

# ソフトウェア開発における不確かさに着目した OSS コミットログ解析

村本 大起<sup>1,a)</sup> 江 冠達<sup>1,b)</sup> 村岡 北斗<sup>1,c)</sup> 深町 拓也<sup>1,d)</sup> 鷗林 尚靖<sup>1,e)</sup> 亀井 靖高<sup>1,f)</sup>  
佐藤 亮介<sup>1,g)</sup>

**概要:** 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つである。本論文では、実際のソフトウェア開発プロジェクトにおいて、どのような不確かさが問題となっているかを明らかにする。オープンソースソフトウェアプロジェクトの開発履歴のコミットメッセージを対象とし、不確かさを表すキーワードを含むコミットメッセージにはどのような特徴があるか分析を行った。結果として、それぞれのキーワードから不確かさに関わる興味深い特徴が分かった。

## 1. はじめに

ソフトウェア工学において、一般的なソフトウェア開発プロセスは大きく3つに分かれる。各段階にはっきりしない要求、定まらない設計候補、未決定のアルゴリズムなど曖昧なところに不確かさが潜んでいる。このような不確かさはソフトウェアのバグや不安定な状態が発生する原因となっている。したがって、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つである。

Garlan は、不確かさという観点から将来のソフトウェア工学について論じた [8]。彼は、「コンピュータ環境が予測可能であり原則として完全に明記でき、また、そういった環境下で動くシステムは障害が無いように設計されている」と主張しており、ソフトウェア工学の研究は行われている」と主張しており、ソフトウェア工学の領域に不確かさを包容させなければならないと述べている。

ソフトウェア開発における不確かさには3つのタイプ (Known Knowns, Known Unknowns, Unknown Unknowns) がある [2]。Known Knowns タイプは、不確かさが存在しない開発であり、多くの伝統的なソフトウェア工学の研究において研究が行われている。Known Unknowns タイプは、ソフトウェア開発プロセス中に不確かな問題が存在する開発である。しかし、Known Unknowns タイプ

では、この不確かな問題は開発者や消費者といったステークホルダー間で認識、共有されている。例えば、複数のどれが選ばれるかわからない要求がこのタイプの不確かさに当たる。一方で、Unknown Unknowns タイプは、何が不確かであるかわからないケースを指す。つまり、このタイプを解決することは難しく、どのような問題が発生するかが予測不可能である。

本研究では、実際のソフトウェア開発プロジェクトにおいて、どのような不確かさが問題となっているかを明らかにする。実際に存在している問題を明確にすることで、ソフトウェア開発を助けることができると考えている。複数オープンソースソフトウェア (OSS) プロジェクトの開発履歴のコミットメッセージを分析することにより、それぞれ不確かさを表すキーワードを含んだコミットの不確かさに関わる傾向を分析する。

以下、第2章では不確かさに関する既存研究を紹介し、その問題点を基に第3章で研究課題について述べる。第4章ではそれぞれ複数の OSS プロジェクトを対象とした不確かさに関する分析を行い、その結果を示す。第5章で妥当性への脅威について述べた後、第6章でまとめとする

