

# 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

大平研

60216114 作島大智

# 目次

- ソフトウェアリポジトリマイニング
- Software Repositoriesとは
- MSRの研究目的
- 研究における背景
- 研究テーマ
- データ取得の一連の流れ
- MSRの発展
- まとめ

# ソフトウェアリポジトリマイニング

- MSR(Mining Software Repositories)
- 2004年より国際会議が開催
- 投稿者・参加者が年々増加
  - 難関で魅力がある

# MSRとは

「ソフトウェアリポジトリ」という鉱山から  
マイニング（宝探し）をする研究分野のこと



# ソフトウェアリポジトリ(狭義)とは

主に以下の**3**つのデータソースを指す

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



蓄積・公開されており研究の題材として  
利用可能な**OSS**がデータソース



**OSS開発**と共に発展

# ソフトウェアリポジトリ(広義)とは

多種多様なソフトウェア開発データを含む  
従来のソフトウェア工学で用いられてきた  
データソースを指す

企業との共同研究によって開発工数の削減や  
品質向上に応用

# MSRの研究目的

大きくこの2つ

- 「何を目的にマイニングするか」
- 「どのようにマイニングするか」

# 何を目的に研究するのか

新しい研究目的を見つけ出すことが

MSR研究の醍醐味

→ 研究内容の自由度が高く 多様性が広がる

|      | 従来の<br>ソフトウェア工学                | MSR                       |
|------|--------------------------------|---------------------------|
| 支援先  | 企業における<br>ソフトウェア開発支援           | OSS 開発者・利用者支援             |
| 開発内容 | 開発工数の予測<br>生産性の要因分析<br>品質評価・予測 | コミッター採用システム<br>バグ原因推測システム |

