様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ６ 月　 ８ 日現在

機関番号：１４４０１
研究種目：若手研究(B)
研究期間：2016～2019
課題番号：１６Ｋ１６０５６

研究課題名（和文）ストリーム環境におけるデータモニタリングに関する研究

研究課題名（英文）Data Monitoring over Stream Environments

研究代表者
天方　大地（Amagata, Daichi）

大阪大学・情報科学研究科・助教

研究者番号：４０７７０６４９

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：本研究では，ストリーミングデータにおいて有用なデータ（知識）をモニタリングする問題に取り組んだ．特に，各データが要素の集合として表される集合データに着目した．取り組んだメインの問題は大きく以下の2つである．1つは複数（大量）のストリームデータで共起しているものをストリーム数でランキングし，その上位をモニタリングする問題，もう一方は，各データの最も類似している他のk個のデータをモニタリングする問題である．これらの問題に対して厳密解をリアルタイムにモニタリングするアルゴリズムを開発し，データベース分野の最難関国際論文誌（IEEE TKDE）および最難関国際会議（IEEE ICDE）に採録された．

研究成果の学術的意義や社会的意義
本研究では，多くのアプリケーションに活用できる問題に着目し，大量のデータに対応できるための効率的なアルゴリズムを開発した．本研究で開発したアルゴリズムは，新たなデータ構造を利用しているものであり，特定の問題に対して効果的に動作するデータ構造の開発は本研究分野における学術的価値が非常に高い．また，実サービスが本研究で提案したアルゴリズムを利用することにより，解析に利用できるデータを10倍以上増やすことができる（既存技術よりも10倍以上高速である）．大量のデータを利用することにより有用な知識が得られることは既に広く知られており，本研究成果の社会的価値も高いことがわかる．

研究成果の概要（英文）：In this work, we addressed the problem of streaming data monitoring. In particular, we focused on dynamic set data. We mainly addressed the following two problems: (i) top-k co-occurrence pattern monitoring across multiple streams and (ii) dynamic set kNN self-join, which monitors the k nearest neighbor set for each set in real-time. For these problems, we proposed fast and exact algorithms with new data structures. These works respectively appear in a top-tier journal (IEEE TKDE) and a top-tier conference (IEEE ICDE) in database filed.

研究分野：データベース

キーワード：ストリーミングデータ　データモニタリング　データ構造　アルゴリズム

様　式　C－１９、F－１９－１、Z－１９（共通）

１．研究開始当初の背景

近年，多くのアプリケーションにおけるデータがストリーム形式で生成されている．例えば，Twitter では 1 秒あたり数万ツイートが生成されており，通信ネットワークや e コマースにおいても，オンライントランザクションデータが絶えず生成されている．また，Internet of Things を実現する動きが活発となり，より多くのモノがインターネットに接続し，データを頻繁に生成することが予想されることから，ストリームデータ処理の重要性は増々大きくなっている．

ストリーム環境では，単位時間（例えば 1 秒）毎に大量のデータが生成されるため，全てのデータ（データの全体集合）をモニタリングすることは非現実的である．そこで，発生したデータをある用途に合わせて処理し，重要な部分集合のみをリアルタイムにモニタリングすることが求められている．例えば，ユーザが指定したスコアリング関数をもとにデータにスコアを付与し，その上位 k 個をモニタリングする技術が開発されている．このように，ユーザが求めるデータを抽出する問い合わせは研究されているが，発生したデータの全体集合の特徴を最もよく表すような k 個のデータ（または部分集合）をモニタリングするための問い合わせについては，これまでに考えられていない．人間が処理可能なサイズ（k 個）のデータ（集合）のみを抽出することは，ストリームデータの解析やマイニングを効率化する働きがあり，先述の多くの現代的なアプリケーションに役立つことが期待できる．そこで本研究では，上記のモニタリングを実現するための問い合わせ方式を数学的に定式化し，リアルタイムモニタリングを実現するための技術の開発を目指した．

２．研究の目的

本研究課題では，Twitter やセンサネットワーク，オンライントランザクション等で発生するデータをストリームデータとして扱い，データ解析やマイニングを効率化するため，データモデルに依存することなく，データ集合全体の特徴を最もよく表す k 個の部分集合を抽出し，それらのモニタリングについて考えた．また，これを実現するための問い合わせ方式の考案，および，リアルタイムモニタリングのための高速アルゴリズムのデザインを目的とした．

