科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　２４　日現在

機関番号：　３２５１５
研究種目：　若手研究(B)
研究期間：　2011～2013
課題番号：　２３７００３４０

研究課題名（和文）データ解析システムにおけるビジュアルプログラミング環境の研究と開発

研究課題名（英文）Developing a visual programming environment in a data analytical system

研究代表者
藤原　丈史（Fujiwara, Takeshi）

東京情報大学・総合情報学部・准教授

研究者番号：６０３４８４５６

交付決定額（研究期間全体）：（直接経費）　　2,400,000円、（間接経費）　　720,000円

研究成果の概要（和文）：　データ解析システムにおけるユーザインタフェースとして，ビジュアルプログラミング環境を実現した．具体的にはアイコンを組み合わせることによりデータ解析の流れを視覚的に組み立てることができる．これにより，初心者においても直感的な操作および解析が可能となる．また，表現方法としては UML (Unified Modeling Language) を採用することにより，他のさまざまなシステムとのデータのやりとりが可能となる拡張性をもたせることも実現した．

研究成果の概要（英文）：The research has designed and developed a visual programming environment as a user interface of a data analysis system. Using this system, a user can assemble and construct analytical flows for data as linking icons visually and easily. Because of adopting UML (Unified Modeling Language) as expression of analytical flows and internal data format, this system has extendibility that is capable of exchanging a program of analytical flows to other systems.

研究分野：　総合領域

科研費の分科・細目：　情報学・統計科学

キーワード：　統計解析システム　ユーザインタフェース　プログラミング言語　UML

１．研究開始当初の背景

近年におけるコンピュータ環境の発展により，多種多様で，かつ膨大な量のデータが取得，蓄積されるようになった．これらの莫大なデータはビッグデータと呼ばれ，そこからさまざまな知識，知見を引き出すことが積極的に行われている．ここで利用されるのが大規模データの蓄積技術であるデータベースとともに，データを分析するデータ解析システムであり，その重要性は日々ますます高まっている．

データ解析システム（統計解析システム）においては，読み込みからクレンジング等のデータの前処理，統計手法やデータマイニング手法等を用いた分析や予測，そしてグラフィクスをはじめとする分析結果の出力など，対象データに対してさまざまな処理を段階を経て行っていく．

このような分析処理の流れは，一般的にはそれぞれのデータ解析システムに固有の統計解析言語を用いたプログラムを記述することが多い．もちろんこの統計解析言語を用いた分析は，システムを使いこなしているユーザや統計解析の専門家であれば，自由度は高く効率的に分析を行うことができる．しかしながら，システムやデータ解析自体の初心者においては，統計解析言語を用いた解析は困難である．

また，そのような解析の流れを表すプログラム，もしくはビジュアルプログラミングで記述した処理手順は，他のシステム間での共有利用は難しく，知識資源の有効活用という面では非効率である．

今後のデータサイエンスのさらなる発展に向けては，このようなデータ解析システムにおける環境を改善することが必要となっている．

２．研究の目的

本研究の目的は，データ解析システムにおける操作環境として，ビジュアルプログラミング環境を実現することにより，データ解析における利便性および効率性を向上させることを目的とする．また，データ解析の中心となる統計学およびデータサイエンスの裾野を広げることで，それらの分野の普及と今後の発展に貢献することである．

具体的には，従来の統計解析言語を用いたCUI (Character User Interface) ベースの解析処理の記述から，アイコンなどを利用したグラフィカルな解析の記述を行えるようにする．もちろん従来の文字ベースの記述はユーザのレベルや解析の状況によっては十分に有効であるが，初学者にとっては適切とはいえない．このため，処理の流れについて，文字を利用したプログラムだけで書く，ということではなく，アイコンのクリックやドラッグといった GUI (Graphical User Interface) 操作により，データ解析の処理を記述するビジュアルプログラミング環境は有効である．またデータ分析の初段階では，試行錯誤的に分析を行うことも多く，分析の履歴という意味でも効果的といえる (図 1)．

このようなビジュアルプログラミング環境を提供するデータ解析システムは，既存のシステムの中にもいくつか存在する．しかしながら，その表現方法および操作方法としては，各システムに固有のものとなっており，システム間に共有できるものではない．そこで本研究では，ビジュアルプログラミング環境の表現方法として，他のシステムにも汎用的に使えるような標準的なものとなるように実現することも重要である．これは現在におけるデータ解析システムの処理の記述が，システム固有の統計解析言語や内部形式で記録されているためであり，例えばあるデータ解析システム上で作成した分析処理はそのシステム上でしか利用できない．これはデータ解析システム間だけではなく，他のコンピュータシステム間においても同様であり，分析処理という知識資源の活用がうまくできている状況ではない．したがって，そのような有用な知識資源である分析処理手順を相互利用できるような表現方法の実現も本研究の目的といえる．

