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　元 年　 ６ 月 １９ 日現在

機関番号：３６１０２
研究種目：基盤研究(C)（一般）
研究期間：2015～2018
課題番号：１５Ｋ００００５９

研究課題名（和文）大規模データからの集約的シンボリックデータの作成に関する研究

研究課題名（英文）Study of generation of aggregated symbolic data from large data

研究代表者
山本　由和（YAMAMOTO, Yoshikazu）

徳島文理大学・理工学部・教授

研究者番号：８０２４０１３３

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：本研究では，大規模データから適切な集約的シンボリックデータを作成できるようにすることを目標として，2つのことを行った．1つは，1台の計算機で処理できないような大規模データから集約によって，インタラクティブに処理できる程度の大きさのデータを作成する．この時に並列分散処理を行う．もう1は，作成したデータについての適切な集約的シンボリックデータを作り出す．これは，データによって異なるために，可視化と対話的な操作による試行錯誤を行う．この時に，必要に応じてデータ作成処理に戻ることもある．

研究成果の学術的意義や社会的意義
適切な集約的シンボリックデータによって，データサイズが大き過ぎるために，今までは，傾向を把握できなかったようなデータの傾向を把握できるようになる．特に，可視化によって，直感的に理解できるようになる．

研究成果の概要（英文）：In this study, we made two things possible with the goal of generating suitable aggregated symbolic data from large-scale data. One is to create data of a size that can be processed interactively by aggregating large-scale data that can not be processed by one computer. Parallel distributed processing is performed at this time. The other one produces appropriate aggregate symbolic data about the created data. Since this differs depending on the data, it performs trial and error by visualization and interactive operation. At this time, it may return to the data creation process as needed.

研究分野：計算機統計学

キーワード：シンボリックデータ解析　可視化　並列分散処理　大規模データ

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

本研究のキーワードは，可視化，　並列分散処理，　大規模データ，　集約的シンボリックデータである．

データ解析のための可視化については，Mondrian や iPlots などのソフトウェアが有名である．われわれも，統計グラフィクスライブラリ Jasplot (Java statistical plot)を開発している．これらのソフトウェアでは，表示できる統計グラフィックス上での対話的な操作やグラフィックス間の連携などを行える．しかし，データ量の増加によって，描画速度や対話的操作に対するレスポンスが急激に低下する．大規模データに対する可視化も行われているが，本研究では，ビッグデータに関連する技術を利用することによって，さらに大規模なデータに対応することを考える．

大規模データの解析については，特に最近の国内・国外の統計関係学会において，ビッグデータに関する特別セッションが設けられ，研究が行われている．本研究では，並列分散処理によって，大規模データを処理して，集約的シンボリックデータを作成することを考える．

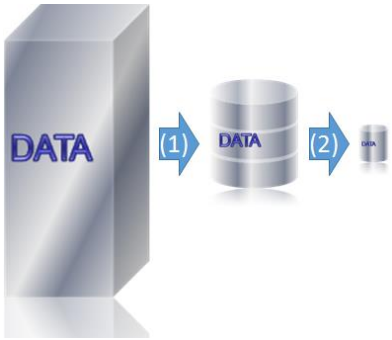
集約的シンボリックデータに関する研究は，現在もわれわれが継続しており，データの特徴を自然に表すことを目的として，シンボリックデータに同時分布の情報も考慮したものである．本研究では，大規模データから可視化を利用して，適切な集約的シンボリックデータを作成することを考えている．

２．研究の目的

本研究の目的は，大規模データから適切な集約的シンボリックデータを作成できるようにすることである．適切な集約的シンボリックデータとは，大規模データの特徴を表す，サイズの小さいデータである．このために，並列分散処理と可視化技術を利用する．

３．研究の方法

本研究では，右の図の(1)と(2)の処理を実現する．



(1)は，並列分散処理によって，大規模データから集計などによってインタラクティブに処理できる大きさのデータを作り出す過程である．

(2)は，(1)で作り出したデータの可視化と対話的な操作による試行錯誤によって，適切な集約的シンボリックデータを作り出す過程である．

(1)は，大規模データのサイズを小さくする過程である．図の左側が大規模データ，中央がこの処理によって作成されるデータを表している．ここで作成されるデータは，可視化と対話的処理が可能な程度の大きさである．このために，何らかの基準によって，いくつかのグループに分けて，そのグループごとの集計を行う．グループを作るための基準は，次のようなものである．

