時間と共に変化する多重集合に対する min-hash の高速計算

古賀研究室 2031136 三原寛寿

2021 年 7 月 12 日

1　はじめに

近年，集合間類似検索が注目されている．集合間類似検索とはクエリ集合と類似している集合をデータベースから検索する問題である．2 つの集合が似ているかどうかを表す類似度は何度も計算する必要があり，Jaccard 係数を用いる．

$$sim(S,Q) = \frac{|S \cap Q|}{|S \cup Q|} \tag{1}$$

Jaccard 係数とは，2 つの集合に含まれる要素のうち共通要素が占める割合である．集合間類似検索はクエリとデータベースの全集合で Jaccard 係数を求めれば解くことができる．しかし，Jaccard 係数による類似度の厳密計算は類似度を計算するオーバーヘッドが大きい．

そのため，集合間類似検索を高速化する手法である Min-hash を用いる．Min-hash[2] は 2 つの集合のハッシュ値が一致する確率が Jaccard 係数と等しいという性質を持ち，ハッシュ値を使って，類似集合を検索できる．

通常，Min-hash は要素が不変な集合を対象とする．要素が時間と共に変化する集合に対しては，ハッシュ値の再計算をする必要がある．Datar ら [1] はスライディングウインドウモデルに従って動的に変化する集合に対して，ハッシュ値を高速更新する手法を提案している．

$$P(mh(A) = mh(B)) = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

そして，本研究の目的は，Datar らの手法を多重集合に拡張することである．

以下、本文の構成を述べる．2 章で，Min-hash の計算方法を説明する．3 章で 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し，それを動的多重集合に拡張する 2 つの問題について説明する．4 章では 3 章であげた 2 つの問題を解決するために行った提案手法である割り当て値の修正と最小値となり得ない要素の発見方法を紹介する．5 章では，実験評価を行い，6 章でまとめと今後の課題を述べる．

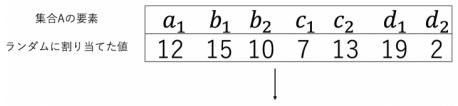
2　Min-hash

Min-hash の計算方法は，集合 A の要素に対して，割り当てられた値から最小値を選び，その値がハッシュ値となる．多重集合の場合，同じアルファベットを複数含む場合には異なる割り当て値を与える．(図 1) 多重集合とは，集合を同一集合の要素を複数持てるようにしたものである．例えば,$\{a, b, b, c, c, d, d\}$ という集合である．

計算例として，図 2 のように，集合 A に対して，図 1 の割り当て値を参照し，その中の最小値 2 がハッシュ値となる。

	a	b	c	d
1 個目	3	15	7	19
2 個目	12	10	13	2

図 1　多重集合の割り当て表



ハッシュ値 ： $h(A) = 2$

図 2　Min-hash によるハッシュ値計算

3　動的に変化する集合に対する類似度計算

最近では動的に変化する集合に対する類似検索も注目されている．ストリームデータとは時間と共に変化するデータであり，直近の w 個の要素をスライディングウインドウとして，動的に変化する集合とみなせる．そして、ストリームデータの類似検索は，集合の要素が変化するため，毎時刻ハッシュ値の再計算が必要となる．

3.1　Datar らによる手法

Datar ら [1] は時間と共に変化する集合に対するハッシュ値更新を効率よく行う方法を提案した．Datar らの手法は，将来的に最小値になり得ない要素を削除，残りを Minlist で管理し，Minlist の最小値をハッシュ値とす

るという方法である．最小値の候補集合の作り方は，図 3 のように，自分より新しい要素に，小さい値がある場合，将来，最小値になる可能性なく，最小値になる可能性のある要素だけの集合 Minlist を作成する．このように作成した Minlist は元の集合より小さくなるため，類似度を高速に計算することができる．しかし，この手法は多重集合で取り扱うことができない．

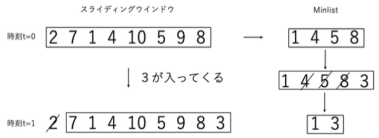


図 3　最小値の候補リスト Minlist の作成

3.2　Datar の手法を拡張する難しさ

多重集合では，スライディングウインドウ内の個数によって，割り当て値が変化する．そのため、Minlist を作る上で，問題点が二つある．1 つ目は，Minlist 内に，同一要素が出現し，Minlist が長くなること．2 つ目は、割り当て値の変化により，最小値が変わる可能性があり，最小値になるかどうかの判断が難しくなることである．

4　提案手法

本研究では，Datar[1] らのハッシュ値更新のアルゴリズムを動的多重集合に拡張する手法を考案した．

提案手法の特色は 2 つである．

（1）同一アルファベットに対する割り当て値を修正し，Minlist 内に同一アルファベットが 1 つしかないことを保証する．ここでは，割り当て値を修正するが，それによって，集合に対するハッシュ値は変化しない．

（2）最小値になり得ない要素を割り当て値の変化があっても特定可能にした．

4.1　割り当て値の修正

Minlist 内の同一アルファベットの要素を 1 つにするために，割り当て値の修正を行った．修正方法は．同じアルファベットの中で，i 番目の割り当て値より $i+1$ 番目の割り当て値が大きければ修正するという方法である．

$$if(\pi(\alpha_i) < \pi(\alpha_i+1))\{if\pi(\alpha_i+1) = \pi(\alpha_i)\} \quad (3)$$

4.2　最小値となり得ない要素の発見方法

Minlist において，到着したアルファベット:α,Minlist 内の要素:β,β よりあとに到着した α の個数:n, 割り当て値:π として，β よりあとに到着した α の個数に応じた π により，将来的に最小値になり得るかどうか判断する．

$$if(\pi(\alpha_n) < \pi(\beta))\{\beta\text{を } Minlist \text{ から削除 }\} \quad (4)$$

5　実験評価

データベース内の集合数を 1000 であり，集合は毎時刻更新，スライディングウインドウサイズを 100 とし，時刻 t=1 から 1000 まで，1000 回の 10nn 検索にかかる処理時間を計測した．Baseline は，毎時刻，スライディングウインドウの割り当て値から最小値を見つけるアルゴリズムである．

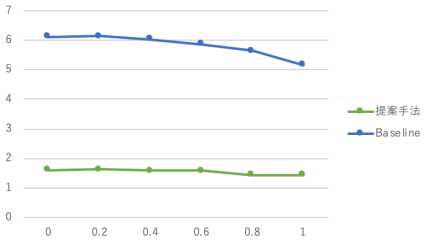


図 4　要素の偏りを変化させた実験

結果として図 4 より，提案手法は，Baseline に対して，4 倍ほどの速度となった．

6　まとめ

本研究では，動的に変化する集合に対して，Min-hash 値を更新するアルゴリズムを多重集合へ対応するように拡張した．そのために，割り当て値の修正と要素の入った時刻によって，最小値となり得ない要素を判断し，将来的に最小値となるリストを作成した．この提案手法は，Baseline の約 4 倍の速度となった．

今後，データへのアクセスの仕方で，時間がかかっていることが判明しているため，変数へのアクセスパターンを考慮しながら，アルゴリズム開発と実装を行っていく．

7　参考文献

[1]Mayur Datar and S Muthukrishnan ”Estimating Rarity and Similarity over Data Stream Window” AT&T Research, Florham Park NJ, USA

[2] 岡野原 大輔, “MinHash による高速な類似検索”, https://research.preferred. jp/2011/02/minhash/