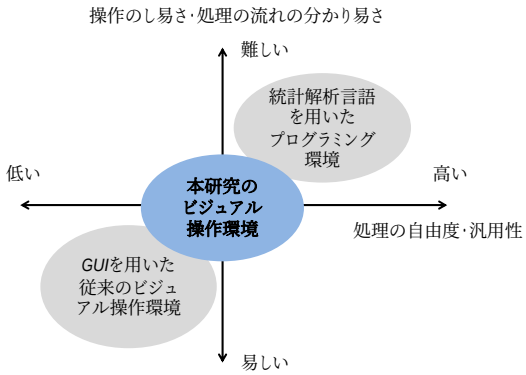

図 1 本研究のビジュアル操作環境

３．研究の方法

ビジュアルプログラミング環境として，どのような表現方法を用いるかは，本研究においての重要なポイントといえる．例えば，データや処理，およびその処理の順序や処理の間でどのようなデータ（オブジェクト）をやりとりするか，などをどのような形式（例えばアイコン）で表すか，どのようにそれらを操作して処理手順を構成していくか，といったビジュアル操作環境全体についてどう実現するかが問題となる．

既存のビジュアル操作環境の多くは，処理やデータをアイコンで表し，それらを線でつなぐことにより処理手順を表している．しかしながら，システムごとにアイコンの形状等

が異なり，統一的な表現方法は実現されていない．本研究の目的のひとつとして，他システムとの相互利用性は重要であり，その点も考慮した上で検討した結果，本研究のシステムではビジュアルプログラミング環境におけるその操作の表現方法として，現在のコンピュータシステムのモデリング言語として標準化されている UML (Unified Modeling Language) (*1) を採用した．

この UML は OMG (Object Management Group) が現在策定を行っている標準化された仕様記述言語であり，システム設計のさまざまな側面を複数のダイアグラムで表現するものである．一般的なソフトウェア設計やシステム設計に広く用いられている．現在のバージョンは 2.4 (2014 年 3 月現在) であり，13 種類のダイアグラムを必要に応じて使い分ける．

この UML の複数のダイアグラムの中でも本研究では，アルゴリズムやプログラムの図示ではよく用いられるフローチャートに似た，処理の流れを記述するためのアクティビティ・ダイアグラムをビジュアルプログラミング環境の表現方法として利用した．

これにより，データ解析の処理の流れを汎用的に表現し，他のシステムとの相互利用が可能となる．もちろん表面的なグラフ表現としての汎用化だけでなく，内部的には XMI (XML Metadata Interchange) というメタデータの情報交換の標準規格を用いることで実現している．XMI は主に UML モデルを異なるアプリケーション間でインポートおよびエクスポートするための XML (Extensible Markup Language) 形式のフォーマットである．この XMI をシステム間でやり取りを行う中間言語として利用することで，今後本研究システムで構築した分析処理手順を他のシステムでも利用できるといった相互共有の拡張性をもたせることができる (図 2).

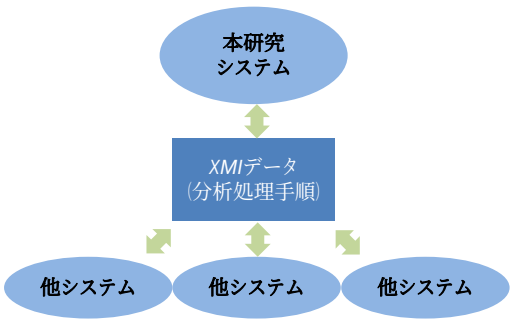

図 2 XMI による拡張性

実際のシステム全体の設計，実装方法の方針としては，現在のコンピュータ環境においては大きく分けて２つあげられる．ひとつは，システムを構成するプログラムのすべてをほぼ最初から作成する方法であり，もうひとつは既存のシステムやライブラリといったコンピュータ資源を最大限活用して実現する方法である．本研究は，研究代表者単独で実施しているものであり，前者の方法は時間的な制約からは難しく，後者の既存システムを有効に活用する実現方法をとった．

具体的には，ベースとなるシステムとして，操作環境としては UML 作成ツールである ArgoUML (*2)，解析処理の実行環境としては統計解析システムである R (*3) を用いることにした．

ArgoUML は Java 言語で開発されているオープンソースの UML 作成ツールで，UML バージョン 1.4 で定義されているアクティビティ・ダイアグラムを含む 9 つのダイアグラムに対応している（2014 年 3 月現在）．

統計解析システム R もオープンソースであり，非商用の統計解析システムとしてはもっとも利用され，開発自体も非常に活発である．世界中の多くの研究者が活用しているため，豊富なライブラリが揃っており，かつ，信頼性も高い．