## 2. ソフトウェア開発における不確かさ

本章では、不確かさに関する既存研究の最新動向を紹介する。その後、既存研究の問題点について検討する。

### 2.1 不確かさに関する既存研究

近年、不確かさは研究者に非常に注目されている。不確かさに関するテーマとしては、ゴールモデリング、UML

<sup>1</sup> 九州大学

a) muramoto@posl.ait.kyushu-u.ac.jp

b) jiang@posl.ait.kyushu-u.ac.jp

c) muraoka@posl.ait.kyushu-u.ac.jp

d) fukamachi@posl.ait.kyushu-u.ac.jp

e) ubayashi@posl.ait.kyushu-u.ac.jp

f) kamei@posl.ait.kyushu-u.ac.jp

g) sato@posl.ait.kyushu-u.ac.jp

モデリング, ソフトウェアアーキテクチャ, モデル変換, テスト, 検証, パフォーマンス工学, 自己適応システムなど多岐にわたる。

代表的な不確かさの研究として, Partial model を用いて Known Unknowns タイプの不確かさを表現する手法が Famelis らによって提案されている [5, 6]. Partial model は要求モデルにおける不確かさの特定, 不確かさを減らすことによるモデルの改善, 不確かさを含むモデル間の関係のトレーサビリティへの意味の付与, そして関係するモデル間での不確かさを減らす変更の波及などの分野で用いられている. Esfahani らはソフトウェアアーキテクチャの分野においての不確かさについて研究を行っている [3, 4]. Letier らは要求やアーキテクチャにおける不確かさがリスクへ与える影響や, 不確かさを減らすことの価値を評価する手助けとなる手法を提案している [9]. Elbaum と Rosenblum は, 不確かさがソフトウェアテストにどのように影響するかという調査を行なっている [2]. その他にも, 自己適応システムにおける不確かさの研究や, 不確かさを含んだ状態での性能や信頼性の分析なども多く行われている。

## 2.2 不確かさの分類

本節では, ソフトウェア工学における代表的な不確かさの分類法を紹介する。

### 不確かさの定義

Perez-Palacin と Mirandola は自己適応システムにおける不確かさについてシステマチックレビューを行い [13], 「最もよく用いられる不確かさの定義は, 偶然や確率的な不確かさと呼ばれる自然現象の変動と, 認識や既知知識の不確かさといったプロセスに関する知識によるものである。」とまとめている。

### 不確かさの三次元分類

さらに, Perez-Palacin らは Walker らの研究 [16] に従って, 位置 (Location), レベル (Level), 性質 (Nature) から構成される三次元分類 (three-dimension classification) を提案した。

位置では, 文脈 (Context), 構造 (Structural), 入力パラメータ (Input parameters) の3つに不確かさが分類される。文脈, 構造, 入力パラメータは, それぞれ周りの環境の違いによって生じる不確かさ, モデル自体の構造の不確かさ, モデルへの入力データによる不確かさ, というようにどの部分に現れる不確かさであるかを指し示す。

レベルは, 4段階 (1st/2nd/3rd/4th) で順序付けられている。1段階目の不確かさは, 対象は何かしらの知識を欠落しているが, 開発者はその欠落について認識している状態 (Known Unknowns に相当) である。2段階目の不確かさは, 知識を欠落していることを認識できていない状態 (UnKnown Unknowns に相当) である。3段階目の不確か

表 1 Perez-Palacin らによる不確かさの要因

Source of Uncertainty	Classification	
	Location	Nature
Simplifying assumptions	Structural/context	Epistemic
Model drift	Structural	Epistemic
Noise in sensing	Input parameter	Epistemic/Aleatory
Future parameters value	Input parameter	Epistemic
Human in the loop	Context	Epistemic/Aleatory
Objectives	Input parameter/context	Epistemic
Decentralization	Context/structural	Epistemic
Execution context/Mobility	context/structural/input parameters	Epistemic
Cyber-physical system	context/structural/input parameters	Epistemic
Automatic learning	Structural/Input parameters	Epistemic/Aleatory
Rapid evolution	Structural/Input parameters	Epistemic
Granularity of models	Context/structural	Epistemic
Different sources of information	Input parameter	Epistemic/Aleatory

さは, 認識できていないということを理解するためのプロセスが欠落している状態である。4段階目の不確かさは, ここで定義した不確かさの段階そのもの (メタレベル) が不確かであるという状態である。

性質は, 偶然 (Aleatory) と認識 (Epistemic) とに分類される。偶然は偶発的な事象によりもたらされる不確かさ, 認識は得られるデータの不完全さ, またはデータ認識の齟齬などによる不確かさを表す。

また, Perez-Palacin らは不確かさの要因 (Source of Uncertainty) についても表 1 のように分類している。それぞれ不確かさの要因の出現位置 (Location) と性質 (Nature) が明記されている。

例えば, ステークホルダーからの不確かな要求は不確かさの要因としては Human in the loop とされ, 三次元分類では, 位置は Context であり, レベルは 1st (Known Unknowns), 性質は認識 (ステークホルダーの将来の決定が不確か) と分類される。三次元分類の主な対象は自己適応システムであるものの, 他のアプリケーションドメインに対しても適用することができる。

## 3. 研究課題

最新の研究において不確かさの様々な側面を扱っているものの, 不確かさに関する実証研究は無い。また, 残念ながら実際のプロジェクトにおいて, どのような種類の不確かさが発生するのかについて研究は行われていない。

