

『ソフトウェア工学の実証的アプローチ』 シリーズ第5回

ソフトウェアリポジトリ マイニング

門田 暁人 伊原 彰紀 松本 健一

コンピュータソフトウェア, VOL.30, NO2(2013), P.52-65.

学籍番号 60206299

氏名 吉富 楓雅

目次

1. リポジトリの定義は何か？
2. 研究のテーマ例
3. マイニングの具体例
4. MSRの発展
5. まとめ

ソフトウェアリポジトリマイニングとは？



「リポジトリ」から「マイニング」する
宝 **発掘**

いわゆる・・・鉱山から宝を探し出す分野

リポジトリの定義は何か？



狭義（もともとは）

3つのデータソース

1. ソフトウェア構成管理システム（CVS, Subversion, Gitなど）



2. バグ管理システム（Bugzilla, Trac, Redmine）



3. メーリングリストアーカイブ

OSS開発の広がりとともにMSRの分野は発展してきた

広義(もっと広く)

実際は対象は3つだけではない

多種多様なソフトウェア開発データ

例えば...

- ・プロジェクト情報
- ・開発工数データ
- ・進捗管理データ
- ・テスト管理データなど

MSRの研究には2つの側面

1. 「何を目的にマイニングするか」
2. 「どのようにマイニングするか」

「何を目的にマイニングするか」

従来のソフトウェア工学・・・企業におけるソフトウェア開発支援が目的

MSRでは・・・主なデータソースはOSS開発プロジェクトであることから、OSS開発者やOSS利用者の支援を目的

研究テーマ：非常にバラエティに富んでいる。

「どのようにマイニングするか」

マイニングの技術的な自由度が高い点が大きな魅力

他分野における新しい技術の取り入れ

例えば・・・

協調フィルタリング,

シーケンシャルパターンフィルタリング

テキストマイニング など

MSR研究の魅力

従来・・・「実際の開発現場では使えない」という
批判を受けていた

現在・・・OSSの開発・利用支援は企業にとって
必要なものとなっている

ここからは研究のテーマ例

EXAMPLE

MSRの研究テーマ

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

Bug localization (バグ位置の推定)

エラーパターンの検出

バグ修正時間の予測

バグ修正誤りの分析

開発プロセスの分析

変更の分析

変更の予測

同時変更コード(ロジカルカップリング)

Patch(パッチ)投稿・受理プロセスの分析

バグ修正プロセスの分析

コードクローン分析

協調作業の分析

エキスパート推薦

コミッター推薦

コミュニケーション分析

Openness(開放性)の分析

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

ソフトウェア部品検索

ソフトウェアの分類

ライセンスの分析

コーディングパターン検出

MSR研究者の支援

リポジトリ間のリンク付け

名寄せ

MSRのサービス化

マイニングの具体例



データ取得方法

MSR分野では、マイニング目的を決定後、どのリポジトリからどのようなデータを取得するかを検討。

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
4. データ整形方法
5. データ取得後のプロセス

ソフトウェア構成管理システム

多くの研究では、プロジェクトが管理するソフトウェア構成管理システムに保存されている開発記録を、Webを介して取得している。

バグ管理システム

開発者あるいはユーザから報告されたバグ情報をバグ表に記録し、開発者がバグ修正の進捗を記録しながら、修正の様子を共有するシステム。

メーリングリストアーカイブ

メーリングリストの使用

オープンソース開発プロジェクトにおいて、
課題、開発計画等をプロジェクト全体に認知させることを目的

データ整形方法

各ソフトウェアリポジトリからデータをダウンロードした後に、分析を容易にするフォーマットに書き換える必要性

・・・開発者はリポジトリ、開発環境、実施する作業の役割に応じてアカウント名やemailアドレスを変えることがある。

“名寄せ”アルゴリズムの利用

“名寄せ”アルゴリズム

Christian Birdが提案した、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.

