターンの集合 $\mathcal{F}_i \subseteq T$ を出力する.

データストリームから継続的に頻出パターンを計算するために, アルゴリズムはデータ木 D の上で掃木枝 (sweep branch) と呼ばれる D の根 v_1 から現在の節点 v_i への経路を右へずらしていき, 掃木枝と交わるすべてのパターンを探索する. これを実現するために, アルゴリズムはステージごとに以下の情報を更新する:

- 候補集合 $C \subseteq T$.
- 掃木枝スタック $B = (B[0], B[1], \dots, B[Top])$.

任意の候補パターン $T \in C$ は, T の D_i 中のルート出現数 $\text{count}(T)$ をもつ. また, 深さ r の最新のルート出現 $Rto_T[r]$ を, 深さごとに保持する.

3.2 掃木枝を用いた漸増的なパターン発見

候補パターンの出現位置を見つけるために, パターンの埋め込みを完全に保持する必要はない. アルゴリズムは, 掃木枝を D 上で右へ走らせて, パターン T の埋め込み $\varphi(T)$ と掃木枝の交差をステージごとに更新しながら保持し, 漸増的に候補パターンの出現を計算する. パターン T の埋め込みと掃木枝の交差に関して, 次の性質が成り立つ.

補題 1 $\varphi(T)$ をパターン T の任意の埋め込み, SB_i を時刻 i の掃木枝とする. このとき, $\varphi(T)$ の最右枝と SB_i との交差 $I_T = \text{RMB}(\varphi(T)) \cap SB_i$ が空集合でなければ, I_T は D の経路となり, $\varphi(T)$ の根を含む. $\square$

この補題にしたがい, 次の定義を与える. パターン T の埋め込み $\varphi(T)$ と掃木枝 SB_i に対して, $\varphi(T)$ と SB_i の交差を $I_T = \text{RMB}(\varphi(T)) \cap SB_i$ とする (図 4). このとき, I_T 中でもっとも浅い節点 v_r と深い節点 v_b を, それぞれ SB_i に関する I_T のルート出現, ボトム出現と呼ぶ. I_T のルート出現は φ に関する T のルート出現と一致することに注意せよ. また, $\text{RMB}(\varphi(T)) \subseteq SB_i$ が成り立つときは, I_T のボトム出現と φ に関する T の最右葉出現は一致する.

次の補題は, 最右葉出現の漸増的な特徴づけを与える.

補題 2 T を大きさ $k > 1$ のパターンとする. 時刻 i において, T が現在の節点 v_i を最右葉出現にもつことの必要十分条件は, 次の (i), (ii) を満たす大きさ $k-1$ のパターン S が存在することである. (i) S はボトム出現 v_b として v_i の親節点をもち, (ii) T は S の $(d-r, \ell)$

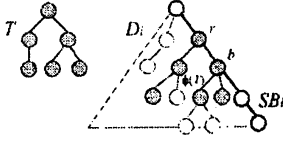


図 4: SB_i に関する T のルート出現 r とボトム出現 b

拡張である。ここに、 d は v_i の深さ、 r は S のルート出現 v_r の深さ、 ℓ は v_i のラベルである。□

この補題により、候補パターン $T \in \mathcal{C}$ のすべての最右葉出現を、 T の後者 S のボトム出現を用いて計算できることが分かる。我々のアルゴリズム StreamT は、掃木枝スタック $B = (B[0], B[1], \dots, B[Top])$ を管理することにより、パターンの埋め込みと掃木枝の交差を記録する。ここで、掃木枝スタックの任意の要素（バケットと呼ぶ） $B[b]$ は、 (T, r, b) の形の三項組を保持する。 $T \in \mathcal{C}$ は候補パターン、 r と b は、 T の SB_i に関するルート出現 v_r とボトム出現 v_b の深さである。

掃木枝スタック $B = (B[0], B[1], \dots, B[Top])$ が以下の条件 (i), (ii) を満足するとき、 B は時刻 i で up-to-date であるという。(i) スタック B の長さが掃木枝 SB_i の長さと同じ、すなわち $Top = |SB_i|$ が成り立ち、(ii) 任意の $0 \leq h \leq Top$ に対して、 B の h 番目のバケット $B[h]$ は、三項組 $\tau = (T, r, b) \in T \times N \times N$ のうちでボトム出現が $b = h$ となるものをすべて含む。