実際に発生している不確かさの問題について明らかにすることで, ソフトウェア開発を行っていく上で有用な知識が得られるのではないかと考えられる。例えば, どのような内容には不確かさが多いのか, また, 要求, 設計, 実装それぞれのどのようなシーンで不確かさは現れるのか知ることで, 不確かさの問題について対応しやすくなる。

そこで実際に存在するプロジェクトの不確かさには, どのようなものが存在しているのかを解明するために本研究を行う。今回は, 複数の OSS プロジェクトの開発履歴を対象として分析する。本研究では, 開発履歴の情報から不確かさをプロジェクトの中から抽出する手法として, コミットメッセージを対象にテキスト分析を行う。このように,

表 2 不確かさを表すキーワード

ambiguous	faltering	might	undecided
ambivalent	fluctuating	obscure	undetermined
arcane	fuzzy	of two minds	unforeseeable
blowing hot and cold	hesitant	open to question	unknown
chancy	in doubt	probably	unpredictable
changeable	in the balance	risky	unreliable
debatable	incalculable	tentative	unsettled
dicey	informal iffy	todo	unsure
doubtful	irregular	unascertainable	up in the air
dubious	irresolute	unclear	vacillating
erratic	may	unconfident	vague

テキスト分析によって不確かさの特定方法を提案し、結果に対して目視で調査を行うことで実際のプロジェクトにどのような不確かさが存在するのかを明らかにする。

## 4. 不確かさを含むコミットメッセージの傾向分析

本研究では、不確かさを表すキーワードを含むコミットメッセージの特徴を分析するためテキスト分析を行う。既存研究 [12] によりテキストの特徴量を機械的に抽出することが可能であるため、この既存研究を用いることで多くの OSS プロジェクトのコミットメッセージを用いて分析する。

### 4.1 データセット

本研究ではコミット単位のバージョン管理システムである Git をベースとしたソースコード共有サービスである GitHub\*1 から、OSS プロジェクトを入手した。一般的な特徴を得るため、GitHub から無作為に 1,444 個のプロジェクトを選出し、データセットとした。

また、コミットには実際のソースコードや変更の差分など多くの情報が存在するが、本研究では、分析する手法として tf-idf [14] を用いたテキスト分析を扱うため、開発者たちがバージョン管理する際の指標とするコミットメッセージだけを対象として分析を行う。コミットメッセージだけを利用した分析の妥当性については後の 5 章で述べる。全てのコミットメッセージの中から、不確かさを表すキーワードを含んでいるコミットメッセージを抽出して分析対象とする。

不確かさを表すキーワードとしては *Oxford American writer's thesaurus* [10] という代表的なシソーラス書籍から *Uncertainty* に関わる類義語を参照し決定した (表 2)。

### 4.2 アプローチ

本節は不確かさの特徴を分析するためのアプローチを示す。本アプローチは 3 つのステップに分ける。ステップ 1: 不確かさを表すキーワードを含むコミットメッセージから、tf-idf 値 [14] によってキーワード毎の特徴単語を抽

出する。ステップ 2: ステップ 1 で抽出した特徴単語から名詞と動詞を抽出する。ステップ 3: ステップ 2 で抽出した不確かさを表すキーワード毎の特徴単語を分類して各不確かさを表すキーワードの特徴を明らかにする。

以下に各ステップの詳細を述べる。

#### 4.2.1 ステップ 1: キーワード毎の特徴単語を抽出する

コミットメッセージには、主にそのコミットで更新された内容や内容に対して開発者が気づいたことについて書かれたものである。コミットメッセージは自然言語で構成されたものであるため、テキスト分析によりその特徴単語を抽出することが可能である。不確かさを表すキーワードを含むコミットメッセージからキーワード毎の特徴単語を抽出するために、まず、プロジェクトごとに不確かさを表すキーワードを含むコミットメッセージから、各キーワードの特徴単語を抽出する。その後、全プロジェクトの特徴単語をキーワード毎に統合して特徴単語を抽出する。

