

# データストリームを対象とした動的多重集合に対する Min-hash の高速計算アルゴリズム

三原 寛寿<sup>†</sup> 古賀 久志<sup>†</sup>

<sup>†</sup> 電気通信大学情報理工学研究科 〒182-8585 東京都調布市調布ヶ丘1丁目5-1

E-mail: <sup>†</sup>{mihara,koga}@sd.is.uec.ac.jp

**あらまし** 近年, IoT や SNS の発展に伴いストリームデータが取り扱われる機会が増え, ストリームデータを対象とする類似検索の重要性も増している. この類似検索ではストリームデータを要素が動的に変わる集合と見なし, 集合間類似検索により類似ストリームデータを探す. 集合間類似度としては Jaccard 係数がよく用いられる. しかし, 集合が変わる度に Jaccard 係数を計算し直すのはオーバーヘッドが大きい. 各集合に対するコンパクトなスケッチを Min-Hash というハッシュ関数により生成し, スケッチ間で Jaccard 係数を近似計算する手法が提案されている. 本研究ではスケッチの更新を効率化するために, 動的に変化する集合に対する Min-Hash のハッシュ値更新方法について論じる. 既存研究では (1) 要素の削除と (2) 多重集合の両方を取り扱える手法がほとんど存在しない. 本研究では, スライディングウィンドウモデル (に基づく要素削除) の条件下で集合を取り扱った Datar らの手法を多重集合に拡張する. 提案手法は, スライディングウィンドウモデルで多重集合を取り扱える初めての手法となる.

**キーワード** Streaming algorithms, Sketch, Min-hash

## 1 はじめに

ストリームデータとは時間経過と共に継続的に次々と生成されるデータのことを言う. 近年, IoT や SNS の発展に伴いストリームデータが取り扱われる機会が増加している. 例えば, IoT におけるセンサからの観測データはストリームデータである. また, 特定ユーザの twitter におけるツイートやウェブページの閲覧履歴も時間と共に新データが追加されるという点でストリームデータである. こうしてストリームデータの増加に連れて, ストリームデータを対象とした類似検索も重要になっている. 例えば, 過去の異常パターンとの類似性に基づいた異常検知 [7] や, SNS のコンテンツが似たユーザを見つけて類似ユーザの挙動からアイテムを推薦するユーザベースの情報推薦 [8] などはストリームデータを対象とした類似検索に帰着して解ける. 後者の例では, ユーザ  $u$  の SNS への投稿内容を時間経過に伴ってデータが増えるストリームデータと見なし, ストリームデータを対象とした類似検索によって  $u$  と嗜好性が似た類似ユーザを見つける. 最近のストリームデータを対象とした類似検索では, ストリームデータを生成されたデータの集合として表現し集合間類似検索によって類似ストリームデータを探すアプローチが主流である. 通常の集合間類似検索と比べると, 新たなデータの生成により集合の要素が動的に変化するため類似検索結果を更新する必要がある点が異なる. 2つの集合  $A, B$  に対する類似度  $\text{sim}(A, B)$  としては Jaccard 係数がよく用いられるが,  $A$  や  $B$  が変化する度に Jaccard 係数を計算するオーバーヘッドは大きい. そこで Min-Hash [10] を用いて  $A$  と  $B$  のコンパクトなスケッチ  $ms_A, ms_B$  を生成し, スケッチ間で Jaccard 係数を近似計算する手法がいろいろ提案されて

いる. これらの手法はいずれも集合の変化に対するスケッチ更新を効率化するが,

- データ削除を取り扱えるか
- 同種類の要素を複数持つ多重集合に対して拡張 Jaccard 係数の近似値を計算できるか

という2点で機能的に異なる. データ削除に関してはストリームデータ内では新しいデータほど最新の状況を反映して価値が高いところから, 古いデータを軽視するモデルが2種類存在する.

(1) スライディングウィンドウモデル: データストリームの直近  $W$  個要素をスライディングウィンドウと定義し, 時刻が進むとウィンドウに到着データを追加し, ウィンドウ内の最古データを廃棄する.

(2) 減衰モデル: データストリームを要素に重みが付与された重み付き集合として扱い, 時間経過に伴って古いデータの重みを減衰する.

MaxLogHash [4] は Min-Hash を省メモリ化する  $b$ -bit Min-hash [5] をデータ追加時に更新可能にしたがデータ削除を取り扱えない. Datar ら [1] はスライディングウィンドウモデルでデータ削除に対してもハッシュ値を更新できるアルゴリズムを考案した. さらにスケッチ更新のために保持しないといけないスライディングウィンドウ内の要素数が  $O(\log W)$  となることを証明した. しかし, Datar らの手法は多重集合を取り扱えない. Bury らは [9] はデータ削除が任意の順序で発生してもハッシュ値を更新できる手法を構築したが, やはり多重集合は取り扱えない. Histosketch [3] は多重集合に関してデータ削除を取り扱える唯一の手法であるが, 減衰モデルを想定している. したがって, スライディングウィンドウモデルで多重集合を取り扱える手法は存在しない. そこで本研究では, スライディング

ウィンドウに対して多重集合を取り扱える手法を実現することを目的とし、Datar らの手法を多重集合を取り扱えるように拡張する。多重集合の場合、スライディングウィンドウ内の同種類の要素数に依って、同一要素に割り当てられるハッシュ値が変化するという難題があるが、本研究では Datar らの手法をハッシュ値の変化に対処できるように修正した。以下、本稿の構成を述べる。

2 節で提案手法の要素技術となる多重集合、集合間類似度、Min-Hash の概念を説明する。3 節では従来手法となる Datar らの動的に変化する集合を対象としたハッシュ値更新アルゴリズムを紹介する。4 節で提案手法となる動的に変化する多重集合に対するハッシュ値更新アルゴリズムを述べる。5 節で 4 節のアルゴリズムを複数個の要素が入り出すウィンドウに対応するように拡張したアルゴリズムを提案する。6 節で提案手法を人工データ、実データを用いて評価する。提案手法がハッシュ値を完全に再計算するベースライン手法よりも Min-Hash スケッチを高速に更新できることを示す。7 章でまとめと今後の課題を述べる。