# どのようにマイニングするか

- リポジトリ自体に様々な種類のデータを含む  
→ 技術的な自由度が高い

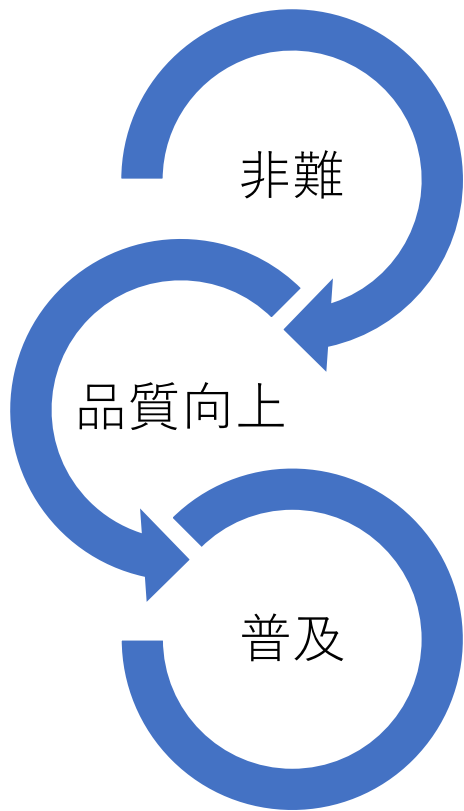


他分野における新しい技術を取り入れることが  
盛んで最先端の技術を試すことができる



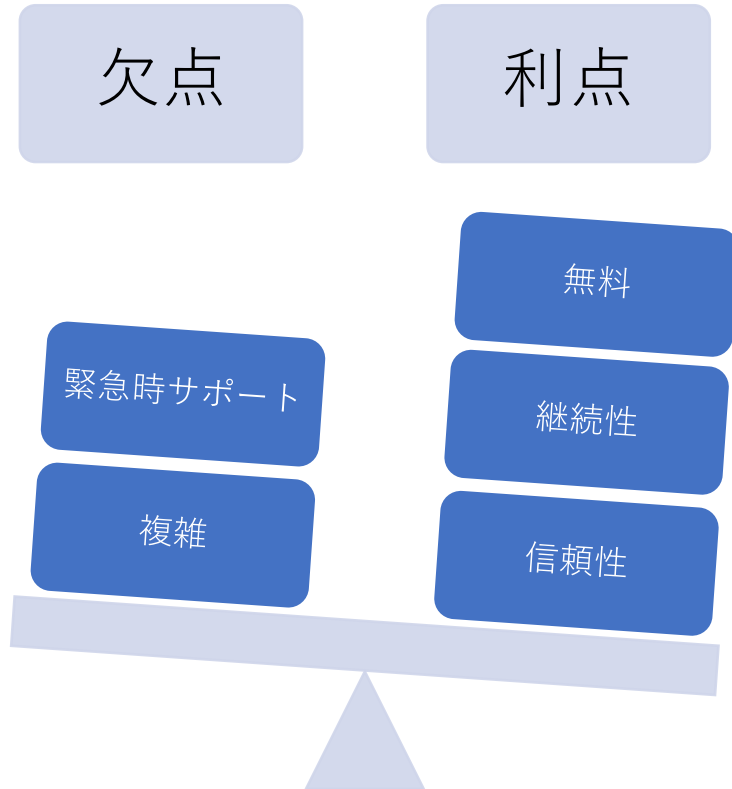
研究者にとって良い環境が整っている

# 研究における背景



- OSS開発がソフトウェア業界で重要な位置を占めるようになった
- 「実際の開発現場では使えない」という非難
- LinuxやAndroidが普及し企業にとっても必要なものに

# OSSについての調査



- OSS利用時の  
多数の課題
- MSR研究の成果の  
活用が期待される

# 研究者の見解

OSS開発データ自体が興味深く研究テーマが多彩

MSR分野の研究者の急増により研究者同士互いに論文を引用しあうことにより更なる発展が期待



技術的進歩の見込みあり

# 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

# 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)
- 開発者の誤りパターン検出
- バグ修正時間の予測
- バグ修正誤りの分析

# 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

## 開発プロセスの分析

- 変更の分析・予測
- 同時変更コードの分析
- Patch投稿受理システム
- バグ修正プロセス
- コードクローン分析

# 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

## 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

# 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

## OSSを利用する開発者

- ライブラリの使用状況の分析
- ソフトウェア部品検索
- ソフトウェアの分析
- ライセンスの分析
- コーディングパターン検出

# 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

## 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

# データ取得の一連の流れ

- どんな管理システムを使うのか
- 情報をどのようにして共有するのか
- データをどのように情報にするのか
- データ取得後のプロセスはどのようなものか

# ソフトウェア管理システム

コミットログから正規表現などを用いて  
目的に応じて必要な情報を抽出する

|        | Subversion | Git |
|--------|------------|-----|
| 編集形態   | 中央集中型      | 分散型 |
| 編集可能人数 | 1人         | 複数人 |

# バグ管理システム

**Bug 346882 - unable to update tickets with Bugzilla**

**Status:** RESOLVED FIXED

**Product:** Mylyn Tasks  
**Component:** Bugzilla  
**Version:** unspecified  
**Platform:** PC Linux

**Importance:** P1 major with 2 votes (vote)  
**Target Milestone:** 3.6.5  
**Assigned To:** Frank Becker

**基本情報**

**Comment 2**  
Andrew R. Thompson 2011-05-23 15:56:43 EDT  
Synchronize works in that the ticket is updated with any change that I make through the bugzilla web interface. But the behavior is the same after synchronization when I attempt to modify that ticket through the mylyn interface.  
I'm completely open to the idea that this is a server side configuration setting. But am stumped on how to debug or get enough information to forward to the team that maintains our bugzilla instance.

**Comment 3**  
Steffen Pingel 2011-05-23 16:40:47 EDT  
Frank, any chance this could be related to the domain setting in Bugzilla (don't remember the bug but I recall that there were problems when the configured domain didn't match the actual domain of the server's address)?

**コメント情報**

|                |                         |                  |                                                                 |          |
|----------------|-------------------------|------------------|-----------------------------------------------------------------|----------|
| Frank          | 2011-10-28 14:51:13 EDT | Status           | NEW                                                             | ASSIGNED |
|                |                         | Assignee         | mylyn-triaged                                                   | Frank    |
| Frank          | 2011-10-28 15:15:07 EDT | Summary          | [upstream] unable to update tickets with Bugzilla 4.0 and Mylyn |          |
| steffen.pingel | 2011-10-28 17:21:35 EDT | Target Milestone | 3.7                                                             | 3.6.4    |
| zerobox        | 2011-11-02 12:41:14 EDT | CC               |                                                                 | zerobox  |
| Frank          | 2011-11-05 12:34:12 EDT | Status           | ASSIGNED                                                        | RESOLVED |
|                |                         | Resolution       | ---                                                             | FIXED    |
| steffen.pingel | 2011-11-06              | Status           | RESOLVED                                                        | REOPENED |

