時間的順序関係を考慮した相関分析手法の提案

山 谷 陽 亮[†1]　大 平 雅 雄[†1]

本稿では，メトリクス間に時間的な順序関係が伴う場合にも分析が可能な相関分析手法を提案する．ワークショップでは，どのようなメトリクスでどのような順序関係を抽出できるかについて議論したい．

A Correlation Analysis Method Considering Temporal Order Relations Between Metrics

Yosuke Yamatani[†1] and Masao Ohira[†1]

This paper proposes a correlation analysis method considering temporal order relations between metrics. In this workshop, we would like to discuss what metrics we choose to extract temporal order relations.

1. はじめに

オープンソースソフトウェア (OSS) を用いたソフトウェア開発が主流となりつつある．OSS をソフトウェア開発に利用することで，システムの低価格化・短納期化が期待できる．しかし，OSS プロジェクトは，ボランティアの開発者によって開発・保守が行われているため，コミュニティが分裂したり，時期によって品質が異なるといったことがしばしば生じる．

このような懸念を払拭するために，OSS プロジェクトを対象とした研究が盛んに行われている．例えば Crowston らは，OSS プロジェクトの成功要因を特定するために，複数のプロジェクトを対象として，様々な種類のメトリクスの相関を分析している[1)]．ただし，既存研究の多くは，ある時点で取得したメトリクスを用いて目的変数（例えば，成功要因）と説明変数（例えば，コア開発者数）との関係を扱っているため，時間的な順序関係を伴う因果関係は観察できないという問題があった．例えば，プロジェクトに参加する開発者が増えると一定期間後にソースコードの規模が増える，などといった関係は，既存の相関分析の枠組みで抽出するのは困難である．

そこで本稿では，メトリクス間に時間的な順序関係が伴う場合にも分析が可能な相関分析手法を提案する．

†1 和歌山大学システム工学部
Faculty of Systems Engineering, Wakayama University

2. 時間的順序関係を考慮した相関分析

提案手法は，竹内ら[2)] によって提案された時系列データ分析手法（以下，遅延相関分析）に基づいている．本章ではまず，竹内らの遅延相関分析について述べた，その後，本研究の提案手法を紹介する．

2.1 遅延相関分析

図 1 は，遅延相関分析の概念図であり[3)]，e_i は時刻 i における説明変数の値を表している．ある加算係数 $(i-j)$ における説明変数の値の累積値を e_{ij} とすると，e_{ij} は (1) 式で定義される．

$$e_{ij} = e_i + e_{i-1} + \cdots + e_j \tag{1}$$

また，r_n は時刻 n における目的変数の値であり，ある差分係数 $(n-m)$ における目的変数の値の変化値

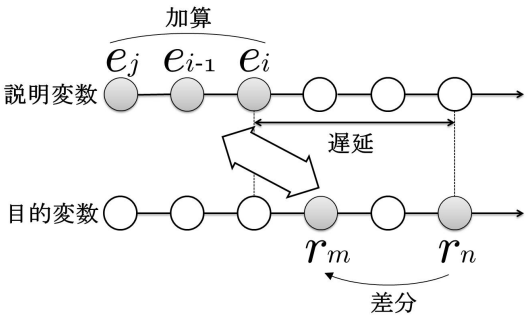

図 1 時系列データ処理の概念 ([3)] を参考に改編)
Fig. 1 Processing time-series data

表 1 各リポジトリから抽出するメトリクス
Table 1 Metrics in each repository

リポジトリ	メトリクス
バージョン管理システム（ソースコード）	総コード行数，複雑度の平均，クラス宣言の数，関数宣言の数，空白の行数，セミコロンの数，宣言部のコード行数，処理を行うコード行数、コメントの行数，宣言部のステートメント数，処理を行うステートメント数，制御構造の最大ネストレベル値，など
バージョン管理システム （変更履歴）	コミッタ数，新しいコミッタ数，コードオーナがコミットしている割合，コミット回数，変更したファイルの数，追加した行数，削除した行数，説明文の文字数，など
不具合管理システム	報告者数，割り当て者数，修正者数，変更を行った人数，報告件数，割り当て数，修正数，再割り当てした回数，Reopen した回数，コメントの数，割り当て時間，修正時間，など
メーリングリスト	メール数，スレッド数，リプライ数，メール投稿者数，ドメイン数，平均文字数，平均行数，リプライ時間，など

を r_{nm} とすると，r_{nm} は (2) 式で定義される．

$$r_{nm} = r_n - r_m \tag{2}$$

相関係数 c_{er} は (3) 式で定義され，S_e および S_r はそれぞれ e_{ij}, r_{nm} の標準偏差，S_{er} は e_{ij}, と r_{nm} の共分散である．

$$c_{er} = \frac{S_{er}}{S_e \times S_r} \tag{3}$$

このように，遅延相関分析は，説明変数の値が累積した結果，一定期間後に目的変数の値の変化に影響を与えると想定した相関分析手法である．

2.2 提 案 手 法

提案手法は，時系列データとして表現されるメトリクスを入力すると，すべてのメトリクスのペアについて遅延相関分析を行う．また，説明変数の値が累積し，一定期間の遅延をもって目的変数の値の変化に影響を与えるということを表現するために，加算係数（説明変数の値を累積させた期間），差分係数（目的変数の値の変化を考慮する期間），遅延係数（説明変数が目的変数に影響を与えるまでの期間）の 3 つのパラメータを定義する．

提案手法には，この 3 つの各パラメータの最適値を，相関係数が最大となるよう求める処理が含まれる．そのため，収集したデータを提案手法を用いて解析することで，分析者は遅延の大きさや分析窓の大きさを気にせず，相関係数が大きくなるメトリクスの組み合わせのみを分析対象とすることができる．

2.3 対象メトリクス

表 1 は，提案手法が現在対応可能なメトリクスの一覧を示したものである．これらのメトリクスは，Git や Subversion などのバージョン管理システム，および，Bugzilla などの不具合管理システムのリポジトリ，メーリングリストのアーカイブから抽出するものと想定としている．

ソースコードに関するメトリクスは，ソースコード解析ツールである Understand [⋆1]を用いて抽出している．その他のメトリクスは，著者らが作成したスクリプトを用いて抽出した．

3. お わ り に

本稿では，メトリクス間に時間的な順序関係が伴う場合にも分析が可能な相関分析手法を提案した．また，本手法が現在対応可能なメトリクスを紹介した．今後は，実際の OSS プロジェクトにおいてケーススタディを行う予定である．

ワークショップでは，どのようなメトリクスを用いてどのような順序関係を抽出するべきかどうかについて議論したい．

謝辞 本研究の一部は, 文部科学省科学研究補助金 (基 盤 (B):23300009) および (基盤 (C):24500041) による助成を受けた．

参 考 文 献

1) Crowston, K., Annabi, H., Howison, J. and Masango, C.: Towards a portfolio of FLOSS project success measures, *Workshop on Open Source Software Engineering, 26th International Conference on Software Engineering*, pp. 29–33 (2004).
2) 竹内裕之，児玉直樹：生活習慣と健康状態に関する時系列データ解析手法の開発，*Proceedings of the 3th Forum on Data Engineering and Information Management(DEIM'08)*, E1-5 (2008).
3) 黛　勇気，竹内裕之，児玉直樹：生活習慣と健康状態の時系列データ解析における重み付けの検討 (I) -日毎の任意係数による重みづけ-，*Proceedings of the 3th Forum on Data Engineering and Information Management(DEIM'11)*, D7-5 (2011).

⋆1 テラマトリックス社：Understand
http://www.techmatrix.co.jp/quality/understand/