３．研究の方法

本研究は，4 年計画で実施した．データモニタリングのための問い合わせの考案，および，効率的なモニタリングアルゴリズムのデザインが研究の主要な部分となるため，各年度においてアプリケーションが求めるデータモニタリングを実現する問い合わせの定式化，ベースラインとなるアルゴリズムのデザイン，効率的なアルゴリズムのデザイン，考案アルゴリズムの評価実験の順に段階的に研究を推進した．

４．研究成果

Top-k Co-Occurrence Pattern Mining across Multiple Streams. 本研究は，複数のストリームが存在する環境を想定している．複数のストリーム間で同じデータが一定の時間窓に出現している場合，それらは共起とみなすことができる．多くのストリーム間で共起しているデータは，何らかのパターンと考えることができ，知識マイニングに有用である．このパターンマイニング方法は，トピック抽出，web 利用パターンマイニング，e コマース，および相関ルールマイニングといった，非常に多くのアプリケーションで用いることができるが，ストリーミングデータという性質からリアルタイムな処理が求められる．そこで，最も多くのストリーム間で共起している上位 k 個（top-k）のパターンをリアルタイムにモニタリングするアルゴリズムを開発した．

本問題は，あるパターンのカウントを，そのパターンが現れるストリームの数と定義している．最も単純なモニタリングアルゴリズムが，全てのパターンのカウントを計算するものであるが，全てのパターンの数え上げ問題は NP 完全であり，リアルタイムに解くことは非現実的である．そこで，本研究では本問題の理論的性質を最初に解析し，解に含まれることがないパターンを無視する理論を証明した．次に，解に含まれるかもしれないパターンのカウントを高速に計算するデータ構造である CP-Graph（図１）を設計した．このデータ構造の主な特徴は，あるパターンのカウントの上界値を O(1)時間で計算できることである．これにより，あるパターンが top-k に入る可能性があるかないかを O(1)時間で判定できる．また，top-k に入る可能性がある場合においても正確なカウントを高速に計算できる．このデータ構造はあるストリームに新たなデータが追加された場合においてもインクリメンタルに更新す