## 2 準備

### 2.1 多重集合

一般にオブジェクトの集まりを集合と呼ぶ。例えば  $\{a, b, c\}$  はアルファベットを要素とする集合である。一般的には要素群を表すアルファベット  $\phi$  に対して、その要素を 0 個以上含むものが集合となる。通常、集合は同じ種類の要素を 1 つしか含まない。集合を同じ種類の要素を複数持てるようにしたもの多重集合という。つまり、 $\{a, a, b, c, d, d, d\}$  というような集合である。

### 2.2 Min-hash

集合間で類似度を計算する方法の 1 つとして、Min-hash がある。Min-hash は、集合に対する確率的なハッシュ関数であり、Jaccard 係数を用いた集合間類似検索を高速化するための技術である [10]。Min-hash は計算されたハッシュ値が一致する確率は Jaccard 係数と一致するという性質を持ち、類似集合ほどハッシュ値が一致しやすいという良い性質を持つ (式 (1))。

$$\Pr[h(A) = h(B)] = \text{sim}(A, B) \quad (1)$$

式 (1) で  $h$  はハッシュ関数であり、 $A, B$  は集合である。

次に、Min-hash によるハッシュ値の計算方法を紹介する。 $\phi = \{x_1, x_2, \dots, x_{|\phi|}\}$  をアルファベット集合とする。ハッシュ関数は  $\phi$  の各アルファベットに対して、 $\{1, 2, \dots, |\phi|\}$  の中から被らないようにランダムな値を割り当てることで決定される。1 つのある集合の中のアルファベットを見て、その中の要素に対応する割り当て値の中から最小値を選ぶ。この最小値が Min-hash によるハッシュ値となる (式 (2))。

$$h(A) = \min_{e \in A} \pi(e). \quad (2)$$

### 2.3 多重集合に対する Min-hash

多重集合に対して Min-hash のハッシュ値を計算する手法を紹介する。文献では、大きく 2 つの方法が知られている。

- (1) 多重集合内の複数個のラベルに異なる値を割り当てる方法
- (2) Consistent Weighted Sampling (CWS) [11]

CWS は、Min-hash の出力の確率分布を模擬して、サンプリングすることでハッシュ値を計算する手法である。明示的に割り当て値を準備する必要がなく、多重集合が整数でなくても適用可能であるという利点がある。一方で確率の重み理論に精通していないと拡張が難しい手法でもある。(1) の同じラベルのアルファベットに異なる値を割り当てるやり方は多重集合の重みが整数であるという制約を受けるが、拡張は比較的容易である。本論文では (1) の多重集合内の同一ラベルの要素に異なる値を割り当てる手法を使用する。

(1) のハッシュ値計算手法を例を用いて説明する。まず、図 2.1 のように、アルファベット  $\{a, b, c, d\}$  に多重度が 2 である時の数値をランダムに割り当てる。

	a	b	c	d
1 個目	3	4	7	6
2 個目	5	1	8	2

図 2.1 多重集合のハッシュ値表

多重集合  $A = \{a, b, b, c, c, d, d\}$  とする。多重集合  $A$  に対して、図 2.1 の割り当て表を用いて数値を割り当てる。そして、その中の最小値が Min-Hash によるハッシュ値となる (図 2.2)。

集合Aの要素 割り当てた値	$a_1$	$b_1$	$c_1$	$c_2$	$d_1$	$d_2$
	3	4	7	8	6	2

↓

ハッシュ値 :  $h(A) = 2$

図 2.2 多重集合のハッシュ値計算

## 3 Datar らのアルゴリズム

本章では、Datar ら [1] によって提案されたデータストリームに対するハッシュ値更新アルゴリズムを記述する。本論文での提案手法は、このアルゴリズムを基に多重集合へ拡張したものである。データストリームとは、毎時刻要素  $e$  が 1 つ到着するデータである。Datar らの手法はデータストリームのスライディングウィンドウモデルに従って集合が変化することを仮定している。

### 3.1 スライディングウィンドウモデル

時刻  $t$  における集合を  $A_t = \{e_{t-w+1}, e_{t-w+2}, \dots, e_t\}$  とする。 $w$  はスライディングウィンドウの幅であり、 $A_t$  は  $w$  個の要素で構成される。図 3.1 では、 $w = 4$  の時に 3 つの時刻  $t, t+1, t+2$  に対する集合  $A_t, A_{t+1}, A_{t+2}$  を図示している。そして、時刻によってスライディングウィンドウは変化していく。ウィンドウサ

イズ  $w = 4$  とすると、時刻  $t$  では、 $A_t = \{e_{t-3}, e_{t-2}, e_{t-1}, e_t\}$  であり、時刻  $t+1$  では、 $A_{t+1} = \{e_{t-2}, e_{t-1}, e_t, e_{t+1}\}$  となる。

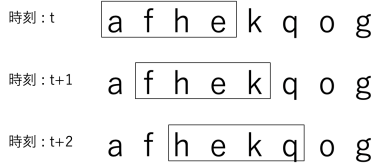


図 3.1 スライディングウィンドウ

動的に変化する集合に対してハッシュ値を計算する自明な手法は、時刻経過によりウィンドウがスライドするたびにハッシュ値を完全に再計算するやり方である。これはすなわち、時刻が  $t$  から  $t+1$  に変化した際に、 $h(A_{t+1})$  を  $A_t$  とは無関係に計算するということである。この手法では、毎時刻ウィンドウ内の全要素をスキャンする必要があるので、 $h(A_{t+1})$  を求めるための時間計算量は  $O(w)$  となる。

しかし、 $A_t$  と  $A_{t+1}$  は  $e_t$  と  $e_{t+w}$  以外の  $w-1$  個の要素が共通であり、 $h(A_t)$  計算時に判明した情報を再利用することで  $h(A_{t+1})$  を計算するオーバーヘッドを減らせる可能性がある。