プロジェクト毎の特徴単語を抽出するために 4 つの段階に分けて操作を行った。各操作は以下の通りである。

(1) 全コミットメッセージの特徴値付き単語リストを抽出する

オープンソースの統計解析システムである R\*2 のテキスト分析における代表的な tm パッケージ [7] を利用して、プロジェクトすべてのコミットメッセージを 1 つ 1 つの文章として集積して 1 つのコーパスを作成する。このコーパスに、各コミットメッセージの句読点や空白等ゴミ情報を除いた後、単語だけを残す。その後、同じ意味の単語を形態の違いにより二重カウントしてしまうことを防ぐため、ステミング処理で単語の語幹だけを残し、各コミットメッセージにある単語の tf-idf の値を計算する。tf-idf は、tf (Term Frequency, 単語の出現頻度) と idf (Inverse Document Frequency, 逆文書頻度) の 2 つの指標にもとづいて計算される値である。tf-idf の値が高いほど単語の特徴が大きく、低いほど単語が一般的となる。最後に単語毎の tf-idf の平均を計算して特徴値として抽出する。

(2) 不確かさを表すキーワードを含むコミットメッセージの特徴値付き単語リストを抽出する

不確かさを表すキーワードを含むコミットメッセージから各キーワードの単語リストを (1) と同様に抽出する。

(3) 特徴単語の差分リストを作る

(2) で抽出した各不確かさを表すキーワード毎の特徴値付き単語リストをステップ 1 で抽出した全コミットメッセージの単語リストと比較して、不確かさを表すキーワードの単語の特徴値から、全コミットメッセージの単語の特徴値を引いて差分の単語リストを作る。

(4) 差分が 0 以上の特徴単語だけ抽出する

(3) で作られた各不確かさを表すキーワードの単語差

\*1 <https://github.com/github>

\*2 <https://www.r-project.org/index.html>

分リストから差分が0以上の条件でさらに抽出する。不確かさを表すキーワードを含むコミットメッセージの特徴単語とは、全コミットメッセージにおいて一般的な単語（特徴値が低い）であるが、不確かさを表すキーワードを含むコミットメッセージにおいては特徴的な単語（特徴値が高い）である。したがって、単語の特徴値の差分が大きいほど不確かさを表すキーワードの特徴単語となる。

各プロジェクトに以上4つの操作を行った結果として、各プロジェクトでの不確かさを表すキーワードの特徴単語が抽出できる。そこから全プロジェクトにおいて、各不確かさを表すキーワードの特徴単語リストが作成できる。しかし、単語の差分が小さすぎると、その単語は特徴的な単語とは言い難いので、不確かさを表すキーワードごとの特徴単語の差分値を降順でトップ20の特徴単語に限定して、全プロジェクトにおいて出現したプロジェクトの数と差分の平均及び変動係数（Coefficient of Variation）[1]を計算する。変動係数は差分の標準偏差を平均値で割った値である。

単語の差分の平均が大きいほど特徴的で、変動係数が高いほど分散の値が大きい。また、最後の分析はマニュアルで行うため、不確かさを表すキーワードの頻出度によりトップ20の不確かさを表すキーワードを抽出する。

#### 4.2.2 ステップ2：特徴単語に品詞タグを付いて名詞と動詞だけ抽出する

表2で示したように、不確かさを表すキーワードは全て形容詞と副詞である。よって、それらが用いられているコミットで何を行ったのかを調査するため、名詞と動詞だけを抽出する。抽出方法としては、RのopenNLPパッケージ[12]を用いる。openNLPパッケージを用いることで、米Pennsylvania大学のRatnaparkhiが開発したMaxentモデル[15]を利用して文章の単語にタグをつけることができる。Maxentモデルは、同じ大学のMarcus氏ら5年間で作った英語品詞タグ（Part of Speech Tags）データベース（The Penn Treebank [11]）に基づいて開発されている。これにより、ステップ1で抽出した不確かさを表すキーワードの特徴単語リストにある名詞単語と動詞単語にタグを付けて抽出する。実際にunknownキーワードに対してタグ抽出までを行った結果を表3に示す。

#### 4.2.3 ステップ3：特徴単語の分析

本ステップでは、ステップ2で抽出した各不確かさを表すキーワードの特徴単語を分類する。その後、各特徴単語の差分の平均と変動係数に合わせて、各不確かさを表すキーワードを分類する。分類結果から不確かさを表すキーワードを含むコミットメッセージがどのような特徴が見られるかを目視で観察する。

