スマートフォンアプリケーションのレビューにおける苦情の分析
―ソフトウェア品質モデルに基づく分類についての考察―

2018SE047 水谷大志　2018SE089 竹下光希　2018SE090 田中雅貴
指導教員：横森励士

1　はじめに

スマートフォンアプリケーションにおけるユーザーレビューは開発者にとって保守における開発の方向性を決めるための重要な参考意見となる．過去の研究では，ユーザーレビューの苦情内容の分類を行っているが，分類の基準が苦情の内容で分類しており，ソフトウェア品質の観点から分析を行うために，分類モデルを再構築する必要がある．本研究では，ユーザーレビューの苦情内容を「製品品質モデル」と「利用時の品質モデル」の各項目に関連付けて分類を行う方法を提案し，いくつかのアプリケーションのレビューに対して実際に分類を行う．品質モデル上でどのように分類されるか，過去の研究で得られた特徴がどのように現れるかなどを調査し，ユーザーが品質モデル上でどの要素を重視しているかを得ることを目的とする．

2　研究背景

2.1　ユーザーレビューを分析した研究について

スマートフォンアプリケーションは一般的にアプリストアを介して配布され，ユーザーは利用したアプリケーションについての評価をアプリストアに投稿できる．投稿したレビューはタイトル，星評価，具体的なコメントで構成され，それらの情報は開発者へのフィードバックや他のユーザーへのアドバイスとなっている．ユーザーレビューは，アプリの開発者にとって，開発や保守の方向性を決めるための重要な参考意見となる．Leonard Hoon ら [1] は App Store の 17,330 のアプリから 870 万件のレビューにおいて使われている単語をすべて抽出し，評価との関連を調査した．その結果，肯定的な意味を表す単語より否定的な意味を表す単語の方が種類数が多いことが分かった．Khalid ら [2] は北米で提供されている無料 iOS アプリを対象に低評価レビューを分析し，ユーザーがどのような要素に対して不満を持ちやすいかを分析した．低評価レビューのコメントの内容に基づき，表 1 で示す 12 種類の苦情を表現するタグを付けた．結果として，各苦情タイプの中で「機能エラー」,「機能要求」が苦情に多く存在することを示した．

我々の研究グループでは，

- 苦情は低評価だけに存在するのではない
- 地域ごとに苦情の傾向に差があるのではないか

と考えて，日英米の市場向けに展開されたアプリケーションに対して，できる限り条件をそろえて，Khalid らの分類方法に基づいて苦情を分類した [3][4][5][6]．その結果，以下のようなことが分かった．

- どの地域における分類結果でも，低評価で「機能エラー」が，中・高評価では，「機能要求」が多く見られ，これらは一般的な傾向として考えられる．
- 3 地域で共通の傾向が見られたが相違点も存在し，米は積極的に意見を言う傾向が強いといった地域性や地域の政治や文化による違いも見られた．

これらの結果からは，レビューを目的に応じて評価帯ごとに見ることが有益であることや，アプリケーションを他地域に向けて展開する際に，同じように考えるべき部分と地域性などを考慮すべき部分があることを示しており，レビューから方針を測る際の参考になると考えられる．

3　品質モデルに基づいた分類

3.1　現状の課題

過去の研究では，Khalid ら [2] の分類方法に基づいて分類を行ってきた．実際に得られた分類結果を品質の観点から評価しようとすると，品質モデル上での評価が必要となる．しかし，1 つの分類項目が品質モデル上の複数の観点にまたがることが多く，分類結果から品質のどのような性質に問題があるかを説明することが難しい．実際にはレビューを品質モデルの評価項目に沿って分類し直す必要がある．今回 Super Mario Run の北米レビュー (286 件) を分類した際，Khalid らの分類項目で分類した際はチーム内で 36 回の意見の相違があったのに対し，品質モデル上で分類した際には 134 回もの意見の相違があり，品質モデル上での分類が難しいことが分かった．

3.2　提案する手法でのアプローチ

本研究では，スマートフォンアプリケーションのレビューを品質モデル上の区分に合わせて分類することで，品質モデル上でレビューを分類評価する方法を提案する．一度 Khalid らの分類項目に基づいて分類を行ったあとで，それぞれの分類項目が品質モデル上の区分である「利用時の品質モデル」のどの項目に当てはまるかを調査する．その上で,「製品品質モデル」のどの項目に関連があるかを調査するというアプローチで分類する．このアプローチを用いることで，関連がある項目だけを見るので正確に分類できることが考えられ，品質モデル上での問題点の整理が行いやすくなると考えられる．実際に，Khalid らの分類項目に基づいて分類を行ったあとで，Khalid らの分類項目がどの品質モデル上の項目に対応するかなどに着目しながら分類を行い，手法の有効性を確認する．