### 3.2 Datar らの手法

Datar[1] らの手法の基本アイデアは

- 将来に割り当て値が最小になる可能性が絶対ない要素をあらかじめ削除することで、次の時刻のハッシュ値  $h(A_{t+1})$  を計算する時に参照する要素数を減らす

というものである。ここで割り当て値が最小になる可能性が絶対ない要素とは、「ウィンドウ内で自分より後方に割り当て値が小さい要素が存在する要素」のことである。例えば、 $A_t$  内の 2 要素  $e_i$  と  $e_j$  ( $i < j$ ) の割り当て値が  $\pi(e_i) > \pi(e_j)$  を満たすとしよう。この時、 $e_i$  は自分より後方に割り当て値が小さい  $e_j$  が存在するという条件を満足する。この条件の下では  $e_i$  がウィンドウ内に存在する期間は  $e_j$  も必ずスライディングウィンドウ内に存在するので、 $e_i$  の割り当て値が最小になることはありえない。Datar の手法ではこの性質を利用して、スライディングウィンドウから最小値になりえない要素を削除し、将来最小値になりうる要素のみを Minlist というリストで管理する。 $A_t$  のハッシュ値は Minlist に残った要素の割り当て値の最小値となる。任意の Minlist 内の要素  $e$  に対して、 $e$  より後方に割り当て値  $\pi(e)$  より小さい要素は存在しないので、Minlist 内の要素群に対する割り当て値は単調増加となる。このため、 $h(A_t)$  は実は Minlist の先頭要素の割り当て値と等しい。

時刻が  $t$  から  $t+1$  に変化した時に Minlist を更新する手順を以下にまとめる。

- $e_{t-w+1}$  がウィンドウから離脱した時の処理  
 $e_{t-w+1}$  が Minlist に含まれる時には  $e_{t-w+1}$  を Minlist から削除する。この時、 $h(A_t) = \pi(e_{t-w+1})$  であるため、ハッシュ値を更新する必要がある。そのため、 $h(A_{t+1})$  を新たに Minlist の先頭になった要素の割り当て値に修正する。
- $e_{t+1}$  をウィンドウに追加した時の処理

Minlist 内で割り当て値が  $\pi(e_{t+1})$  より大きい要素は、今後、割り当て値が最小になりえないので削除する。 $e_{t+1}$  により Minlist 内の他の要素がすべて削除された結果、Minlist に  $e_{t+1}$  のみ残った場合は  $h(A_{t+1}) = \pi(e_{t+1})$  とする。

このアルゴリズムの動きを図 3.2 に例示する。

時刻  $t+1$  の時、スライディングウィンドウから一番前の要素  $e_{t-w+1}$  を削除し、新しい要素  $e_{t+1}$  のアルファベット  $k$  が入ってくる。 $k$  の割り当て値は 3 であるので、Minlist の中から 3 より大きい割り当て値を持つ要素を削除し、候補リストの一番後ろに  $k$  の割り当て値 3 を加える。



図 3.2 候補リストの作成

このように最小値の候補リスト Minlist のみを保持することで、元のスライディング全体を保持しない。ウィンドウサイズ  $W$  の場合、候補リストは  $\log(W)$  で保持することができ、Minlist 内で  $e_t$  より大きい要素と小さい要素の境界を 2 分探索で探すことにより実行時間は  $O(\log \log W)$  となる。

## 4 多重集合に対するハッシュ値更新アルゴリズム

本稿では、Datar らのアルゴリズムを動的に変化する多重集合に対して拡張したハッシュ値更新アルゴリズム SWMH (Sliding-Window Min-Hash) を提案する。Datar ら [1] のアルゴリズムを多重集合に拡張する場合、スライディングウィンドウ内の要素  $e$  への割り当て値が  $e$  と同じラベルを持つウィンドウ内の要素数によって変化する点が難しくなる。つまり、 $e$  への割り当て値が時間経過によって変化することが起きる。スライディングウィンドウ内の要素  $e$  のラベルをアルファベット  $l(e)$  とする。割り当て値  $\pi(e)$  はスライディングウィンドウ内に  $l(e)$  が何個あるかによって変わる。例えば、図 4.1 の場合だと、時刻  $t=0$  の時は、スライディングウィンドウ内にアルファベット  $b$  は 2 つあり、 $\pi(e_i) = 4, \pi(e_j) = 1$  となる。時刻  $t=3$  でスライディングウィンドウから要素  $e_i$  が抜けた場合、 $e_j$  への割り当て値  $\pi(e_j)$  は 1 から 4 に変化する。

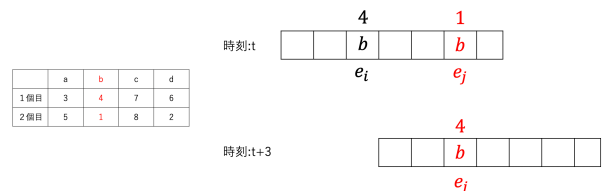


図 4.1  $\pi(e)$  の変化

Datar らのアルゴリズムでは最小値になり得ない要素を Min-

list から削除する。多重集合の場合、割り当て値が変化することを考慮して最小値になり得るかどうかを判定をする必要がある。提案手法を基盤とする要素技術は以下の2つであり、それぞれを順に説明する。

- (1) 同種アルファベットへの値の割り当て
- (2) スライディングウィンドウ更新時の処理

#### 4.1 同種アルファベットへの値の割り当て

多重集合に対する Min-Hash では、多重集合内に同種アルファベットが複数存在する場合それらを区別して値を割り当てる。例えば、ラベルが  $a$  のアルファベットが  $n$  個存在する場合、それらを  $\{a_1, a_2, \dots, a_n\}$  のように区別する。 $a_i$  ( $1 \leq i \leq n$ ) は  $i$  番目の  $a$  という意味であり、各  $a_i$  ( $1 \leq i \leq n$ ) に異なる値  $\pi(a_i)$  を割り当てる。