### 4.3 分析結果

本節ではデータセットである1,444件のOSSに対して、

表3 品詞タグ付きの特徴単語リスト（unknown）の例

Word	Mean (平均)	CoV (変動係数)	プロジェクト数	POS タグ
type	0.01874571	0.853236928	808	NN (名詞)
error	0.018047575	0.913466452	903	NN (名詞)
messag	0.016683953	0.901754048	590	NN (名詞)
ignor	0.016163827	1.080568182	528	NN (名詞)
return	0.015338545	0.8766055	677	NN (名詞)
handl	0.014811577	1.105870303	605	NN (名詞)
reason	0.012827254	0.873758822	666	NN (名詞)
name	0.010824656	0.98421489	531	NN (名詞)
report	0.009979287	1.080975554	539	NN (名詞)
case	0.009530822	1.082011154	500	NN (名詞)
caus	0.007997205	1.138618122	538	NN (名詞)
will	0.00708426	0.902794009	592	NN (名詞)
warn	0.016426316	0.911621722	479	VBP (動詞)
fail	0.01033176	1.012432553	543	VB (動詞)
tri	0.008507029	1.229532611	507	VBD (動詞)

表4 特徴単語の類別

類別	説明
Programming	プログラミング用語
Programming Operation	プログラミングの操作に関する用語
Self-Reference	不確かさを表すキーワード自身に関する単語 多くの場合は無視して良い
Uncertain	不確かさを表す単語
General	特徴がない、一般的な用語
Bug	error, warn, bug のいずれか

第4.2章で述べたアプローチを適用し抽出した不確かさを表すキーワードの特徴単語、及び、その不確かさを表すキーワードに対する分類と分析結果を示す。

#### 4.3.1 キーワードの分類

キーワードの分類について、まず特徴単語を分類して、それから不確かさを表すキーワードを分類する。

##### 特徴単語の分類

表2に示した20個の不確かさを表すキーワードの特徴単語の差分の平均が高いほど特徴的であり、変動係数が高いほど分散度が高い。また、差分の平均が高く、かつ変動係数が高い単語はあらゆるプロジェクトに存在し、そのような単語は不確かさを表すキーワードと結びつきが強いことが分かる。今回、我々は各不確かさを表すキーワードの特徴単語を6つの類別に分類した。表4は今回用いた類別である。

不確かさを表すキーワードの特徴単語と各特徴単語の分類は表5で示す。

##### 不確かさを表すキーワードの分類

我々は前節で分類された特徴単語を元に不確かさを表すキーワードを分類を行なった（表6）。それぞれの類別は表7で示す。

#### 4.3.2 不確かさを表すキーワードの特徴分析

表6で示した結果と特徴単語の分類を合わせてキーワード毎に分析した。今回分析を行った結果、表7のそれぞれの類別について特に特徴的な傾向が見られたものについて挙げる。

**changeable** 図1はchangeableを含むコミットメッセージの一例である。fieldやstate等、変数に関する単語が出現する傾向が見られた。パラメータの使用等の不

表 5 不確かさを表すキーワードの特徴単語とその分類

Uncertain Word	Programming	Programming Operation	Self-Reference
ambiguous	call, valu, type, case	renam	ambigu
arcane	trigger, damag, stack	build, make	
changeable	field, set, cach, default	make, offset	changeabl, chang
debatable	case, system	get, put	debat
dubious	valu, code, bit, case	replac, get	
doubtful			doubt
erratic	enabl, issu, behavior, result, logic		errata, erratum, errat
fuzzy	test, input, requir, code, number, unit	got	fuzzi, fuzz
irregular	break, switch, pass, valu		
may	case		
might	case, tri	get	
obscure	code, case, problem, result, need	get, call	
probably		get	
risky	secur, case, data	run	risk, riski
tentative	method, implement, problem	call	tentat
unclear	valu		
unknown	type, messag, return, fail, case	tri	
unreliable	test, case	detect, run, tri, set	
unsure	case		
vague	return, line, function	set, get	
Uncertain Word	Uncertain	General	Bug
ambiguous		avoid, name, differ, one	warn, error
arcane		take, shot, entri, seem, empti, column, will	error, problem
changeable		disk, select, pass, check, addit, start, found	error
debatable	seem	name, place, want, think, realli	
dubious	seem	use, remov, avoid	warn, error
doubtful		see, realli, one, work, want, look	
erratic	workaround	handl, appli, arm, caus, affect, part	bug
fuzzy	option	translat, way	
irregular		part, usag, approach, place, help, condit, work	
may	mayb	caus, need, want, sinc, contain, time, differ, happen, mean	
might		want, need, caus, differ, look, happen	
obscure		caus	bug, error
probably	seem	need, want, caus, work, doesnt, way	
risky	potenti	reduc, dont, avoid, need	
tentative		support, definit, work, like, allow, way, done	bug
unclear	seem	name, reason, user, differ, like	
unknown		ignor, handl, reason, name, report, caus, will	error, warn
unreliable		remov, check, chang, time, dont, result, need	
unsure	seem	work, dont, need, think, result	
vague	attempt	document, time, think, caus	error



