不確かさ分析用公開データベースの作成に向けて

村岡 北斗[1,a)]　亀井 靖高[1,b)]　佐藤 亮介[1,c)]　鵜林 尚靖[1,d)]

概要：ソフトウェア開発において「不確かさ」は避けることができない．我々はこの不確かさの解決に向けた分析をサポートするために，不確かさについての公開データベースの作成を提案する．本論文では，公開データベースに登録する最初の不確かさ検出方法と，公開データベースの運用方法について説明する．本セッションでは，リポジトリマイニング研究を実施する上で重要となるデータ品質について議論したい．

1. はじめに

ソフトウェア開発において，「不確かさ」とは避けられないものである [2]．ソフトウェア開発における不確かさとは，要求，設計，実装方法やテストなど様々な開発工程に現れる不確実性についての問題である．不確かさには，多種多様な問題が存在する（例：要求や仕様が曖昧なため2パターンの実装をした，昔書いたコードがどういう使い方をするかわからなくなってしまった，などが存在する）．Perez-Palacin らの研究 [4] では，場所，レベル，性質といった 3 つの観点から不確かさの分類を定義している (表 1)．このような不確かさは，開発者にとって扱いにくい問題であり，一時的な措置をとることも多く，バグやソースコードの煩雑化の原因となる．近年，ソフトウェア工学において，不確かさを包容したソフトウェア開発は重要な研究課題の 1 つとされている [1]．

不確かさを包容したソフトウェア開発の参考として，まずは実際のプロジェクトで発生している不確かさがどのように発生し，対応されているかを知りたい．そのために，不確かさについて機械的な分析を行う際に，不確かさを開発履歴から検出する手法について妥当性を持たせることが難しいという問題が存在する．この問題は，不確かさの基準が曖昧で，不確かさか否かの判断が見た人の主観によって変わってしまうことが原因である．Perez-Palacin らの分類も，実際に分類するための基準が明確に設けられているわけではないため，彼らの分類について不確かさを判別した時に，人によって異なるレベルの不確かさと判断してしまうことがある．

[1] 九州大学 大学院システム情報科学府
[a)] muraoka@posl.ait.kyushu-u.ac.jp
[b)] kamei@ait.kyushu-u.ac.jp
[c)] sato@ait.kyushu-u.ac.jp
[d)] ubayashi@ait.kyushu-u.ac.jp

表 1　Perez-Palacin らによる不確かさの分類

観点	性質	性質の説明
場所	コンテキスト	環境に関する不確かさ
	モデル構造	モデル自体の構造に現れる不確かさ
	入力パラメータ	モデルへの入力に関する不確かさ
レベル	レベル 0	確定している知識
	レベル 1	知識の不足を認知している状態. 既知の不確かさ.
	レベル 2	知識の不足を認知できていない状態. 未知の不確かさ.
	レベル 3	不確かさを認知するプロセス自体が 不足している状態.
	レベル 4	不確かさのレベル自体が不確か.
性質	認識的	十分なデータや知識が無いために 発生する不確かさ
	偶発的	物理現象等の確率的不確かさ

不確かさの基準が曖昧な問題を解決するために，不確かさについての評価を共有する環境が必要である．バグ予測やリファクタリングの分野では，公開された分析用のデータセットが存在し，評価を共有することができるが，不確かさの分野には存在しない．データセットを互いに評価する環境があると何を基準に不確かさを判断するかの手がかりにできる．

本論文では，不確かさについて分析するための公開データベースを提案する．データベースの作成は，Palomba らがコードのにおいについての分析用公開データセットを作成した手法 [3] を参考とする．まず，不確かさを機械的に特定し，データベースに登録する初期データセットとする．次に，機械的に特定した不確かさの候補内容を手作業で正しいかどうか判断する．そして，第三者がデータセットを評価するための環境について説明し，公開データベースの活用方法について述べる．最後に，今後の予定と本セッションで議論したい内容について説明する．