ここで  $\pi(a_i) \leq \pi(a_{i+1})$  ならば、 $a_{i+1}$  の割り当て値は最小には絶対ならない。その理由は  $i+1$  番目の  $a_{i+1}$  が存在する条件下では、 $i$  番目の  $a_i$  も必ず存在するからである。このように割り当て値  $\pi(a_{i+1})$  は Min Hash のハッシュ値に影響を与えないので

$$\pi(a_i) \leq \pi(a_{i+1})$$

を条件を満足する限りは別の値に変更しても構わない。そこで、 $\pi(a_i) \leq \pi(a_{i+1})$  が成立する場合は  $\pi(a_{i+1})$  を  $\pi(a_i)$  に修正する。修正前と修正後を区別するため、修正前の割り当て値を  $\pi(a_{i+1})$  とし、修正後の割り当て値を  $\pi'(a_{i+1})$  とする。 $\pi'$  は  $i$  に対して増加することがなく、広義単調減少である。

**Lemma 1.**

$$\begin{aligned} & \min\{\pi(a_1), \pi(a_2), \dots, \pi(a_n)\} \\ &= \min\{\pi'(a_1), \pi'(a_2), \dots, \pi'(a_n)\} \end{aligned}$$

また、Lemma 1 が成立するため割り当て値の修正によってハッシュ値は不変である。実際には多重度の上限值  $n$  をパラメータとして  $\pi'$  を保持する表を事前計算し、スライディングウィンドウに到着した要素の割り当て値は表を参照して決定する。割り当て表の修正例を図 4.2 に示す。

	a	b	c	d
1個目	3	4	7	6
2個目	5	1	8	2

↓

	a	b	c	d
1個目	3	4	7	6
2個目	3	1	7	2

図 4.2 割り当て値の修正

次にスライディングウィンドウ内の  $n$  個の  $a$  のインスタンスをそれぞれ何番目の  $a$  とするかを考える。つまり、 $n$  個の  $a$  のインスタンスのインデックスをどう定めるかという問題である。通常の多重集合であれば要素間に時間順序がないのでこの問題は重要でない。しかし、スライディングウィンドウの場合、要

素間に時間順序があるためインデックスの決め方によって挙動が変わる。自然な方式としては以下の2つが考えられる。

- 一番古い要素を  $a_1$  とし、 $a$  のインデックスを到着時刻の昇順とする
- 一番新しい要素を  $a_1$  とし、 $a$  のインデックスを到着時刻の降順とする

提案手法では、「一番古い要素を  $a_1$  とし、 $a$  のインデックスを到着時刻の昇順とする」という方式を採用する。この時、

- (1)  $a_i$  は  $a_{i+1}$  より先にデータストリームに到着した
- (2)  $\pi'(a_i) \geq \pi'(a_{i+1})$

$\pi'$  の広義単調減少性より、古い  $a_i$  はより新しい  $a_{i+1}$  の存在によって Minlist から削除される。このことが任意の  $i$  に対して成立するので、以下の Lemma2 が保証される。

**Lemma 2.** Minlist の中に同種アルファベットは最新の 1 要素しか存在しない。

以降では記述の単純化のため、修正後の値割り当て  $\pi'$  を単に  $\pi$  と記載する。

#### 4.2 ヒストグラムの作成

スライディングウィンドウが多重集合である場合、スライディングウィンドウ内にいくつ同一要素が含まれているかわからないと Min-hash のハッシュ値を計算できない。そこで要素のヒストグラムを保持する。さらに各要素の到着時刻も保持するためヒストグラムのビンを到着時刻のリストとして管理し、リストのサイズにより各アルファベットのスライディングウィンドウ内の個数を取得する。リストは到着時刻が昇順で保持され、新たにウィンドウに到着した要素の到着時刻がエンキューされ、ウィンドウから出ていく要素の到着時刻がデキューされる。

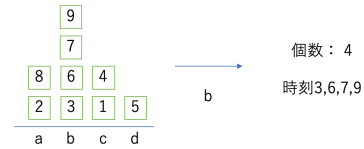


図 4.3 ヒストグラム

#### 4.3 スライディングウィンドウの更新

提案手法 SWMH でも Datar らの手法と同様に最小値になりうる要素のリスト Minlist を管理する。Datar らのアルゴリズムでは Minlist に割り当て値のみを保持していた。しかし多重集合の場合、割り当て値が変化が発生すると現在の割り当て値から変化後の値を計算できない。変化後の値を計算するため、Minlist は

- 現在の割り当て値:  $\pi(e)$
- 要素のラベル:  $l(e)$
- 要素の到着時刻:  $t(e)$

の3つ組を要素とするリストとする。さらに SWMH では多重集合  $A_t$  が  $A_{t+1}$  に変化する時、(1) ウィンドウから一番古い要素  $e_{t-w+1}$  が離脱する時の処理と (2) ウィンドウに新要素  $e_t$  に入ってくる時の処理を拡張しなくてはならない。



#### 4.3.1 要素 $e_{t-w+1}$ がウィンドウから出ていく処理

時刻  $t$  に割り当て値が最小でハッシュ値となる要素を  $\alpha$  とし、そのラベルを  $l(\alpha)$  とする。

(Case 1): Minlist の先頭要素の時刻が  $t-w+1$  ならば Minlist の先頭要素は  $e_{t-w+1}$  そのものなのでデキューする。この時、 $l(e_{t-w+1}) = l(\alpha)$  ならば Lemma2 より、Minlist 内に同じラベルを持つ要素は複数存在しないので  $e_{t-w+1}$  は  $\alpha$  と一致する。つまり、ハッシュ値となる  $\alpha$  がウィンドウから離脱するため、ハッシュ値の更新が必要となる。そこで、新たに最小割り当て値となる要素を Minlist から探索して更新する。Minlist は割り当て値順にソートされていないので、この操作は Minlist の全要素のスキャンを伴う。