図 1 changeable が含まれたコミットメッセージの例

確かさに関わる可能性が高い。

**debetable** リファクタリングでよく出現すると考えられる name, place, put, get 等の特徴単語が出現するため、それに関連する不確かさが存在する可能性は高い。図 2 は debetable を含むコミットメッセージの一例である。開発者はラベルの処理ルールに疑問が持っていたため、ルールを変更した。

**dubious** 差分平均が大きい warn や error 等の特徴単語が出現する。それらとともに code, value, remove, re-

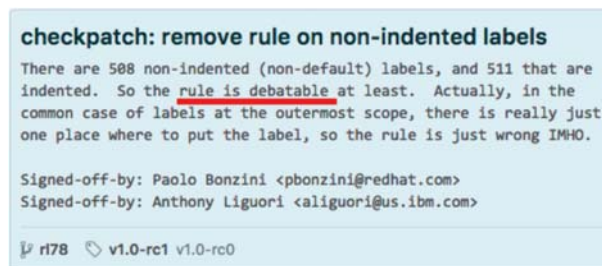


図 2 debetable が含まれたコミットメッセージの例

place 等の特徴単語が出現する。疑わしいコードや値が発見した際、問題となったところに remove や replace 等の操作を行うことが想定できる。図 3 は dubious を含むコミットメッセージの一例である。開発者はバグ発生可能なコードを削除したが、この操作によりまた

表 6 不確かさを表すキーワードの分類

キーワード	分類 (開発工程)	分類 (コード要素)	バグ関連語の有無
ambiguous	実装	型, 変数, call	o
arcane	実装	リスト	o
changeable	実装	変数	o
debatable	実装	リファクタリング	x
dubious	実装	変数	o
doubtful	-	-	x
erratic	実装	振る舞い	o
fuzzy	テスト	-	x
irregular	-	変数	x
may	要求	-	x
might	要求	-	o
obscure	実装	call	o
probably	要求	-	x
risky	実装	セキュリティ, データ	x
tentative	実装	メソッド, call	o
unclear	-	変数	x
unknown	実装	型	o
unreliable	要求, テスト	-	x
unsure	-	-	x
vague	実装	ドキュメント, 関数	o

表 7 不確かさを表すキーワードの類別

類別	説明
分類 (開発工程)	一般的な開発工程 (要求, 設計, 実装, テスト) のどれに関わりがあるか
分類 (コード要素)	上項目が「実装」のものがどのコードの要素 (型, メソッド, など) に関わりがあるか
バグ関連語の有無	bug, error, warn のどれかの有無

```

Clean up error cases in psql's COPY TO STDOUT/FROM STDIN code.
Adjust handleCopyOut() to stop trying to write data once it's failed
one time. For typical cases such as out-of-disk-space or broken-pipe,
additional attempts aren't going to do anything but waste time, and
in any case clean truncation of the output seems like a better behavior
than randomly dropping blocks in the middle.

Also remove dubious (and misleadingly documented) attempt to force our way
out of COPY_OUT state if libpq didn't do that. If we did have a situation
like that, it'd be a bug in libpq and would be better fixed there, IMO.
We can hope that commit fa4448f took care
of any such problems, anyway.
    
```

図 3 dubious が含まれたコミットメッセージの例

```

feat(http): set custom default cache in $http.defaults.cache
When we need more control over http caching, we may want to provide
a custom cache to be used in all http requests by default.

To skip default cache, set {cache: false} in request configuration.
To use other cache, set {cache: cache} as before.

See #2879
    
```