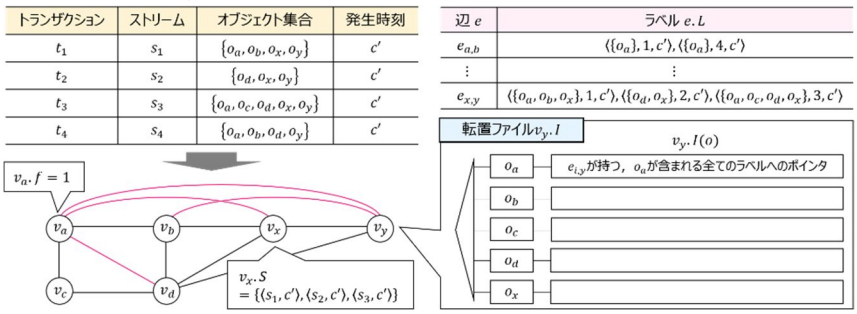

図 1：CP-Graph の例

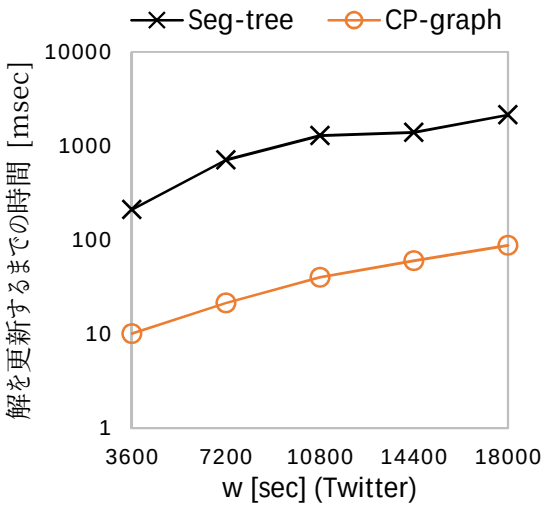

図 3：計算時間（Twitter）

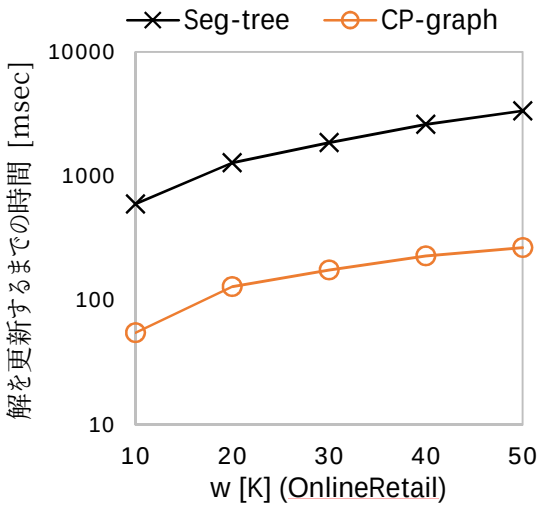

図 2：計算時間（OnlineRetail）

ることできる．こういった性質から，提案したアルゴリズムは解を高速に出力し，リアルタイムモニタリングを可能としている．

実データを用いた実験を行い，提案アルゴリズムの性能評価を行った．その結果の一部を図２および３に示す．ここで，Seg-tree は別の問題の既存アルゴリズムを本問題に適応させたものである．図２および３から，提案アルゴリズムは既存技術の約 10 倍高速であることがわかる．また，横軸はウィンドウ（窓）サイズを示しており，窓内に大量のデータが存在する場合においてもスケールしていることが分かる．本研究の成果は，データベース分野のトップ国際論文誌である IEEE Transactions on Knowledge and Data Engineering に採録されている．

Dynamic Set kNN Self-Join. 本研究では，データ＝要素の集合と定義しており，集合の類似度を Jaccard 類似度で計算する．このとき，各集合に対して k 最近傍が得られるが，各集合が動的に変化する（新たな要素を得たり，要素が削除される）環境において，全ての集合に対する k 最近傍をモニタリングする問題に取り組んだ．本問題は，推薦システムやデータクリーリングといった実サービスのプリミティブオペレータであるが，これまでに本問題に対する解法は提案されていなかった．本研究はこの問題を効率的に解いた最初の研究である．

この問題の課題は，ある集合 s の要素が追加・削除された場合，s と他の集合の類似度が全て変わってしまう点にある．つまり，一つ要素が追加されただけで s の k 最近傍が大きく変わってしまう可能性があり，s 以外の集合の k 最近傍も変わる可能性がある．こういった問題の性質を理論的に解析し，要素が追加または削除された際，どういった集合と類似度が大きくなる，または小さくなるかを証明し，どのような集合にアクセスしなければならないか，またはアクセスしなくてよいかを証明した．これをもとにデータ構造を設計し，必要最小限の類似度計算で k 最近傍を更新するアルゴリズムを設計した．また，提案アルゴリズムの妥当性（考えられる別のアプローチとの比較）を理論的に分析し，効率性も示した．

提案アルゴリズムの性能評価を行うため，10 個の実データを用いた実験を行った．図 4 は結果の一例を示している．データの分布に依存はするものの，提案アルゴリズムは既存の静的データに対する join アルゴリズムよりも（大幅に）高速である．本研究の成果は，データベース分野の最難関国際会議の一つである IEEE International Conference on Data Engineering にフルペーパーで採択された．