この 2 つのシステムをベースに，ビジュアルプログラミング環境としてのデータ解析の操作および実行ができるようにシステムを設計および構築を行った．実行および開発環境としては，操作環境のベースとなる ArgoUML に合わせ Java 言語を用いている．これにより，Windows や Mac をはじめ，さまざまなプラットフォームで利用可能となっている．統計処理の実行を担う R とのやり取りは，Java 環境から R を実行できるライブラリ JRI (Java/R Interface) (*4) を使うことで実現している．

システム内部の具体的な処理の実行は以下のように行っている．まず，ビジュアルプログラミング環境におけるユーザインタフェース上では，データ解析の処理手順は UML のアクティビティ・ダイアグラムを表すオブジェクト，より具体的には Action State や Object Flow といったオブジェクトで内部的には構成されている．処理の実行時にはそれらのオブジェクトツリーが XMI 形式に変換される．さらにその XMI で表された処理構造を解析し，R の統計解析言語である R 言語に変換を行う．その変換後のコマンドについて，JRI を通して実際に統計処理として R 上で実行する．その結果のオブジェクトを再び JRI を通して受け取り，実行結果の表示等を行っている (図 3).

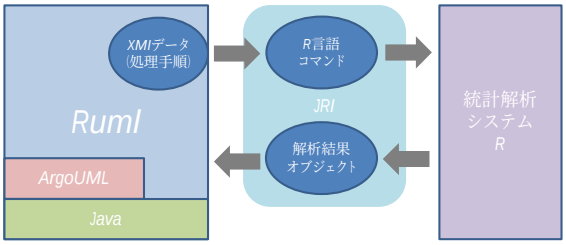

図 3 Ruml 内部構成・処理

*1・・・http://www.uml.org/
*2・・・http://argouml.tigris.org/
*3・・・http://www.r-project.org/
*4・・・http://rforge.net/JRI/

4．研究成果

研究成果として，UML を活用したビジュアルプログラミング環境をもつ，あらたなデータ解析システムである Ruml (R visual programming environment with UML) (図 4) を実現した．

このシステムでは，ユーザは処理の流れおよびデータ（オブジェクト）についてアイコンを使ってつなぎ合わせることで，基本的なデータ解析を行うことができる．これにより直観的な操作でシステムを利用でき，かつその処理手順は XMI 形式となるので，他のシステムとの相互利用が可能な汎用性をもつ．また，実行エンジンとして統計解析システム R をベースとしているので処理結果の信頼性は高く，豊富なライブラリを利用できる拡張性ももつ．

今回の研究期間内においては，研究目的として当初目標としていたシステムのベースとなる設計および実装までは行うことができ，システムの方向性は十分に示せた．しかしながら，研究途中においての，より実用的なシステムにするための大きな設計方針の変更等があり，論文発表等の研究成果の報告はこれからである．具体的には，当初は本研究代表者が開発に参加しているプロジェクトである統計解析システム Jasp を統計処理の実行エンジンとして採用していた．Jasp は現在のコンピュータ環境の技術を積極的に取り入れた先進的なシステムであるものの，現在も開発中であり，実用システムとしての統計ライブラリの充実と信頼性という意味では現在でも発展途中である．したがって，本研究の目的であるビジュアル操作環境の実現という観点からは，実行エンジンとしてはすでに信頼性が高く，実用上豊富なライブラリが利用可能である統計解析システム R を採用する方針へと変更した．

今後継続的に実用システムとなるようなチューニングを行いながら，システム全体の信頼性および実用性を高めるとともに，研究結果の公表を行っていく予定である．

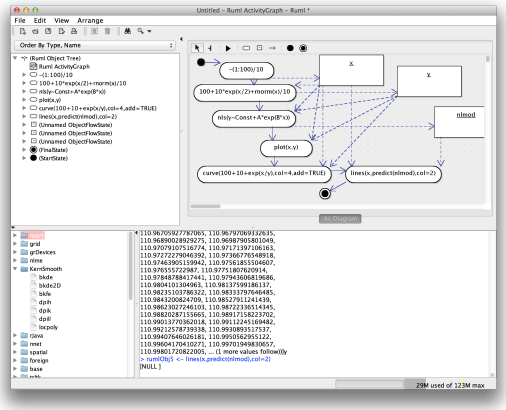

図 4 Ruml

5．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　0　件）

〔学会発表〕（計　0　件）

〔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：
出願年月日：
国内外の別：

○取得状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

6．研究組織
(1)研究代表者
藤原　丈史（FUJIWARA TAKESHI）
東京情報大学・総合情報学部・准教授
研究者番号：60348456

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者

(　　)

研究者番号：