リポジトリマイニング

松　下　　誠[†]

本稿では，ソフトウェア工学の分野で最近注目されつつある，ソフトウェア工学に関連する種々のデータに対する分析技術である，リポジトリマイニングについて紹介する．リポジトリマイニングでは，分析対象となるリポジトリに対してデータマイニング手法を用いることにより，新たな知見を得ることを目標としている．

Mining Software Repositories

Makoto Matsushita[†]

This article explains *repository mining*, a technique to analyse data of software engineering. The main focus on repository mining is that we could uncover a novel findings from applying data mining techniques to lots of repositories.

1. はじめに

現在行われているソフトウェア開発，あるいは，すでに行われたソフトウェア開発が「うまくいっている」かどうかを判断することは，開発組織はもちろん，ソフトウェアを利用する側や開発にたずさわる開発者にとっても有用である．これまでにも，ソフトウェアメトリクスに関する研究をはじめとして，ソフトウェア自体やソフトウェアプロセスを計測するための研究が行われてきた．

しかしながら，ソフトウェア開発の形態は多種多様であるため，どのような手法が対象を計測するにあたり有効であるか，はっきり断言するのは困難といえる．そのため，ソフトウェアやその開発に関する実証データに基づいた地道な努力が継続的に行われている．例えば，プロセスやプロダクトを何らかの形で抽象化した上で,「プロセスあるいはプロダクトはこのようなものである」という仮定を置いて理論を構築し，その理論を用いて既存の開発データを用いて検証することによって，理論の正しさを示す，というような，トップダウン的な方法が取られている．

一方，ソフトウェア開発時に広く用いられている種々のツール群は，一般的にいろいろな形でデータを蓄積する．例えば，版管理ツールでは，管理対象となったソースコードやドキュメントの変更履歴がデータとして蓄積されており，電子メールによる開発グループ内での連絡内容は，電子メールアーカイブ等に蓄積されている．近年の計算機システムの進化により，ハードディスク等は非常に安価で潤沢となっており，文字通り「開発の際に用いられたすべてのデータ」を何らかの形で保存することが可能となりつつある．また，オープンソースソフトウェアの開発では，分散した開発者が協調しながら作業を行う必要があるため，開発に関わるデータをどこかに蓄積しておき，適宜必要なデータをとりだして利用するといったことが行われている．このような開発が広まった結果，ネットワーク上にソフトウェアの開発データが大量に提供されるような状況となってきている．

ソフトウェアの開発データが大量に利用可能となってきたことに伴い，そのようなデータを既存のデータマイニング手法等を用いて分析することにより「その開発データがどのようなものであるか，あるいは，それを生みだしたプロセスや組織がどのようなものであるか」をボトムアップ的に明らかとする研究，リポジトリマイニングに関する研究が行われはじめている．リポジトリマイニングの研究では，リポジトリの存在が前提となるため「結果として何が得られるか，なぜそれが得られるのか」をはっきりとさせにくいものの，仮に何らかの結果が得られるならば，その結果は必ず既存の開発データから導かれたものであることが自明であり，従って得られた結果はすでに検証されていると考えることができる．リポジトリマイニングの研究では，分析対象とマイニング手法を効果的に組みあ

† 大阪大学 大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka University

わせることによって，開発データが蓄えられているリポジトリから新しい知見を得られる可能性や，分析によって雑多な情報が蓄えられているリポジトリから適宜必要な情報を抽出できる可能性がある．

以下，リポジトリマイニングにおいて対象となるリポジトリやデータマイニング手法，最近の研究の動向等について述べる．

2. リポジトリ

リポジトリマイニングが分析対象とするリポジトリは多岐に渡るが，一番広く用いられているのは，版管理ツールが蓄積するデータ (以下，開発リポジトリ) であろう．

開発リポジトリには，過去に行なわれたソフトウェア開発において，管理されているプロダクトに対して加えられた修正内容すべてが蓄えられている．また，版管理ツールが修正内容を蓄える際には，開発者名や時刻，開発者によって記入される当該修正に対するコメント (コミットログ) といったメタ情報をあわせて記録するのが一般的である．版管理ツールは，プロダクトの内容を管理するものと考えられるが，リポジトリマイニングにおいては，プロダクトの内容のみならず，メタ情報も分析対象として用いることが多い．マイニング結果としては，プロダクトに関するものはもちろんのこと，既存の複雑度メトリクス等によってわかるプロダクトの性質と開発者や開発期間との相関といった，プロセスに関するものがある．

