

ソフトウェアリポジトリマイニング

コンピュータソフトウェア, Vol.30, No.2(2013), pp.52-65.
[解説論文]2012年10月15日受付

門田 暁人



伊原 彰紀



松本 健一



はじめに

- ソフトウェアリポジトリマイニング
(Mining Software Repositories, MSR)
- MSRという名の国際会議が2004年から開催
- MSRを研究テーマとする研究者はICSEなど**トップカンファレンス**で論文発表

MSRとは(1)

- ソフトウェアリポジトリ

- (狭義)

1. ソフトウェア構成管理システム
2. バグ管理システム
3. メーリングリストアーカイブ

- (広義)

多種多様なソフトウェア開発データ

(プロジェクト情報、開発工数、進捗管理、テスト管理など)

MSRとは(2)

- 「何を目的にマイニングするか」

新しい研究目的を見つけることがMSR研究の醍醐味

- 従来:

企業におけるソフトウェア開発支援

開発工数の予測, 生産性の要因分析, 品質評価・予測

- MSR:

OSS開発者やOSS利用者の支援

研究目的の例:コミッターを開発者から見つけ出す

バグが含まれるソースコードを推定する

MSRとは(3)

- 「どのようにマイニングするか」

非常に詳細なプロセス・プロダクトデータ(ノイズも多い)

取り出すデータはさまざまな種類がある

技術的に自由度の高いマイニングができる

クラスタリングや機械学習といった従来のデータマイニングに加えて他分野における新しい技術も取り入れる

MSR研究の魅力

- **企業**においても構成管理システムやバグ管理システムが使われている。

MSR分野の成果が**企業**にとって魅力的

- OSS開発のデータ自体が面白く、研究が多彩で、**研究のしがい**がある
- この分野の論文の引用回数が多くなっている。
“Who should fix this bug?”

MSRの研究テーマ

- MSR分野の代表的な研究テーマ
 - ソフトウェア品質
 - 開発プロセスの分析
 - 協調作業の分析
 - OSSを利用する開発者の支援
 - MSR研究者の支援

ソフトウェア品質

[バグモジュール予測]

- 多数のモジュールの中から**潜在バグを含むモジュール**を判別することが目的
- より少ないリソースでより多くのバグを検出することを目指している

[Bug localization(バグ位置の推定)]

- バグ管理システムに登録されたバグ報告を手掛かりにして、ソースコード中の**どの位置にバグが潜在しているか**を推定する研究

開発プロセスの分析

[変更の分析]

- OSSの開発はソースコードの変更の繰り返しのプロセス
- **どのような変更, といった種類の変更**
- バグとの関係や変更の粒度を分析

[バグ修正プロセスの分析]

- バグ修正のプロセス
バグ報告, バグ承認, 修正担当者決定, バグ修正, 修正確認
- 修正が**早いバグ**と**遅いバグ**のプロセスの違い

協調作業の分析

[コミッター推薦]

- すぐれた開発者のみにコミットする権限をあたえられる
- 大規模プロジェクトでは1万人を超える一般開発者
- 機能拡張や不具合修正を分析し、コミッターを推薦

[コミュニケーション分析]

- 人と人との関係をネットワーク構造と捉えて分析
- OSSのコミュニケーション分析に用いる
- 開発者, ユーザ, バグ報告者の3者間を橋渡しする人物
(コーディネータ)

OSSを利用する開発者の支援

[ライブラリの使用状況の分析]

- どのバージョンのプロダクトを利用すべきか
- どのバージョンが最も使われているか分析
- ユーザ数の多いバージョンほど機能と信頼性のバランス

[ソフトウェア部品検索]

- 利用者が必要なソフトウェア部品を探すのは容易でない
- ソフトウェア部品検索エンジンが多数開発されている
- キーワードベース, ソフトウェア部品の入出力ベース

MSR研究者の支援

[リポジトリ間のリンク付け]

- 異なるリポジトリに含まれる関連情報のリンク付け
- バグ管理システム上のバグ修正情報と構成管理システムのソースコードのリンク付けが盛ん

[名寄せ]

- OSS開発者は、リポジトリや作業内容に応じて、異なるアカウント名やEmailアドレスを使う
- 同一人物を特定してマージする

マイニングの具体例

リポジトリからデータを**取得**し、
必要な情報を**抽出**して**整形**する作業が必要不可欠

- データ取得と整形作業の例

データ取得方法

- マイニング目的決定後、どのデータを取得するか
- 例: bug localization
欠陥を含むソースコードを探す手掛かり
→バグ管理システムから**バグを再現する情報**
- 例: 開発者の協調, 共同作業の関係を明らかにする
→メールアーカイブをダウンロードし、メールの**ヘッダ情報**や**内容**を抽出する

データ整形方法

- データを分析が容易になるフォーマットに書き換える
- 名寄せアルゴリズム
類似するアカウント, Emailアドレスが同一人物のものか否かを判断するためのアルゴリズム
現在提案されている方法の中で最も信頼性がある
- ほかに
データセットの作成技術, 形態素解析技術

MSRの発展

- MSR vision 2020(カナダ)
 - Large/Small-Scale mining:
大規模だけでなく小規模プロジェクトを対象とした研究も重要ではないのかという議論
 - IR(Information retrieval):
ソフトウェア工学に限らず, ビッグデータの中から目的に合致したものを取り出す
 - Github:
ソフトウェア構成管理システムの一種
従来のリポジトリを全員で共有するのとは異なり、個人で保有できる

まとめ

- MSR研究, 主な研究テーマ, マイニング具体例, MSR分野の今後の発展
- マイニング対象を単一バージョンから複数バージョンに広げる
- 異なるプロダクト間の比較を行い, マイニングにより得られた知見の一般性を評価する