図 4 may が含まれたコミットメッセージの例

バグが発生する可能性がある。正しい対応法は分かっていない。

**may** 原因や要求に関する用語が多いため、要求に関わる不確かさが見られた。図 4 は may を含むコミットメッセージの一例である。開発者は将来必要となる機能のためにパラメータの設定を行った。

**obscure** bug, error, message 等バグに関連する単語が多く出現する。また、それらの特徴単語の変動係数も非常に高いため、バグに関わる不確かさが多く含まれ

```

Comment out "METHOD BY UUID CALLED" WriteLine as it is
obscuring the UserServer console for now.
    
```

図 5 obscure が含まれたコミットメッセージの例

```

diag: Remove world read/write permissions from /dev/diag
/dev/diag has world read/write permissions which is considered
a security risk. Remove these permissions. Add new qcom_diag
group for access to /dev/diag

CRs-Fixed: 356415
(cherry picked from commit 14c85ec56fd47c5eb143617a3bd5a6e7ee39cb6e)

Change-Id: I536323a97e19a07bbd8eadffa0265d1955590285
    
```

図 6 risky が含まれたコミットメッセージの例

```

Fast inverse square root method
Use well-known Quake3 hack.

FastMath: fast integer vector normalisation

Fast normalisation

Fast inv sqrt

fast sqrt

fast sqrt

fixed: implement fast inverse sqrt function

fixed: return zero on fast_sqrt(x=0)

test_normalise: only do fast sqrt tests if fixed

[mk: removed timing tests, as these failed on my machine, and
comparing the wallclock time is inherently unreliable]
    
```

図 7 unreliable が含まれたコミットメッセージの例

ている可能性が高いと推測できる。図 5 は obscure を含むコミットメッセージの一例である。開発者はバグが発生可能なコードを一時的にコメントアウトした。

**risky** セキュリティに関わるのが一般的に認識されている。したがって、risk や secure 等セキュリティに関する特徴単語が出現することが想定できる。図 6 は risky を含むコミットメッセージの一例である。開発者はセキュリティ問題を防ぐためにファイルの権限を変更した。

**unreliable** 信頼性に関わることが多い。time や detect 等テストに関する単語が特徴的である。図 7 は unreliable を含むコミットメッセージの一例である。開発者は信頼できないテストを除いた。

#### 4.3.3 傾向分析

以上のように不確かさを表すキーワードごとに特徴的なキーワードをいくつか分類できた。キーワード may は要求に関する不確かさが多く見られた。changeable, debatable

はプログラムの実装でパラメータや API 等に関する不確かさが存在していた。その他にも erratic, vague などでも同じ傾向が見られた。dubious, obscure はバグ関連の内容が多いため、バグに関する不確かさが存在する可能性が高いと考えられる。同じく, ambiguous, arcane, might, tentative, unknown でもバグを取り除くためのコミットが見られた。unreliable はプログラムのテストに関連する可能性があり, fuzzy もテスト関連の用語が見られた。risky はセキュリティに関連するコミットメッセージが確認された。

## 5. 妥当性への脅威

分析にて不確かさを表すキーワードを分類した際に、幾つか特徴的な傾向が見られないキーワードが存在したため、傾向を捉える手法の改善が必要である。特徴単語の抽出は既存研究に基づいて行ったが、特徴単語の分類と不確かさを表すキーワードの分類では、そういったキーワードに特徴的な傾向が存在する可能性があるとしても、分類者のプログラミング経験や知識などにより発見できない可能性もあると考えられる。

また、この分析はコミットメッセージのみを対象として行ったもので、ソフトウェアにおいて実際のコードとその変更履歴を見ない限り、分析の結果が正しいとは言えない。そのため、実際のコードや変更履歴と合わせてそれらの特徴的な傾向を検証することが必要となる。また、コミットコメントには開発者が認識している問題などが書かれるものなので、認識されているレベルの不確かさ、つまり Known Unknowns は多く現れていると予想されるが、開発者が認識していない不確かさ、つまり Unknown Unknowns を抽出する方法としては十分でない可能性がある。こういった理由から今回の手法で発見した不確かさは、傾向に偏りが存在しているかもしれないと考えられる。