**進捗情報**

Webを介して  
誰でも閲覧可能

図 2 バグ管理システム Bugzilla のバグ票

# メーリングリスト

- 大規模ソフトウェア開発では複数の開発者が情報を共有する必要がある
- パッチをメールで添付してプロジェクトで確認される場合があるので添付ファイルも研究対象

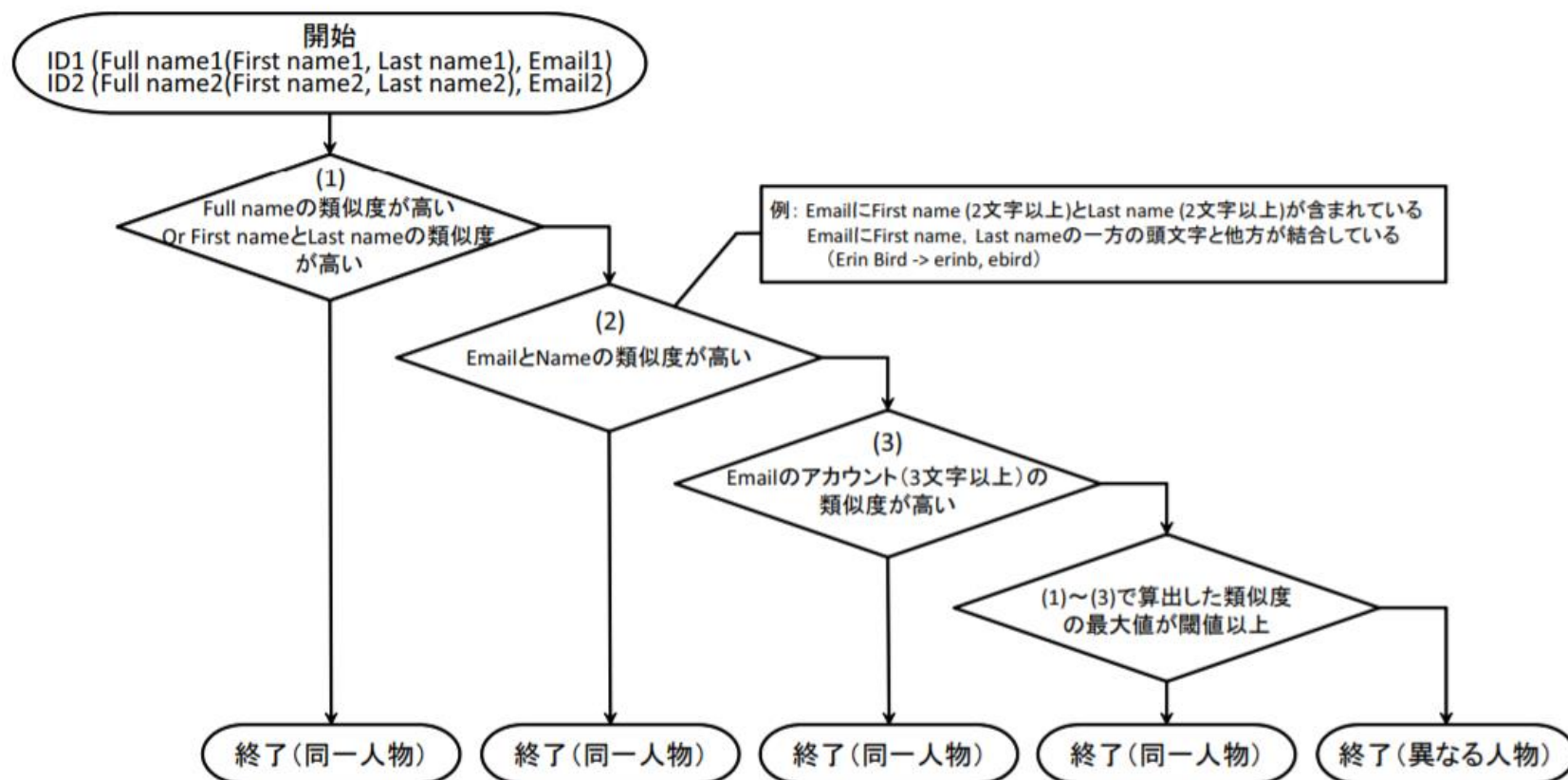
# データ整形方法

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



名寄せアルゴリズム

# 名寄せアルゴリズム



# MSRの発展

MSR分野の発展に向けて効率的な実験や質の高い論文の執筆を支援することをも含む国際会議が多く開催されている

MSR visionで議題となったMSR研究トピック

- Large/Small-Scale mining
- IR(Information retrieval)
- GitHub

# まとめ

- ソフトウェアリポジトリの研究分野を活用した応用分野は多岐にわたる
- MSRは主にOSS関係者をターゲットにした研究分野
- マイニングにより得られた知見の一般性を評価することが重要である