その他，前述した電子メールや，障害記録といった情報も分析対象として用いられている．また，複数のリポジトリを対象とした分析も行われている．

3. データマイニング

データマイニング[1] は，一般的には大量の分析対象 (データ) から，そのデータを見ただけでは自明ではないような，有用な情報を抽出する手法である．リポジトリマイニングの研究では，前述したリポジトリを対象として，データマイニングの手法を適用することにより，有用な情報を発見することを目指す．

データマイニングには多くの手法がある．リポジトリマイニングにおいて，主に用いられている手法として，クラスタリング，頻出パターン検出，クラス分類がある．

3.1 クラスタリング

クラスタリングは，与えられたデータの集合を複数のクラスタと呼ばれる集合に分類する手法である[2]．このとき，同一のクラスタに含まれるデータは，ある基準で互いに類似しており，かつ，異なるクラスタ同士のデータは類似していないものとなるよう，クラスタを作成する．

例えば，ソースコードに対する修正内容を要素とした集合に対し，ある修正に対して依存して修正される他の修正をまとめるようなクラスタを作成するようなクラスタリングを用いることにより，別の修正に対する修正漏れを防ぐことができるようになる[3]．

3.2 頻出パターン検出

頻出パターン検出とは，与えられたデータの集合の中に高い確率で存在している，特徴的な性質を発見する手法である (例えば[4])．特徴的な性質は，各データの中に含まれている場合や，複数のデータの相関事象として現われる場合がある．

例えば，与えられたソースコードから，外部の関数の呼び出し系列をあらかじめ抽出しておき,「同一の呼び出し系列が高い確率で存在している」ことを発見することによって，イディオムのように用いられる関数呼び出し順番を明らかにすることができる[5]．

3.3 クラス分類

クラス分類は，クラスタリングと同様，与えられたデータの集合をクラスとよばれる複数の集合に分類する手法である (例えば[6])．クラスタリングは「結果としていくつのクラスタが発見できるか」を事前に知ることができないが，クラス分類では，分析前に「いくつのクラスが存在するか」を事前に決定しているという点が異なっている．

例えば，ソースコードの集合が与えられた時,「不具合が含まれているのが明らかなソースコード」と「不具合が含まれていないことが明らかなソースコード」の 2 種類に分類するデータマイニング手法を用いることにより，新たなソースコードに対して，不具合が含まれているか，あるいは含まれていないか，を判定することができる[7]．

4. 研究の動向

リポジトリマイニングに関する研究報告は，ソフトウェア工学一般の会議等で発表されることが多い．しかしながら，リポジトリマイニングに関する国際会議が最近の動向を追うには最適であろう．

2004 年に開始された，ソフトウェアリポジトリのマイニングに関する国際会議 (Working Conference on Mining Software Repositories, MSR)[8] は，リポジトリマイニング関連の研究発表の場として有名である．開始当初から，ソフトウェア工学に関する国際会議 (International Conference on Software Engineer-

ing, ICSE)[9] の併設イベントとして開催されている．本年の MSR2009 は，ICSE2009 併設の国際会議として，2009 年 5 月 16,17 日にバンクーバーで開催された[10]．会議の概要については，参考文献[11] を参照されたい．本年の技術論文採択率は 3.8 倍と比較的高く，この分野において質の高い発表が行われている．

MSR2009 では，技術論文の発表は 5 つのセッションで構成されており，それぞれ Version control and infrastructure (4 篇), Defect Prediction (3 篇), Text analysis (2 篇), Topic mining (1 篇), Developers (3 篇) となっている．最初の Version control and infrastructure セッションでは，版管理ツール git を対象としたデータマイニング[12] や，MapReduce を用いたリポジトリマイニング手法[14] についての発表などが行われており，Defect Prediction セッションでは，コード規約違反とフォールトの相関関係[15] に関する発表などが行われた．MSR2009 の技術論文 13 篇のうち，この 2 セッションで半分の 7 篇が発表されており，分析対象となるリポジトリとして版管理ツールが注目されていること，リポジトリマイニングの応用事例として欠陥の検出や予測が重要であると考えられていることを伺い知ることができる．また最後の Developers セッションでは,「リポジトリを分析して開発者の情報を得る」研究について発表が行われており，例えば，開発者が作業履歴を記録する際の特徴分析[16] などが行われている．