(Case 2): Minlist の先頭要素の時刻が  $t-w+1$  でない場合は  $e_{t-w+1}$  は Minlist のメンバーではないため Minlist からのデキューは不要である。しかしこの場合も Minlist 内で最小割り当て値が変化する場合がある。具体的には離脱要素  $e_{t-w+1}$  のアルファベットが  $l(\alpha)$  である場合、 $l(\alpha)$  の多重度が 1 減るため  $\alpha$  の割り当て値が増加し、 $\alpha$  の割り当て値がウィンドウ内で最小でなくなる可能性がある。この場合も Minlist をスキャンし新たな最小割り当て値を探索する。

$e_{t-w+1}$  のアルファベットが  $l(\alpha)$  でない場合、最小割り当て値は不変である。しかし、Minlist 内にラベルが  $l(e_{t-w+1})$  の要素  $\beta$  が存在した場合、 $\pi(\beta)$  は同じラベルの要素数が減少することに伴い本来増加する。しかし、最小割り当て値には影響を与えないので Minlist 内では  $\beta$  の割り当て値を更新しない。この結果、Minlist 内に保持されている  $\beta$  の割り当て値は(一時的に)不正確になる。 $\beta$  の割り当て値の修正は、次回にハッシュ値を更新するために Minlist をスキャンするときに行う。正しい割り当て値は  $\beta$  のラベル  $l(\beta)$  の多重度をヒストグラムから得ることで算出する。

#### 4.3.2 要素 $e_{t+1}$ がウィンドウに入る時の処理

$e_{t+1}$  がスライディングウィンドウに入る時に、Minlist 内の要素が同じラベルを持つか違うラベルかで分けて処理を行う。

##### (1) 同じラベルを持つ要素の削除

Minlist 内の要素が  $e_{t+1}$  と同じラベルを持つ場合、割り当て値表を広域単調減少で保持しているため、必要なく、削除することができる。

##### (2) 違うラベルを持つ要素の削除

違うラベルを持つ要素の削除の場合、Datar らの手法では Minlist の中で  $\pi(e_{t+1})$  より割り当て値が大きい要素を消すだけで十分だった。しかし、多重集合の場合は  $\pi(e_{t+1})$  が将来増加する可能性があり、単に  $\pi(e_{t+1})$  より割り当て値が大きい要素を消すと割り当て値が最小となる可能性がある要素も消してしまう。そこで、 $\pi(e_{t+1})$  ではなく、 $e_{t+1}$  の将来の割り当て値の上限を上回る要素だけ削除する。 $e_{t+1}$  のラベルを  $l(e_{t+1})$  とする。また、Minlist 内の要素  $\gamma$  の到着時刻を  $t_\gamma$  として、 $t_\gamma$  よりあとに到着したアルファベット  $l(e_{t+1})$  の個数を  $n$  とする。要素  $\gamma$  がウィンドウ内に存在する間、 $\gamma, e_{t+1}$  間にラベルが  $l(e_{t+1})$  となる要素が  $n$  個ある。よって、要素  $\gamma$  の存在期間中の  $e_{t+1}$  の割り当て値は  $\pi(l(e_{t+1})_n)$  より大きくなならない。よって

$$\pi(l(e_{t+1})_n) < \pi(\gamma)$$

であれば、要素  $\gamma$  を Minlist から削除してよい。

##### (3) ハッシュ値の更新

最後に Minlist の一番後ろに  $e_{t+1}$  を挿入し、 $\pi(e_{t+1})$  が Minlist の最小値を更新するかをチェックする。現在のハッシュ値より  $\pi(e_{t+1})$  が小さいならば、ハッシュ値を更新する。

## 5 バッチ SWMH

前章で説明した SWMH では、データストリームに毎時刻要素が 1 つだけ到着することを仮定したが、現実のデータストリームでは、1 度に複数個の要素が到着することも普通である。そこで、そのような場合に SWMH を拡張することを目指す。本章では、データストリームに毎時刻  $c$  個の要素が到着するモデルを想定する。

### 5.1 自明な手法

データストリームに毎時刻  $c$  個の要素が到着するモデルにおいての自明な手法は、4 章で説明した SWMH を  $c$  回適応する手法である。つまり、スライディングウィンドウが  $c$  回ずれる事象を、スライディングウィンドウが 1 つずれる事が  $c$  回繰り返されると判断して、SWMH を  $c$  回適用する、1 回 SWMH を適用する度に Minlist が 1 回スキャンされるので、1 時刻あたり Minlist を  $c$  回スキャンすることになる。

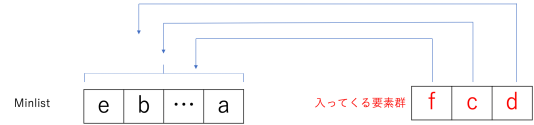


図 5.1 Minlist の更新

### 5.2 バッチ SWMH のアルゴリズム

本章では Minlist を  $c$  回スキャンせず 1 回だけスキャンするバッチ SWMH を提案する。バッチ SWMH のアルゴリズムでは、ウィンドウから要素が出ていく場合と入ってくる場合に処理を分けて考える。ウィンドウに  $c$  個の要素が出入りする場合、ウィンドウから出ていく要素群を  $E_{t-w/c+1} = \{e_{t-w/c+1}^1, e_{t-w/c+1}^2, \dots, e_{t-w/c+1}^c\}$ , 入ってくる要素群を  $E_t = \{e_t^1, e_t^2, \dots, e_t^c\}$  とする。

#### 5.2.1 要素群 $E_{t-w/c+1}$ が出ていく処理

バッチ SWMH の  $E_{t-w/c+1}$  が出ていく処理では、4.2.2 の SWMH のウィンドウから要素が出ていく時に対する処理と同じように場合分けをして処理を行う。Minlist 内で最小割り当て値となる要素  $\alpha$  のアルファベットを  $x$  と記述する。