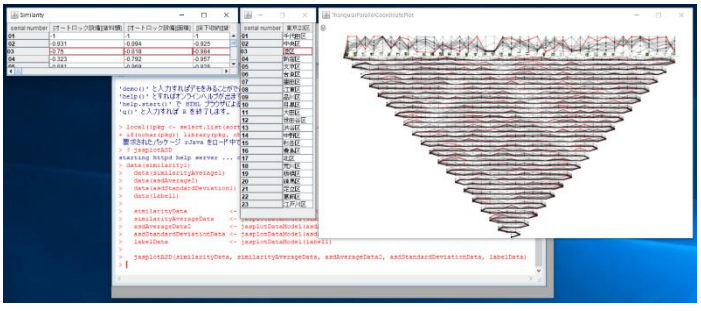
- 実数値変数の値を区間に分けたもの
- カテゴリ変数が含まれているデータであればその変数の値
- これらの両方が含まれていれば，それらの組み合わせ

集計は，グループごとに次のようなものである．

- 各実数値変数の平均値と標準偏差
- 各カテゴリ変数のカテゴリごとの度数
- 2 変数間の類似度

特に，グループの分け方や集計処理は，それぞれのデータに対して変更できるようにする．これは，大規模データの処理になるので，並列分散処理を行う．この処理のためには，ビッグデータに関する並列分散処理技術である MapReduce を利用する．

(2)は，(1)で作成したデータから適切な集約的シンボリックデータを作り出す過程である．図の右側が適切な集約的シンボリックデータを表している．この集約的シンボリックデータは，元の大規模データの特徴や構造を自然に表す必要がある．さらに，元のデータの代わりに解析を行えるようなものを作成したいと考えている．そのために，可視化とそのグラフィックス上での対話的操作を利用する．



右の図は，本研究の可視化に関する部分において利用する予定の Jasplot によって表示した，集約的シンボリックデータの例である．このグラフィックス上では，マウス操作による対話的操作と複数のグラ

フィックス間の連携が可能である．この機能によって，グループ間の関係を見ることができる．この機能を利用して，適切な集約的シンボリックデータを作成する．具体的には，次のような操作を行う．

- 同じ構造のグループがあれば，1 つにまとめる
- 1 つのグループに異なる構造が含まれていれば，分割する
- 2 変数ごとの関係から，重要な変数や省略できる変数などを選定する

４．研究成果

「3.研究の方法」で説明した(1)と(2)を実現した．(1)については，MapReduce と Apache Spark による並列分散処理の設計・実装を行った．特に，可視化を利用して，より適切なシンボリックデータを作り出すための試行錯誤を可能にすることを考えた．このために，大規模データに対する処理を高速に行うことが必要である．これを実現するために，Apache Spark を使用した．さらに，様々な統計計算を可能にすることを目的として，R から Apache Spark を利用した並列処理を行うためのソフトウェアである SparkR も使用できるようにした．この結果として，R のプログラムを利用した計算を並列分散処理によって大規模データに適用できるようになった．(2)については，Java 言語による可視化ソフトウェアである Jasplot を R から利用できるようにした．これによって，SparkR で計算した結果を分散ファイルシステムからローカルファイルシステムに集めて簡単な操作で可視化できるようになった．この結果を実データに対して適用して，集約的シンボリックデータの作成を行った．

５．主な発表論文等

〔雑誌論文〕（計 4 件）

1. 清水信夫，中野純司，山本由和．集約的シンボリックデータのカイ 2 条統計量を用いた非類似度とその不動産情報データへの適用．統計数理，査読有．第 66 巻，第 2 号，279-294，2018
2. 山本由和，高野慎也．アニメーションを利用した購買情報の表示．日本計算機統計学，査読有．第 29 巻，第 1 号，77-87，2016
3. 尾﨑皇彦，高野慎也，山本由和．Web サイトにおける閲覧行動の可視化．日本計算機統計学，査読有．第 28 巻，第 2 号，155-164，2015
4. 山本由和，尾﨑皇彦．大規模データの集計と予測のための MapReduce アプリケーション．日本計算機統計学，査読有．第 28 巻，第 1 号，19-27，2015