次の補題は、時刻 i における掃木枝スタック B_i の各バケットに関する再帰的な特徴づけを与える。

補題 3 $v_i = (d, \ell)$ をデータストリーム S における現在の節点、 $B_k = (B_k[0], B_k[1], \dots, B_k[Top_k])$ を時刻 $k \in \{i-1, i\}$ における掃木枝スタックとする。このとき、掃木枝スタック B_{i-1} と B_i が共に up-to-date であると仮定すると、次の 1-4 が成り立つ。

1. 任意の $0 \leq b < d-1$ に対して、 $\tau \in B_i[b]$ の必要十分条件は $\tau \in B_{i-1}[b]$ である。
2. $b = d-1$ のとき、 $\tau = (T, r, d-1) \in B_i[d-1]$ の必要十分条件は、以下の (i), (ii) のどちらかが成り立つことである。
 - 2-(i) $\tau \in B_{i-1}[d-1]$ 。
 - 2-(ii) $r \leq d \leq h$ なる整数 h が存在し、 $(T, r, h) \in B_{i-1}[d] \cup \dots \cup B_{i-1}[Top_{i-1}]$ を満たす。
3. $b = d$ のとき、 $\tau = (T, r, d) \in B_i[d]$ の必要十分条件は、以下の条件 (i), (ii) のどちらかが成り立つことである。
 - 3-(i) T はラベル ℓ をもつ大きさ 1 の順序木である。
 - 3-(ii) 三項組 $(S, r, d-1) \in B_i[d-1]$ が存在し、 T は S の $(d-r, \ell)$ 拡張である。
4. 任意の $b > d$ に対して、 $B_i[b]$ は未定義である。□

図 5 に、補題 3 に基づいて掃木枝スタック B_{i-1} を更新する様子を示す。

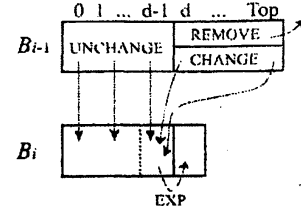


図 5: 時刻 $i-1$ と時刻 i における掃木枝スタック

Algorithm UpdateB($B, C, (d, \ell), i$)

入力: 掃木枝スタック $B = (B[0], B[1], \dots, B[Top])$, 候補集合 C , 節点 $v_i = (d, \ell)$, 現在の時刻 i 。

出力: 三項組の集合 $EXP \subseteq T \times N \times N$ 。

手法:

1. もし $d \leq Top$ ならば、以下を行なう:
 - $BELOW := B[d] \cup \dots \cup B[Top]$; $Top = d-1$;
 - $REMOVE := \{(T, r, b) \in BELOW \mid r \geq d\}$;
 - $CHANGE := \{(T, r, b) \in BELOW \mid r \leq d-1\}$;
 - $B[d-1] := B[d-1] \cup \{(T, r, d-1) \mid (T, r, b) \in CHANGE\}$;
2. ラベル ℓ の 1-パターン T_ℓ に対して、 $EXP := \{(T_\ell, d, d)\}$;
3. 任意の $(S, r, d-1) \in B[d-1]$ に対して以下を行なう:
 - S の $(d-r, \ell)$ 拡張 T に対して、 $EXP := EXP \cup \{(T, r, d)\}$;
4. (B, EXP) を出力する;

図 6: 掃木枝スタック更新アルゴリズム

これに基づき、掃木枝スタックを再帰的に更新するアルゴリズム UpdateB を、図 6 に示す。このアルゴリズムは、任意の時刻 i において、古い掃木枝スタック B_{i-1} と現在の節点 $v_i = (d, \ell)$ を受け取り、新しい掃木枝スタック B_i の中で最初の d 個のバケット $B_i[0], \dots, B_i[d-1]$ を出力する。

次の系は、補題 3 よりただちに示される。

系 4 図 2 のアルゴリズム StreamT における while ループが時刻 i にて発動したと仮定する。このとき、アルゴリズム UpdateB($B, C, (d, \ell), i$) は、以下の (i), (ii) を出力する。(i) d 番目のバケット $B[d]$ を除くすべてのバケット $B[0], \dots, B[d-1]$ が時刻 i で up-to-date であるような掃木枝スタック B , (ii) 時刻 i で up-to-date であるような d 番目のバケット $B[d]$ 。□