(Case1): Minlist の先頭要素の時刻が  $t-w/c+1$  ならば Minlist の先頭要素をデキューする。さらに、Minlist の要素が  $t-w/c+1$  と違う時刻要素が出てくるまで Minlist をスキャンし、 $t-w/c+1$  と同じ時刻を持つ要素を削除する。この時、削除される要素が  $\alpha$  と同じアルファベット  $x$  を持つ場合、新たな最小割り当て値となる要素を Minlist から探索して更新する。Minlist は割り

当て値順にソートされていないので、この操作は Minlist の全要素のスキャンを伴う。

(Case2):Minlist の先頭時刻が  $t - w/c + 1$  でない場合、Minlist からのデキューは不要である。不要である理由は、4.2.2 で説明した理由と同じである。

### 5.2.2 要素群 $E_t$ が入ってくる処理

Minlist に  $E_t$  が入ってくる処理では、以下の 3 つの手順で行う。

- $E_t$  から代表アルファベットの選択
- 代表アルファベットと比較して、Minlist をスキャン
- Minlist へ  $E_t$  を追加

代表アルファベットとは、 $E_t$  の中で割り当て値が最小のアルファベット  $x$  と  $E_t$  の中でスライディングウインドウ内の多重度最大のアルファベット  $y$  である。

そして、Minlist のスキャンでは、 $E_t$  内の要素と同じラベルの要素の削除と違うラベルの削除の 2 つの処理を行う。Minlist の要素を  $\lambda$  とする。

#### (1) $E_t$ 内の要素と同じラベルの要素の削除

同じアルファベットを消すために、従来の手法では、 $\lambda$  に対して、 $E_t$  全体をスキャンし、同じアルファベットが複数存在するか確認する必要があった。しかし、バッチ SWMH では先に保持させたヒストグラムの情報の 1 つである時刻を用いて、まず、ヒストグラムに  $E_t$  内の要素のアルファベットの最新時刻を追加し、 $\lambda$  の最新時刻  $T(l(\lambda))$  と Minlist 内の  $\lambda$  の時刻  $t(l(\lambda))$  を比較し、時刻が違う場合  $\lambda$  を削除する。同じアルファベットの中で一番最後尾の要素のみを残し、ほかの要素を削除することができる。

#### (2) $E_t$ 内の要素と違うラベルの要素の削除

$x, y$  の 2 つのアルファベットを用いてスキャンを行なっていく。Minlist 内の要素  $\lambda$  の入ってきた時刻  $t_\lambda$  とし、 $t_\lambda$  よりあとに到着した  $x, y$  の個数  $n_x, n_y$  とする。時刻に応じて、 $x, y$  のうち小さい割り当て値を持つアルファベットの割り当て値を用いて、Minlist から将来の割り当て値の上限を上回る要素を削除していく。

$$\pi_{\min} = \min(\pi(x_{n_x}), \pi(y_{n_y}))$$

if ( $\pi_{\min} < \pi(l(\lambda))$ ) Minlist から  $\lambda$  を削除

手順 3:Minlist への追加

最小値となり得ない要素を削除した Minlist に対して、 $E_t$  を追加する。

## 6 実験

本章では提案手法である SWMH と、バッチ SWMH を人工データと実データを用いて実験的に評価する。まず、データストリームの到着レートが 1 である状況で SWMH を評価する。ここでは、時刻変化のたびにスライディングウインドウ全体をスキャンして、Min-hash のハッシュ値を再計算する手法を Baseline とし、SWMH が Baseline より高速であることを示す。

次にデータストリームの到着レートが  $c > 1$  である条件でバッチ SWMH を評価する。ここでは、5.1 節で述べた SWMH を  $c$  回適用する単純手法とバッチ SWMH の実行時間を比較する。

### 6.1 データセット

#### 6.1.1 人工データセット

人工データセットは、アルファベット  $\phi$  から、zipf 分布に従ってサンプリングを繰り返し、長さ 100,000 の文字列  $st$  を生成する。到着レートを  $c$  として、データストリームは  $st$  の先頭から順に毎時刻  $c$  個取り出すことで、シミュレートする。データセットの特性を決定するパラメータは

- $\alpha$ : zipf 分布の偏りをコントロールするパラメータ
- $\phi$ : アルファベットの種類数

であり、スライディングウインドウの長さを  $W$  とする。zipf 分布とは、 $\alpha$  を  $[0, 1]$  の範囲で指定し、アルファベットの出現頻度に偏りを持たせた分布である。アルファベットの出現順位を  $k, k$  番目のアルファベットの個数を  $N_k$ 、一番出現頻度の高いアルファベットの個数を  $N_{max}$  とすると、zipf 分布は以下の式 3 で表せる。

$$N_k = N_{max} \times 1/k^\alpha \quad (3)$$

#### 6.1.2 実データセット

実データによる実験は、connect dataset と mushroom dataset の 2 種類のデータ [4] [6] からデータベースを作成した。まず、connect dataset とは 117 種類のラベルが出現し、約 20 の長さのラベルから成る集合が約 4,000 個で構成されるデータである。次に、mushroom dataset とは 127 種類のラベルが出現し、約 20 の長さのラベルから成る集合が約 60,000 個で構成されるデータである。どちらのデータも 1 つの集合に同じラベルは含まれない。それらのデータからランダムに集合を 1 つ選ぶ処理を、選ばれた集合をつなげた長さがデータストリームの長さ 100000 +  $W$  を超えるまで行い、文字列  $st$  を生成する。人工データと同じように、到着レートを  $c$  として、データストリームは  $st$  の先頭から順に毎時刻  $c$  個取り出すことで、シミュレートする。

### 6.2 SWMH の実験評価

まず、人工データのパラメータは以下の 3 つである。

- $\alpha$
- $\phi$
- $W$

$\alpha = 1, |\phi| = 100, W = 100$  という組み合わせをデフォルトパラメータとし、それぞれのパラメータを変更し、実験を行なった。