〔学会発表〕（計 15 件）

1. 陶山瑞樹，川田成美，後藤健文，山口良三，山本由和．D3.js を利用した可視化ソフトウェアの作成．日本計算機統計学会第 32 回シンポジウム 講演論文集，pp.69-72，滋賀大学，11 2018
2. 川田成美，陶山瑞樹，後藤健文，山口良三，山本由和．訪日外国人流動データの可視化ソフトウェアの作成．平成 30 年度電気関係学会四国支部連合大会 講演論文集，pp.212，愛媛大学，9 2018
3. 清水信夫，中野純司，山本由和．集約的シンボリックデータの変数選択．2018 年度統計関連学会連合大会，p.339，中央大学，9 2018.
4. 陶山瑞樹，石川まりな，山本由和．顧客情報の可視化と予測．日本計算機統計学会第 32 回大会講演論文集，pp.80-83，山口大学，5 2018.
5. Yoshikazu Yamamoto, Junji Nakano, and Nobuo Shimizu. Interactive visualization of aggregated symbolic data for summarizing huge datasets. In Conference of 6th International IBM Cloud Academy Conference 2018, Tokyo, 5 2018.
6. 陶山瑞樹，山本由和．アンケートと購買金額による会員のクラスタリング．日本計算機統計学会第 31 回大会講演論文集，pp.119-120，東京理科大学，5 2017.
7. 山本由和，中野純司，清水信夫．グループの特徴の可視化と対話的処理．2017 年度統計関連学会連合大会，p.298，南山大学，9 2017.
8. Yoshikazu Yamamoto, Junji Nakano, and Nobuo Shimizu. Interactive visualization of aggregated symbolic data. In New Zealand Statistical Association and the International Association of Statistical Computing (Asian Regional Section) Joint Conference 2017, p.282, Auckland, 12 2017.
9. Yoshikazu Yamamoto, Junji Nakano, and Nobuo Shimizu. Interactive visualization of characteristics of groups. In Conference of the International Federation of Classification Societies, p. 300, Tokyo, 8 2017.
10. 山本由和，松田真実，藤本祐規．MapReduce と Spark を利用した大規模データのクラスタリング．日本計算機統計学会第 30 回シンポジウム講演論文集，pp.139-142，プラサヴェルデ，11 2016.
11. Yoshikazu Yamamoto, Mami Matsuda, Yuki Fujimoto, Nobuo Shimizu, and Junji Nakano. Clustering large data sets using MapReduce and Apache Spark. In Proceedings of the

2016 International Conference for JSCS 30th Anniversary in Seattle, pp. 62{65, Seattle Central Library, 10 2016.
12. Yoshikazu Yamamoto. Clustering huge data sets using Hadoop and Spark. In Proceedings of the KSS Autumn Conference 2016, p. 51, Statistics Korea, Statistics Center, 11 2016.
13. 清水信夫, 中野純司, 山本由和. カテゴリ変数を含む集約的シンボリックデータの非類似度の性質. 2016 年度統計関連学会連合大会, p.94, 金沢大学, 9 2016.
14. 高野慎也, 中田誠人, 松田真実, 山本由和. 購買情報の時間変化の可視化. 日本計算機統計学会第 29 回大会講演論文集, pp. 81-84, 山梨大学, 5 2015.
15. 中田誠人, 山本由和. JavaFX 3D Graphics を利用した購買情報の可視化. 日本計算機統計学会第 29 回大会講演論文集, pp. 93-96, 山梨大学, 5 2015.

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年：
国内外の別：

○取得状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年：
国内外の別：

〔その他〕
ホームページ等

６．研究組織

(1)研究分担者
研究分担者氏名：
ローマ字氏名：
所属研究機関名：
部局名：
職名：
研究者番号（8 桁）：

(2)研究協力者
研究協力者氏名：
ローマ字氏名：