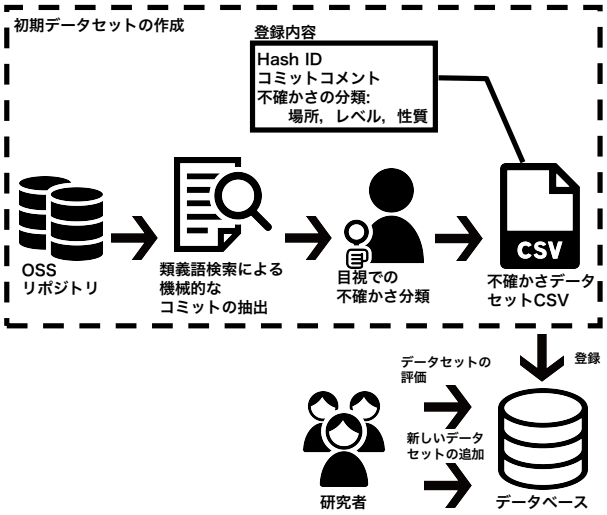

図 1　初期データセット作成手順

2. 初期データセットの作成

本章では，公開データベースに初期データセットとして登録するデータセットの作成手順について述べる．図 1 の点線枠内にデータセット作成の一連の流れを示す．

2.1 対象プロジェクト

GitHub にて公開されている OSS プロジェクトを対象とする．ソフトウェア開発を目的とするプロジェクトを対象としたいので，開発人数が 10 人以上かつリポジトリがコピーされた回数が 10 回以上のプロジェクトを使用する．

2.2 不確かさ検出方法

対象となる OSS リポジトリのうち，コミットメッセージに不確かさの類義語が含まれるコミットを機械的に検出する．uncertainty に関する類義語を Oxford American Writer's Thesaurus から参照した．

2.3 不確かさの手作業での分類

対象プロジェクトから類義語を用いて検出したコミットに対して，Perez-Palacin らの分類 (表 1) にしたがって，目視による分類を行う．分類から主観を取り除くために，複数人のプログラミング経験がある学生で分類を行う．

3. 公開データベースの活用方法

本章では，公開データベースへのデータセット登録，評価，利用方法について説明する．

3.1 データベースの登録内容

データベースには，不確かさについて集めた分析用のデータセットが登録できるようにする．各データセットには，データセット名，著者の一覧，データセット作成手順，検出結果の csv ファイルを登録する．検出結果の csv ファイルは要素として不確かさを持つコミットの hash id，コミットコメント，不確かさの存在する場所，レベル，性質の項目を持つ．

3.2 データセットの評価方法

公開データベースは，不確かさ分析用のデータセットを集めて提供するだけのものではない．図 1 にも表しているように，登録されているデータセットを誰でも評価できるようにすることで，データセットの品質を向上させる目的がある．登録されている不確かさ単位で，不確かさの分類が正しいと思うか否かを投票できる機能を用意し，投票結果を表示することを考えている．また，正しくないと判断した際にはその根拠をコメントとして報告が可能にする．

データセット単位でも，登録された内容のうち何割が不確かさであるかを表示するようにする．これは，データセットを作成した手法が不確かさを分類する際にどの程度有効なものかを表すためのものである．

4. おわりに

今後の予定として，まず，今回提案したデータベースを作成する．初期データセットを作成し，登録することで，データセットを評価できる環境を完成させたいと考えている．その後登録したデータセットの一部に対して実際に評価を行うことで，データセットの作成手法がどの程度有効なものであるかを判断する予定である．本セッションでは，データセット作成時の手作業での分類方法とデータセット評価方法の妥当性について議論したい．また，Palomba ら [3] の分類方法の妥当性，データセットの有用性に関する議論も行いたい．

謝辞

本研究は，JP26240007 による助成を受けた．

参考文献

[1] Sebastian Elbaum and David S Rosenblum. Known unknowns: Testing in the presence of uncertainty. In *Proceedings of the 22nd International Symposium on Foundations of Software Engineering*, pp. 833–836, 2014.

[2] David Garlan. Software engineering in an uncertain world. In *Proceedings of the FSE/SDP workshop on Future of software engineering research*, pp. 125–128, 2010.

[3] Fabio Palomba, Dario Di Nucci, Michele Tufano, Gabriele Bavota, Rocco Oliveto, Denys Poshyvanyk, and Andrea De Lucia. Landfill: An open dataset of code smells with public evaluation. In *Proceedings of Mining Software Repositories (MSR), 2015 IEEE/ACM 12th Working Conference on*, pp. 482–485, 2015.

[4] Diego Perez-Palacin and Raffaela Mirandola. Uncertainties in the modeling of self-adaptive systems: A taxonomy and an example of availability evaluation. In *Proceedings of the 5th ACM/SPEC International Conference on Performance Engineering*, pp. 3–14, 2014.