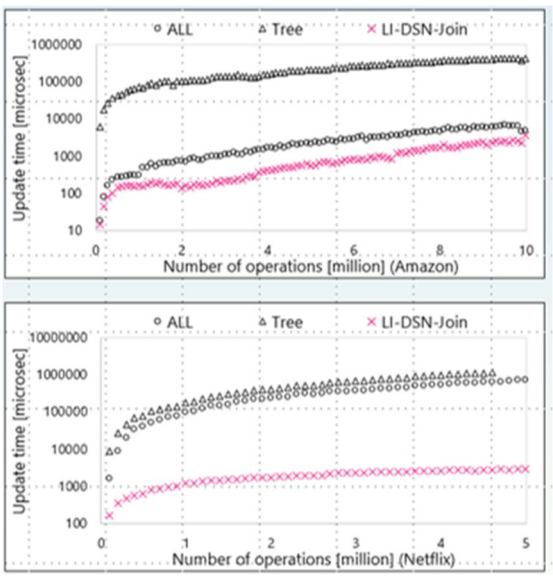

図 4：Dynamic Set kNN Self-Join の実験結果の一例

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　4件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 加藤 慎也，天方 大地，西尾 俊哉，原 隆浩	４．巻 60
２．論文標題 ストリーミング時系列データに対するモチーフモニタリング	５．発行年 2019年
３．雑誌名 情報処理学会論文誌	６．最初と最後の頁 1260-1269
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Daichi Amagata, Takahiro Hara	４．巻 29
２．論文標題 Mining Top-k Co-Occurrence Patterns across Multiple Streams	５．発行年 2017年
３．雑誌名 IEEE Transactions on Knowledge and Data Engineering	６．最初と最後の頁 2249-2262
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TKDE.2017.2728537	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Daichi Amagata and Takahiro Hara.	４．巻 3
２．論文標題 A General Framework for MaxRS and MaxCRS Monitoring in Spatial Data Streams	５．発行年 2017年
３．雑誌名 ACM Transactions on Spatial Algorithms and Systems	６．最初と最後の頁 1-34
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 加藤 慎也，天方 大地，西尾 俊哉，原 隆浩	４．巻 61
２．論文標題 ストリーミング時系列データに対するディスコードモニタリング	５．発行年 2020年
３．雑誌名 情報処理学会論文誌	６．最初と最後の頁 510-519
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計10件（うち招待講演　1件／うち国際学会　6件）

1．発表者名
Shinya Kato, Daichi Amagata, Shunya Nishio, Takahiro Hara

2．発表標題
Monitoring Range Motif on Streaming Time-Series.

3．学会等名
International Conference on Database and Expert Systems Applications (DEXA)（国際学会）

4．発表年
2018年

1．発表者名
加藤 慎也, 天方 大地, 西尾 俊哉, 原 隆浩

2．発表標題
ストリーミング時系列データの効率的なディスコードモニタリングアルゴリズム

3．学会等名
第11回データ工学と情報マネジメントに関するフォーラム（DEIMフォーラム2019）

4．発表年
2019年

1．発表者名
Daichi Amagata, Takahiro Hara, Chuan Xiao

2．発表標題
Dynamic Set kNN Self-Join

3．学会等名
IEEE International Conference on Data Engineering (ICDE)（国際学会）

4．発表年
2019年

1．発表者名
Daichi Amagata, Takahiro Hara

2．発表標題
Mining Top-k Co-Occurrence Patterns across Multiple Streams (Extended abstract)

3．学会等名
IEEE International Conference on Data Engineering（国際学会）

4．発表年
2018年

１．発表者名
天方 大地，原 隆浩

２．発表標題
相関時系列データ集合の計算のための高速アルゴリズム

３．学会等名
FIT2017 第16回情報科学技術フォーラム

４．発表年
2017年

１．発表者名
Daichi Amagata and Takahiro Hara

２．発表標題
Diversified Set Monitoring over Distributed Data Streams

３．学会等名
ACM International Conference on Distributed and Event-Based Systems (DEBS)（国際学会）

４．発表年
2016年

１．発表者名
天方 大地，原 隆浩

２．発表標題
複数ストリーム環境におけるTop-k共起パターンマイニング

３．学会等名
第9回データ工学と情報マネジメントに関するフォーラム（DEIMフォーラム2017）

４．発表年
2017年

１．発表者名
天方 大地

２．発表標題
Monitoring MaxRS in Spatial Data Streams

３．学会等名
日本ソフトウェア科学会第33回大会（招待講演）

４．発表年
2016年

１．発表者名 Daichi Amagata and Takahiro Hara
２．発表標題 Correlation Set Discovery on Time-series Data
３．学会等名 International Conference on Database and Expert Systems Applications (DEXA)（国際学会）
４．発表年 2019年

１．発表者名 Shinya Kato, Daichi Amagata, Shunya Nishio, Takahiro Hara
２．発表標題 Discord Monitoring for Streaming Time-series
３．学会等名 Shinya Kato, Daichi Amagata, Shunya Nishio, Takahiro Hara（国際学会）
４．発表年 2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考