次に、実データのパラメータは、

- $W$

のみであり、 $W = 100$  をデフォルトパラメータとし、実験を行なった。

それぞれの実験において、実験のばらつきを減らすために、

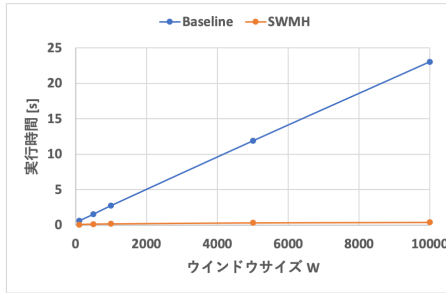


図 6.1 人工 dataset

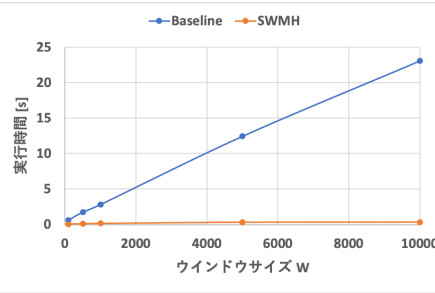


図 6.2 mushroom dataset

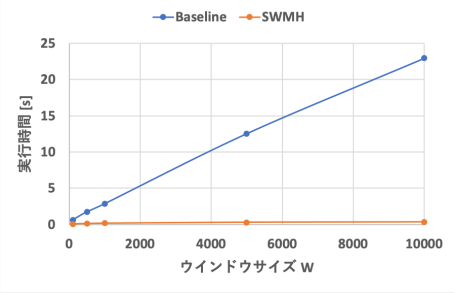


図 6.3 connect dataset

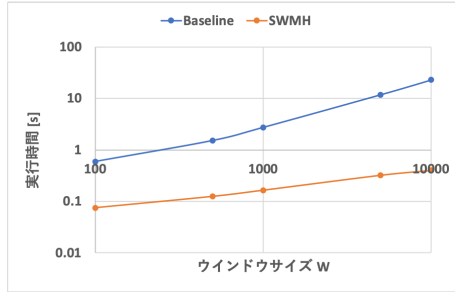


図 6.4 人工 dataset における対数グラフ

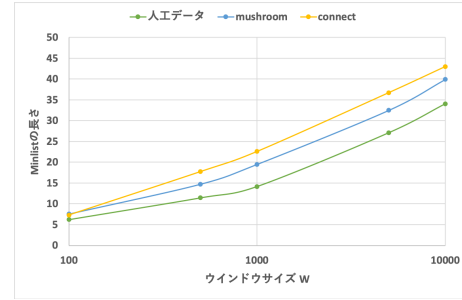


図 6.5 W に対する Minlist の長さ

ハッシュ関数を 10 個生成し、SWMH10 回の合計を実行時間とし、実験を行なった。

実験 1): ウィンドウサイズ  $W$  を変えて実験

$W = 100, 500, 1000, 5000, 10000$  と変えて、提案手法の実行速度と最小値候補リスト Minlist の長さを計測した。人工データ、connect、mushroom いずれのデータセットにおいても、SWMH が Baseline より圧倒的に早かった (図 6.1, 6.2, 6.3)。SWMH の実行時間は、Baseline と比べて、 $W=100$  の時、約 9 倍、 $W=1000$  の時、約 17 倍、 $W=10000$  の時、約 60 倍となった。実行時間と  $W$  の関係をより詳しく調査するため人工データに対する実行時間を対数グラフで表現したものを図 6.4 に表す。Baseline では、実行時間が  $W$  に対してリニアに増加するのに対して、SWMH では実行時間が  $\log W$  に比例して増加する。SWMH の実行時間は Minlist の長さに関係があると考えられる (図 6.5)。

実験 2): 偏りパラメータ  $\alpha$  を変えて実験

偏り  $\alpha = [0, 1]$  の幅で 0.1 ごとに変えて、SWMH の実行時間と Minlist の長さを計測した。

人工データ、connect、mushroom いずれのデータセットにおいても、SWMH が Baseline より早かった (図 6.6)。Baseline では、偏りが大きくなるにつれて実行時間が短くなることにに対して、SWMH では実行時間が長くなっていく。これは、偏りが大きいほど要素の出現に偏りができ、ハッシュ値計算時の割り当て値表へのアクセスが局所的になり、キャッシュ効率が改善したため Baseline は実行時間が短くなったと考えられる。しかし、SWMH では、偏りが大きくなるごとに同じ要素が多く出やすくなる影響から最小値と同ラベルの要素がウィンドウから離脱することが増えて、ハッシュ値の更新回数が増加したことが原因と考えられる。

実験 3): ラベル種類数  $|\phi|$  の変更

要素の種類数  $|\phi| = 10, 50, 100, 500$  と変えて、提案手法の実行速度と最小値候補リスト Minlist の長さを計測した。図 6.7 より、Baseline では、 $|\phi|$  が増加すると実行時間が増加する。この理由は、ラベルの種類数が多いほど、割り当て値表の広範囲にアクセスし、アクセスが局所的でなくなった結果、キャッシュ効率が限定的になるためと思われる。一方で、SWMH では、ラベルの種類数が増えるほど、Minlist 更新の際に、同じラベルを持つ要素同士の削除が減少し、実行時間が減ったと考えられる。

### 6.3 バッチ SWMH の実験評価

本実験では、人工データと connect、mushroom の 3 つのデータセットを使用した。まず、人工データのパラメータは到着レート  $c = 5$ 、 $\alpha = 1$ 、 $|\phi| = 100$ 、 $W = 100$  という組み合わせをデフォルトパラメータとし、実験を行なった。次に、実データのパラメータは、 $c = 5$ 、 $W = 100$  をデフォルトパラメータとし、実験を行なった。実験のばらつきを減らすためにハッシュ関数を 10 個生成し、バッチ SWMH の 10 回の試行の回数を実行時間として、実験を行なった。