fullname, first name, last name, emailアドレスのアカウント名の類似度合いをレーベンシュタイン距離を用いて算出

→閾値(いきち)以上であれば同一人物と判断

“名寄せ”アルゴリズム

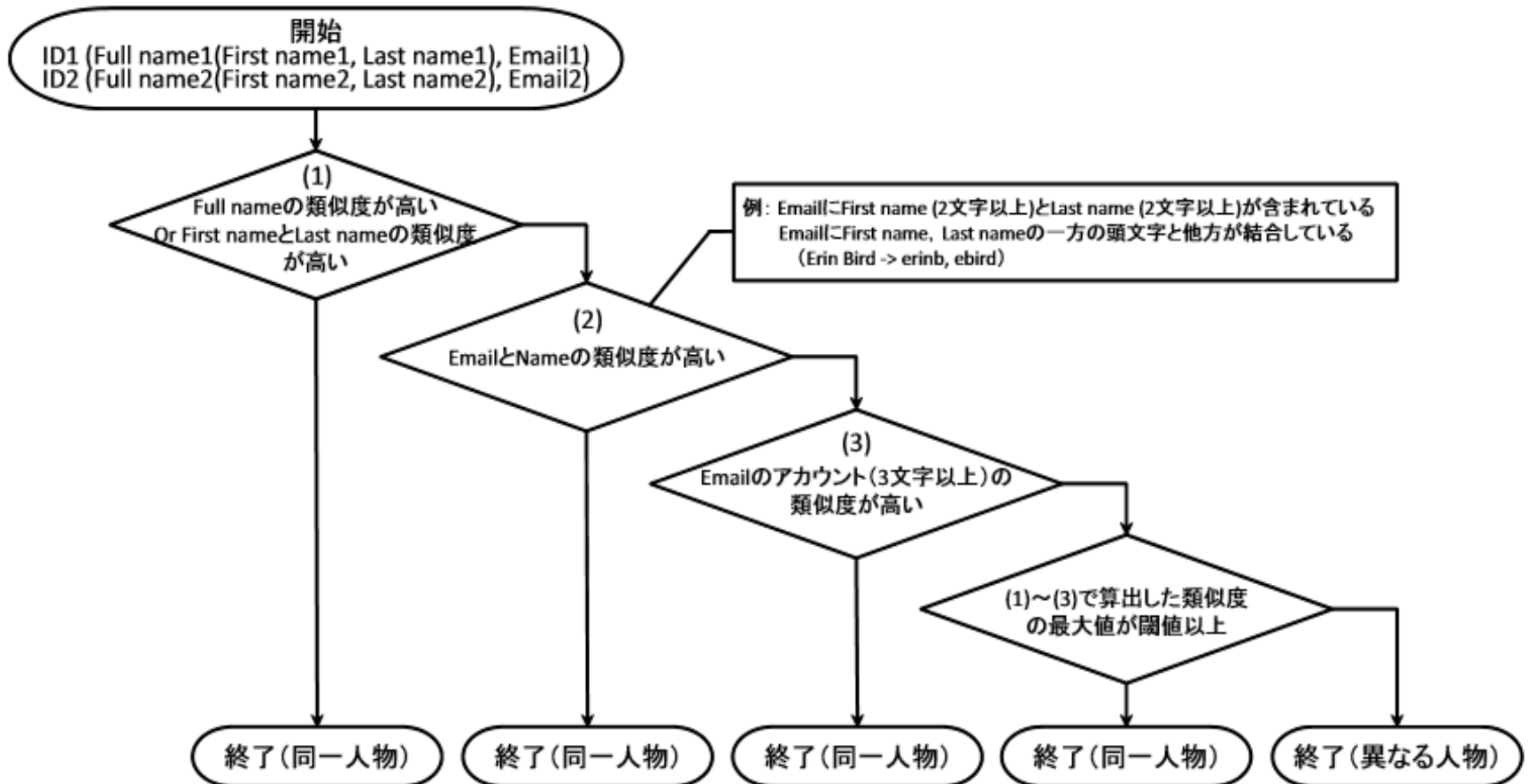


図3 “名寄せ” アルゴリズム

データ取得後のプロセス

データ処理プロセス (Processing) と分析プロセス (Analysis) が続く。

- ・データ処理プロセス

ソースコードやテキストデータなどデータの種類ごとに異なる

- ・分析プロセス

分類, 回帰, ルール抽出, クラスタリングの為の様々なデータマイニング技術が用いられる

MSRの発展



MSRの発展

将来のMSR研究者の育成，およびMSR分野の発展に向けた活動が積極的に行われている。

例：2012年8月 MSR vision 2020 カナダ

2020年に向けて何にチャレンジしていくかを議論

＜議題となったMSR研究トピックの例＞

Large/Small-Scale mining

IR(Information retrieval)

Github

まとめ

- ・MSR研究の魅力に焦点を当て、研究テーマを紹介
- ・マイニングの具体例とMSRの発展について

MSR研究は今後ますます盛んになっていく