3.3 候補集合の管理

本節では、候補集合 C の管理法について述べる。

図 7 に、候補集合 C を管理するアルゴリズム UpdateC を示す。我々のアルゴリズムは、Hidber [11] と同様に、候補パターンとして頻出パターン集合の *negative border* [15] (後者パターンが頻出となるような非頻出

Algorithm UpdateC($EXP, B, C, (d, \ell), i$)

入力: 三項組の集合 EXP , 掃木枝スタック $B = (B[0], B[1], \dots, B[Top])$, 候補集合 C , i 番目の節点 $v_i = (d, \ell)$, 現在の時刻 i .

出力: 更新後の掃木枝スタックと候補集合 (B, C) .

手法:

1. 任意の三項組 $(T, r, b) \in EXP$ に対して, そのルート出現 ρ が $\rho \neq Rto_T[r]$ を満たすならば以下を行なう:
 - $Rto_T[r] := \rho$;
 - $T \in C$ ならば, $count(T) := count(T) + 1$;
 - $T \notin C$ かつ T の後者が頻出ならば, $count(T) := c_i$ かつ $C := C \cup \{T\}$;
 2. $B[d] := \emptyset$; $Top := d$; $freq_{\gamma, i}(T) := count(T)/i$;
 3. 非頻出な後者をもつパターン $T \in C$ に対して, 以下を行なう:
 - $C = C \setminus \{T\}$;
 4. 任意の三項組 $(T, r, b) \in EXP$ に対して, 以下を行なう:
 - $T \in C$ ならば $B[d] := B[d] \cup \{(T, r, b)\}$;
 5. (B, C) を出力する;
-

図 7: 候補集合の更新アルゴリズム

パターン) までを保持する. 候補集合の管理は, 以下の 4 つの手順で行われる.

初期化 大きさ 1 の全パターンを候補集合 C に入れる. これは StreamT (図 2) の最初に行なわれる.

インクリメント 現在の節点 v_i が最右葉出現となる候補パターン $T \in C$ に対して, $count(T)$ を 1 増やす.

挿入 パターン $T \notin C$ がストリームに出現したとき, その後者 $P \in C$ が頻出ならば, C に T を加える.

削除 頻出な候補パターン $T \in C$ が非頻出になったら, T のすべての前者 (最右拡張) を C から削除する. ただし, 大きさ 1 のパターンは削除しないものとする.

3.4 忘却モデルへの拡張

本節では, アルゴリズム StreamT を, 定義 3 の忘却モデルに適用する方法を示す.

忘却モデルにおけるパターン T の頻度は, 定義 3 の (2) 式で定義される. すなわち, 忘却モデルでは, 時刻 i における頻度は各々のルート出現が発見された時刻 j に関するすべての重み γ^{i-j} に依存する.

$freq_{\gamma, i}^{fg}(T)$ と $hit_j^{(i)}(T)$ を, それぞれ fr_i , hit_j と略記する. また, 時刻 i におけるパターン T が最後に出現した時間を $lt(i)$ とする. このとき, 次の補題が成り立つ.

補題 5 任意のパターン T と時刻 i に対して, 以下の漸化式が成り立つ.

$$\begin{aligned} fr_0 &= 0, \\ fr_i &= \frac{lt(i)}{i} \gamma^{i-lt(i)} fr_{lt(i)} + \frac{1}{i} hit_i \quad (i > 0) \end{aligned}$$

□

この補題から, $lt(i)$ を保持することにより, 忘却モ

デルにおける T の頻度 $freq_{\gamma, i}^{fg}(T)$ を効率よく計算できることが分かる.

3.5 アルゴリズムの解析

ステージ i におけるアルゴリズム StreamT の計算量は以下のとおりである.

時刻 i における掃木枝スタックを B_i とする. このとき, ステージ i におけるアルゴリズム UpdateB と UpdateC の時間計算量は, それぞれ $O(\sum_{j=d-1}^{Top} |B_{i-1}[j]|)$ 時間, $O(kC + D)$ 時間となる. ここに, d は時刻 i の入力節点 v_i の深さ, k はパターンの最大サイズ, C は v_i に最右葉出現をもつ候補パターン数, D はステージ i で削除された候補パターンの数である. また, ステージ i における領域計算量は, $O(\sum_{j=1}^{Top} |B_i[j]| + k|C|)$ 領域である.

