2022 年 9 月 29 日

令和 4 年度　先端工学基礎課程　中間発表

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の改良

I 類 CS　学籍番号：1920031　古賀研究室　山川 竜太郎

1　はじめに

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が増している。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みなし、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 係数が良く用いられる。しかし、集合が変わるたびに Jaccard 係数を計算しなおすのはオーバーヘッドが大きい。そのため、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、「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」の研究を引継いでいる。

1.1　スライディングウィンドウモデル

ストリームデータをスライディングウィンドゥモデルで多重集合を扱える手法が先行研究としてある。スライディングウィンドウモデルは、データストリームの直近 W 個の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のデータを破棄する手法である。また、Jaccard 係数の算出を高速化するために、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、それを利用する。(1)

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

式 (1) で h はハッシュ関数であり、A,B は集合である。Min-hash によるハッシュ値の計算方法は、$\Phi = \{x_1, x_2, \cdots x_{|\Phi|}\}$ をアルファベット集合としたとき、ハッシュ関数はΦの各アルファベットに対して $\{1, 2, \cdots |\Phi|\}$ 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。個の最小値が Min-hash によるハッシュ値となる。

1.2　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として、多重集合内の複数個のラベルに異なる値を割り当てている。その時に将来に割り当て値が最小になる可能性が絶対にない割り当て要素は、直前の割り当て値と同じなるように割り当て表を編集する。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。

2　過去研究の課題

過去の論文で提案されている手法 SWMH(Sliding-Window min-Hash) 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。したがって課題として、メモリ使用量削減のためにいくつかの手法を用いて空間計算量を削減することで、メモリの使用量の削減を図る。また、現在の割り当て表は要素の種類数 ×多重度となっているため、それを減らして空間計算量を削減して、メモリ使用量削減を図る。

3　改善する意義

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ていなかった。本研究では、過去に行われたスライドウィンドゥに対して多重集合を取り扱える手法を実現する方法を改良して、高速化したアルゴリズムにすることを目的としている。

4　改善案

実現するアイデアとして、Active Index と Count-Min Sketch を用いる。Active Index と用いて、割り当て表に数値を割り当てるとき、多重度によって割り当て値が切り替わる境界値のインデックスと割り当て値を持つ。例として、多重度 1 のときに割り当て値が 10、多重度 2 以降の割り当て値が 9 の場合は、1,10,2,9 のように割り当て表を持つ。Count Min Sketch とは、1 つのハッシュ関数だけでなく、異なるハッシュ関数を 2 つ以上使用することである。これらのハッシュ関数はペアごとに独立である。カウントを更新するには、項目 a を d 個のハッシュ関数でハッシュし、その後、この方法で得た全てのインデックスをインクリメントする。2 つのハッシュ関数が同じインデックスに対応する場合、そのセルを 1 回だけインクリメントする。使用可能な領域を増やさない限り、この方法はハッシュの衝突の数を増やすだけである。もし今カウントを取得したいので

あれば、最大で d 個の異なるセルを見る必要がある。自然な解決策は、これらのすべての最小値を取ることである。これは他のセルとのハッシュの衝突が最も少なかったセルとなる。これが Count-Min Sketch の基本的な考え方である。この名前は、最小値を取ることによってカウントを求めることに由来する。

5　進捗状況と今後の予定

現在は、Active Index と Count-Min Sketch を SWMH モデルに適用できるかどうか、検討している段階である。参考文献を探して、SWMH の空間計算量を削減することを目標としている。したがって、まずは SWMH の中に組み込めるか否かを疑似コードを作成して検証する。Active Index は割り当て表に対して実装し、割り当て表の空間計算量を削減できるようにする。Count-Min Sketch は、スライドウィンドゥ内の要素の種類をカウントするヒストグラムに対して実装して、Minlist の持っている要素を少なくして、最小値を近似できるようにする。

参考文献

[1] 三原寛寿, 古賀久志, “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,“ 電気通信大学情報理工学研究, 2022

[2] Florian Hartmann, “Count-Min Sketch,“ Google Research, 2019