人工データ、connect、mushroom いずれのデータセットにおいても、SWMH を  $c$  回適応する方法より、バッチ SWMHの方が約 2 倍早かった (表 1)。図 2 より、Minlist の長さはバッチ SWMHの方が長くなった。SWMH では、 $c$  回スキャンするため、 $c$  個の要素のアルファベットを使用するが、バッチ SWMH では 2 個の代表アルファベットだけを使用して、最小値にならない不要な要素を削除するために Minlist が長くなっていると考えられる。しかし、Minlist の各要素を入ってきた  $c$  個の要素全部と比較するのではなく、2 個の代表要素としか比較しないので、比較回数を減らすことができる。この結果、実行時間を短縮することができたと考えられる。

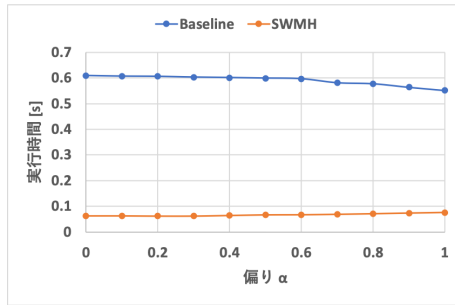


図 6.6 傾き  $\alpha$  を変えた実行時間

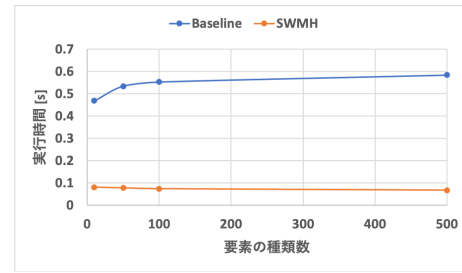


図 6.7 ラベルの種類数  $|\phi|$  を変えた実行時間

表 1 パッチ SWMH の実行時間

データセット	人工	connect	mushroom
SWMH	0.0742	0.0743	0.073
パッチ SWMH	0.042	0.0365	0.037

表 2 パッチ SWMH の Minlist の長さ

データセット	人工	connect	mushroom
SWMH	6.2883	6.2082	6.296
パッチ SWMH	6.5501	7.5545	7.4214

## 7 ま と め

本研究ではデータストリームに対する類似検索の高速化を念頭に、データストリームに対するハッシュ値の更新アルゴリズムを取り扱った。とくにスライディングウィンドウを動的に変化する多重集合と見なして、スライディングウィンドウモデルで多重集合を取り扱える初の手法である SWMH を提案した。

SWMH の特筆すべき点は、動的な多重集合に対して Min-hash を計算する場合、1つの要素への割り当て値が多重度に影響されて動的に変化するという難点に対応したことである。さらに要素への割り当て値をハッシュ値が変化しない範囲で修正することにより、SWMH が管理する必要がある要素数を削減させて、実行時間を短縮した。さらに、SWMH を拡張し、一度に複数個の要素が到着するモデルに対応したパッチ SWMH を提案した。パッチ SWMH では、ウィンドウに入ってきた要素群の中で、最小の割り当て値を保持する要素と最大多重度のラベルを保持する要素の2つに絞って、Minlist へアクセスする回数を減らすことで実行時間の短縮を狙った。人工 dataset と実データセットを用いて Min-hash 計算時間を測る実験を行い、提案手法を評価した。まず、SWMH は、毎時刻 Min-hash のハッシュ値を再計算するベースラインより、高速にハッシュ値を算出できることを示した。さらに、パッチ SWMH は、複数個の要素がウィンドウを出入りする場合に SWMH より短い時間でハッシュ値を高速計算できることを示した。

最後に、提案手法 SWMH では、データを保持するためにヒストグラムを多用しているため、多くのメモリを消費している。従って、今後の研究課題としては、メモリ使用量の削減のために近似ヒストグラムを用いてハッシュ値を計算する手法の実現が望まれる。

## 謝 辞

本研究は JSPS 科研費 JP21K11901 の助成を受けたものである。

## 文 献

- [1] Mayur Datar and S Muthukrishnan "Estimating Rarity and Similarity over Data Stream Window" AT&T Research, Florham Park NJ, USA.
- [2] X, Xu, C. Gao, J. Pei, K. Wang, and A. Al-Barakati, "Continuous similarity search for evolving queries," Knowledge and Information Systems, vol.48(3), pp.649-678, September 2016.
- [3] Yang D, Li B, Rettig L, Cudre-Mauroux P. "HistoSketch : fast similarity preserving sketching of streaming histograms with concept drift", 2017 IEEE international conference on data mining (ICDM); 2017. p. 545-54.
- [4] Wang P, Qi Y, Zhang Y, Zhai Q, Wang C, Lui J, Guan X. "A Memory-Efficient Sketch Method for Estimating High Similarities in Streaming Sets", 2019.
- [5] Ping Li, Arnd Christian König, "b-Bit Minwise Hashing", 2010 Nineteenth International World Wide Web Conference (WWW 2010).
- [6] Michael M, Rasmus P, Ninh, "Efficient Estimation for High Similarities using Odd Sketches", In WWW, pages 109-118, 2014.
- [7] W PM Rowe, A Paola Carrieri, C Alcon-Giner, S Caim, A Shaw, K Sim, J Kroll, L j. Hall, E O. Pyzer-Knapp, M D. Winn, "Streaming histogram sketching for rapid microbiome analytics" Rowe et al. Microbiome 2019.
- [8] D. Yang, B. Li, and P. Cudre-Mauroux, "Poisketch: Semantic place labeling over user activity streams," in Proc. of IJCAI, 2016, pp. 2697-2703.
- [9] M Bury, C Schwiegelshohn, M Sorella, "Efficient Similarity Search in Dynamic Data Streams", 2021.
- [10] A Z Broder, M Charikar, A M Frieze, M Mitzenmacher, "Min-Wise Independent Permutations", Journal of Computer and System Sciences Volume 60, Issue 3, June 2000, Pages 630-659.
- [11] M Manasse, F McSherry, K Talwar, "Consistent weighted sampling", 2010.