したがって, 我々のアルゴリズムは, 掃木枝スタックと候補集合の大きさ程度の時間と領域しか使用しない.

4 実験

我々は, 基本型と忘却型のアルゴリズムを SAX と Java で実装し, PC 上で実験を行った. 実験には, 以下のスペックのマシンを用いた.

- PentiumIII 1GHz, 1500 megabytes RAM, Windows2000 (4.2 節 規模耐性実験)
- PentiumIII 1.2GHz, 1000 megabytes RAM, Windows2000 (4.3 節 アルゴリズムの比較実験)

4.1 データ

我々は, 実験データとして, 2 種類のデータセット *dblp* と *wsu* を用意した.

dblp は計算機科学分野の論文目録サイト DBLP [8] で公開されている XML データ *dblp.xml* (130MB) であり, 対応する半構造データストリームの節点数は 3,185,138, 異なるタグ数は 22 である. このデータは, 4.2 節の規模耐性実験に用いる.

もう 1 つのデータセット *wsu* は, [17] で提供されている XML データ $w = \text{weblog.xml}$ と $s = \text{soap.xml}$ を, w, s, w の順に結合して得られた節点数 22,830 のデータである. w と s に含まれるタグの種類は, それぞれ 13 種類と 11 種類であり, 双方に共通するタグは存在しない. このように, データセット *wsu* は, 時間変化する半構造データストリームを意図して人工的に作成されたデータである. このデータは, 4.3 節のアルゴリズムの比較実験に使用する.

4.2 規模耐性

最初に, 大規模な XML データを用いた規模耐性実験を行い, アルゴリズム StreamT の性能を評価する.

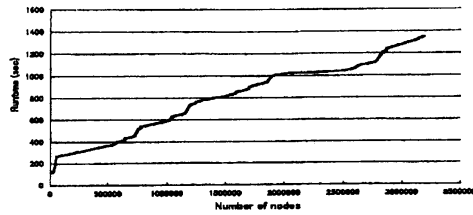


図 8: 規模耐性実験

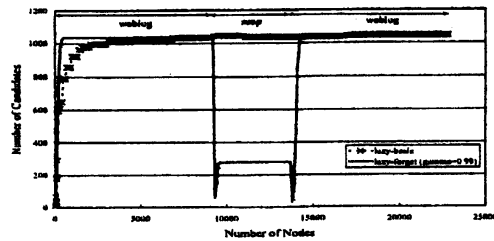


図 9: 基本型 vs 忘却型

この実験では、dblp データに対して最小サポートを $\sigma = 1(\%)$ に固定し、データストリームの節点が 2000 個処理されるごとにアルゴリズムの計算時間を測定した。実験結果を図 8 に示す。

実験より、アルゴリズム StreamT は、入力サイズに対してほぼ線形時間で動作する。この結果は、提案するオンラインアルゴリズムが、高速に流れ続ける大規模な半構造データストリームに対して、効率よく働き続けることを示している。

4.3 アルゴリズムの比較

本節では、基本型と忘却型のアルゴリズムを実験により比較し、忘却の有無がアルゴリズムの実際の振る舞いに与える影響を検証する。本節の実験では、データセット *wsu* を用いた。

図 9 に、最小サポートを $\sigma = 5(\%)$ に固定して、基本型アルゴリズムと、忘却定数 $\gamma = 0.99$ に対する忘却型アルゴリズムが、各時刻にそれぞれ計算した候補パターン数を示す。

この結果から、忘却型のアルゴリズムは、データストリームの傾向が変わると、速やかに過去のパターンを削除し、代わりに現在のパターンを候補集合に加えることが分かる。これは、忘却型のアルゴリズムが、過去に出現したパターンの出現頻度を指数的に減減させるからである。

このように、忘却型アルゴリズムは、基本型アルゴリズムよりストリームデータの時間変化に柔軟に対応する。したがって、忘却の手法は、異常検出やトレンド変化検出などの応用に有効である。