最後に、今回使用した不確かさを持つ

## 6. まとめ

本論文では、オープンソースソフトウェアの開発履歴を分析対象として、実際のソフトウェアプロジェクトにおいてどのような不確かさが存在するかを調査した。

不確かさを表すキーワードを含むコミットメッセージにはどのような特徴があるか、ということに関しては、複数のプロジェクトの分析により特徴を持つキーワードをいくつか見つけることができた。キーワード may は要求に関する不確かさが存在する可能性がある。changeable, debatable, erratic, vague はプログラムの実装でパラメータや API 等に関する不確かさが存在する可能性がある。ambiguous, arcane, dubious, might, obscure, tentative, unknown はバグ関連キーワードである可能性が高い。fuzzy, unreliable はプログラムのテストに関連する可能性がある。risky は

セキュリティに関連する可能性が高い。また、テキスト分析の際に制作した不確かさを表すキーワードごとの特徴単語の差分リストは、不確かさが存在するコミットの特徴を捕えたりリストであるので、機械的に不確かさを調査する際に不確かさを見つけるための指標として利用できると考えている。

## 謝辞

本研究は、文部科学省科学研究補助費基盤研究 (A) (課題番号 26240007) による助成を受けた。

## 参考文献

- [1] Brown, C. E.: *Coefficient of Variation*, pp. 155–157, Springer Berlin Heidelberg (1998).
- [2] Elbaum, S. and Rosenblum, D. S.: Known Unknowns: Testing in the Presence of Uncertainty, *Proceedings of the 22nd International Symposium on Foundations of Software Engineering*, pp. 833–836 (2014).
- [3] Esfahani, N., Malek, S. and Razavi, K.: GuideArch: guiding the exploration of architectural solution space under uncertainty, *Software Engineering (ICSE), 2013 35th International Conference on*, pp. 43–52 (2013).
- [4] Esfahani, N., Razavi, K. and Malek, S.: Dealing with uncertainty in early software architecture, *Proceedings of the ACM SIGSOFT 20th International Symposium on the Foundations of Software Engineering*, p. 21 (2012).
- [5] Famelis, M., Ben-David, N., Sandro, A. D., Salay, R. and Chechik, M.: MU-MMINT: an IDE for Model Uncertainty, *Proceedings of the 37th International Conference on Software Engineering*, pp. 697–700 (2015).
- [6] Famelis, M., Salay, R. and Chechik, M.: Partial Models: Towards Modeling and Reasoning with Uncertainty, *Proceedings of the 34th International Conference on Software Engineering*, pp. 573–583 (2012).
- [7] Feinerer, I.: Introduction to the tm Package Text Mining in R (2015).
- [8] Garlan, D.: Software engineering in an uncertain world, *Proceedings of the FSE/SDP workshop on Future of software engineering research*, pp. 125–128 (2010).
- [9] Letier, E., Stefan, D. and Barr, E. T.: Uncertainty, risk, and information value in software requirements and architecture, *Proceedings of the 36th International Conference on Software Engineering*, pp. 883–894 (2014).
- [10] Lindberg, C. A.(ed.): *Oxford american writer's thesaurus*, Oxford University Press (2012).
- [11] Marcus, M. P., Marcinkiewicz, M. A. and Santorini, B.: Building a large annotated corpus of English: The Penn Treebank, *Computational linguistics*, Vol. 19, No. 2, pp. 313–330 (1993).
- [12] Meyer, D., Hornik, K. and Feinerer, I.: Text mining infrastructure in R, *Journal of statistical software*, Vol. 25, No. 5, pp. 1–54 (2008).
- [13] Perez-Palacin, D. and Mirandola, R.: Uncertainties in the Modeling of Self-adaptive Systems: A Taxonomy and an Example of Availability Evaluation, *Proceedings of the 5th ACM/SPEC International Conference on Performance Engineering*, pp. 3–14 (2014).
- [14] Ramos, J.: Using tf-idf to determine word relevance in document queries, *Proceedings of the first instructional conference on machine learning* (2003).

- [15] Ratnaparkhi, A.: A maximum entropy model for part-of-speech tagging, *Proceedings of the conference on empirical methods in natural language processing*, Philadelphia, USA, pp. 133–142 (1996).
- [16] Walker, W. E., Harremoës, P., Rotmans, J., van der Sluijs, J. P., van Asselt, M. B., Janssen, P. and Kreyer von Krauss, M. P.: Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support, *Integrated assessment*, Vol. 4, No. 1, pp. 5–17 (2003).