修　士　論　文　の　和　文　要　旨

|  |  |  |  |
| --- | --- | --- | --- |
| 研究科・専攻 | 大学院　情報理工学研究科　　情報・ネットワーク工学専攻　博士前期課程 | | |
| 氏　　　　名 | 三原寛寿 | 学籍番号 | 2031136 |
| 論文題目 | 時間と共に変化する多重集合に対するmin-hashの高速計算 | | |
| 要　　旨 | | | |
| 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，集合が変化する度にJaccard係数を計算し直すのはオーバヘットが大きい．そこで集合に対して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のハッシュ値計算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)の条件下で多重集合を取り扱える初めての手法となるSWMH (Sliding Window Min Hash)を提案する．SWMHはスライディングウィンドウで（多重集合ではなく）集合を取り扱った手法を拡張した．とくにＳＷＭＨは同一ラベルの多重度に応じて集合の要素へのMin-Hashの割り当て値が動的に変化するという難しさを解決した点が優れている．さらにMin Hashの要素の割り当て値を、ハッシュ値が変化しない範囲で修正することで、SWMHでは同一ラベルの要素群に対して1要素だけ管理することを実現している，実験評価により，SWMHが毎時刻Min-hashのハッシュ値を再計算するベースライン手法より圧倒的に高速であることを示す．ウィンドウサイズWとすると、実験的には、SWMHの実行時間はlog Wに比例する．  さらに，SWMHをデータストリームに毎時刻複数個(の要素が到着するモデルに拡張した．この手法をバッチSWMHと呼ぶ．バッチSWMHはSWMHが管理するMinlistというリスト構造をスキャンする回数を削減することで高速処理を実現する．1度に5個の要素が到着する状況で，SWMHを5回適応する手法よりバッチSWMHの方が約2倍早いことを実験的に示した． | | | |