5 おわりに

本稿では、半構造データストリームからのオンラインデータマイニングを考察し、効率よいオンライン半構造データマイニングアルゴリズム StreamT を与えた。アルゴリズムに忘却の概念を導入した結果、過去の影響を受けずにデータの時間変化に柔軟に追従することを実証した。

今後は、半構造データストリームからの異常検出やトレンド発見などに、提案手法を応用していきたい。

参考文献

- [1] K. Abe, S. Kawasoe, T. Asai, H. Arimura, and S. Arikawa. Optimized Substructure Discovery for Semi-structured Data, In *Proc. PKDD'02*, 1-14, LNAI 2431, Springer, 2002.
- [2] S. Abiteboul, P. Buneman, D. Suciu, *Data on the Web*, Morgan Kaufmann, 2000.
- [3] R. Agrawal, R. Srikant, Fast Algorithms for Mining Association Rules, In *Proc. the 20th VLDB*, 487-499, 1994.
- [4] Aho, A. V., Hopcroft, J. E., Ullman, J. D., *Data Structures and Algorithms*, Addison-Wesley, 1983.
- [5] T. Asai, K. Abe, S. Kawasoe, H. Arimura, H. Sakamoto, S. Arikawa, Efficient Substructure Discovery from Large Semi-structured Data, In *Proc. the 2nd SIAM Int'l Conf. on Data Mining (SDM2002)*, 158-174, 2002.
- [6] T. Asai, H. Arimura, K. Abe, S. Kawasoe, and S. Arikawa. Online Algorithms for Mining Semi-structured Data Stream, In *Proc. the 2002 IEEE Int'l Conf. on Data Mining (ICDM'02)*, 27-34, 2002.
- [7] R. J. Bayardo Jr., Efficiently Mining Long Patterns from Databases, In *Proc. SIGMOD98*, 85-93, 1998.
- [8] DBLP, <http://dblp.uni-trier.de/>
- [9] L. Dehaspe, H. Toivonen, R. D. King, Finding Frequent Substructures in Chemical Compounds, In *Proc. KDD-98*, 30-36, 1998.
- [10] P. B. Gibbons, Y. Matias, Synopsis Data Structures for Massive Data Sets, In *External Memory Algorithms, DIMACS Series in Discr. Math. and Theor. Comput. Sci.*, Vol. 50, AMS, 39-70, 2000.
- [11] C. Hidber, Online Association Rule Mining, In *Proc. SIGMOD'99*, 145-156, 1999.
- [12] A. Inokuchi, T. Washio, H. Motoda, An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data, In *Proc. PKDD 2000*, 13-23, LNAI, Springer, 2000.
- [13] H. Mannila, H. Toivonen, A. I. Verkamo, Discovering Frequent Episode in Sequences, In *Proc. KDD-95*, 210-215, 1995.
- [14] T. Miyahara, Y. Suzuki, T. Shoudai, T. Uchida, K. Takahashi, H. Ueda, Discovery of Frequent Tag Tree Patterns in Semistructured Web Documents, In *Proc. PAKDD-2002*, 341-355, 2002.
- [15] S. Parthasarathy, M. J. Zaki, M. Ogihara, S. Dworkadas, Incremental and Interactive Sequence Mining, In *Proc. CIKM'99*, 251-258, 1999.
- [16] R. Rastogi, Single-Path Algorithms for Querying and Mining Data Streams, In *Proc. SDM2002 Workshop HDM'02*, 43-48, 2002.
- [17] Sosnoski Software Solution, Inc., XML benchmark, <http://www.sosnoski.com/opensrc/xmlbench/>
- [18] W3C, Extensive Markup Language (XML) 1.0 (Second Edition), *W3C Recommendation*, 06 October 2000. <http://www.w3.org/TR/REC-xml>
- [19] K. Wang, H. Liu, Discovering Structural Association of Semistructured Data, *TKDE*, 12(2), 353-371, 2000.
- [20] K. Yamanishi, J. Takeuchi, A Unifying Framework for Detecting Outliers and Change Points from Non-Stationary Time Series Data, In *Proc. SIGKDD 2002*, ACM, 2002.
- [21] M. J. Zaki. Efficiently Mining Frequent Trees in a Forest, In *Proc. SIGKDD 2002*, ACM, 2002.