この他にも MSR2009 では，ポスター発表や GNOME 環境のソースコードを対象とした分析データ発表など，多様な研究発表が行なわれており，リポジトリマイニングに関する研究が多岐に渡って活発に行われていることを知ることができる．来年の MSR2010 は ICSE2010 併設の会議としてケープタウンで行われる予定で，さらに今後の動向を知るには有用ではないかと思われる．また，過去の MSR については，2007 年までの予稿集が MSR の web ページ上で入手可能であるので，興味のある方は参考にされたい．

5. おわりに

本稿では，リポジトリマイニングの研究について，かなり大雑把であるが紹介を行った．

現在のようなリポジトリマイニングの研究は，歴史が比較的浅いものの，多くの分析対象と多様なデータマイニング手法により，幅広く奥行きのある研究がさかんに行なわれている．リポジトリマイニングを行うための分析基盤に関する研究など，今後もリポジトリマイニングに関する研究分野は，既存の要素技術をとりこみながら発展するのではないかと考えられる．将来研究が発展することにより，大量の開発データから自分の欲しい情報を「掘りあてる」ことがさらに容易になるのではないだろうか．

参 考 文 献

1) 福田剛志, 森本康彦, 森下真一, 徳山豪: データマイニングの最新動向：巨大データからの知識発見技術, 情報処理, Vol.37, No.7, pp.597–603 (1996).
2) Brian Everitt: Cluster Analysis, Edward Arnold, third edition (1993).
3) 早瀬康裕, 今枝誉明, 市井誠, 松下誠, 井上克郎: 潜在的意味解析手法を用いたソフトウェア変更情報のクラスタリング手法, 情報処理学会論文誌, Vol.48, No.10, pp.3352-3356 (2007).
4) Jian Pei, Jiawei Han, Behzad Mortazavi-asl, Helen Pinto, Qiming Chen, Umeshwar Dayal, Mei-chun Hsu: PrefixSpan: Mining Sequential Patterns by Prefix-Projected Growth, Proc. 17th International Conference on Data Engineering, pp.215–224 (2001).
5) 石尾隆, 伊達浩典, 三宅達也, 井上克郎: シーケンシャルパターンマイニングを用いた コーディングパターン抽出, 情報処理学会論文誌, Vol.50, No.2, pp.860–871 (2009).
6) Pedro Domingos and Michael Pazzani: On the Optimality of the Simple Bayesian Classifier under Zero-One Loss, Machine Learning, Vol.29, No.2–3, pp.103–130 (1997).
7) Osamu Mizuno and Tohru Kikuno: Prediction of Fault-Prone Software Modules Using a Generic Text Discriminator, IEICE Trans. on Information and Systems, E91-D(4), pp.888–896 (2008).
8) MSR Home Page, http://www.msrconf.org/ (2009).
9) ICSE Home Page, http://www.icse-conferences.org/ (2009).
10) MSR2009: 6th IEEE Working Conference on Mining Software Repositories, http://msr.uwaterloo.ca/msr2009/ (2009).
11) 横森励士，青山幹雄，井上克郎: 第 31 回ソフトウェア工学国際会議 (ICSE2009) 参加報告, 情報処理学会研究報告, Vol.2009-SE-165, No.10, pp.1-8 (2009).
12) Christian Bird, Peter C. Rigby, Earl T. Barr, David J. Hamilton, Daniel M. German, Prem Devanbu: The Promises and Perils of Mining Git, Proc. of MSR2009, pp.1–10 (2009).
13) J. Dean and S. Ghemawat: MapReduce: Simplified Data Processing on Large Clusters, Communications of ACM, Vol.51, No.1,

pp.107–113 (2008).
14) Weiyi Shang, Zhen Ming Jiang, Bram Adams, Ahmed E. Hassan: MapReduce as a General Framework to Support Research in Mining Software Repositories (MSR), Proc. of MSR2009, pp.21–30 (2009).
15) Cathal Boogerd and Leon Moonen: Evaluating the Relation Between Coding Standard Violations and FaultsWithin and Across Software Versions, Proc. of MSR2009, pp.41–50 (2009).
16) Walid Maalej and Hans-Jorg Happel: From Work to Word: How Do Software Developers Describe Their Work?, Proc. of MSR2009, pp.121–130 (2009).