表 1 Khalid らの分析における苦情の種類

苦情タイプ	苦情の詳細	レビュー例
強制終了	アプリケーションが強制終了する	起動後，すぐに落ちる
互換性	特定のデバイスや OS のバージョンに問題がある	ipod touch では画面の半分しか見れない
機能削除	特定の機能がアプリケーションを台無しにしている	アプリ自体は素晴らしいが広告を除いてほしい
機能要求	より良くなるために，機能を追加する必要があると感じている	アラートを設定できる機能がない
機能エラー	アプリケーションの特定の問題に言及し，不満を感じている	アプリケーションを開かないと通知が来ない
隠されたコスト	全てを経験するために追加の隠されたコストが必要	リアルマネーを使い，コインの購入を強いてくる
インターフェース設計	デザイン，制御，映像について不満がある	デザインが小奇麗でなく，わかりづらい
ネットワーク問題	ネットワークに問題があるか，応答速度が遅い	新しいバージョンがサーバーにつながらない
プライバシーと倫理	プライバシーを侵す，または反倫理的である	あなたとの接触が目的なアプリケーション
アプリが応答しない	入力の応答が遅い，または全体的に遅い	古いバージョンに戻したい！ スクロールが遅い
魅力のない内容	特定のコンテンツが魅力的ではない	画面の見栄えは良いが，退屈でつまらないゲーム
重いリソース	アプリケーションがバッテリーまたは容量を消費しすぎる	常時 GPS を使い，バッテリーが消費される
特定できない	ただ単にアプリケーションが悪いと言っている	正直なところ，最悪のアプリケーション

3.3 ソフトウェア品質モデルの分類

JIS X 25010 ではソフトウェアの品質モデルを，表 2 の「製品品質モデル」と表 3 の「利用時の品質モデル」の 2 つの品質モデルとして定義している．「製品品質モデル」は製品そのものが持つ品質を 8 種類に分類している．「利用時の品質モデル」は，利用者がある特定の状況で製品を利用した観点からの品質を 5 種類に分類している．これらの主特性の多くは副特性として更に細かく分類されている．

表 2 「製品品質モデル」の項目

主特性	副特性	主特性	副特性
機能適合性	①機能完全性 ②機能正確性 ③機能適切性	性能効率性	④時間効率性 ⑤資源効率性 ⑥容量満足性
保守性	㉔モジュール性 ㉕再利用性 ㉖解析性 ㉗修正性 ㉘試験性	使用性	⑨適切度認識性 ⑩習得性 ⑪運用操作性 ⑫ユーザエラー防止性 ⑬ユーザインタフェース快美性 ⑭アクセシビリティ
信頼性	⑮成熟性 ⑯可用性 ⑰障害許容性 ⑱回復性	セキュリティ	⑲機密性 ⑳インテグリティ ㉑否認防止性 ㉒責任追跡性 ㉓真正性
互換性	⑦共存性 ⑧相互運用性	移植性	㉙適応性 ㉚設置性 ㉛置換性

表 3 「利用時の品質モデル」の項目

主特性	副特性	主特性	副特性
㉜有効性		㉝効率性	
満足性	㉞実用性 ㉟信用性 ㊱快感性 ㊲快適性	リスク回避性	㊳経済リスク緩和性 ㊴健康・安全リスク緩和性 ㊵環境リスク緩和性
利用状況網羅性	㊶利用状況完全性 ㊷柔軟性		

3.4 Khalid らの分類と品質モデルの対応について

Khalid らは内容を中心に 12 種類に分類している一方で，「利用時の品質モデル」は特性を中心に分かれており，似た項目が多い．Khalid らの分類項目が「利用時の品質モデル」のどの区分に該当するかを事前に表 4 のように検討し，分類時にはその表を参考に分類する．同様に，「利用時の品質モデル」の項目が「製品品質モデル」のどの項目に対応し得るかも事前に調査し，その表を基に分類を行う．

表 4 表 1 の苦情の種類と「利用時の品質モデル」の関係

苦情タイプ	㉜	㉝	㉞	㉟	㊱	㊲	㊳	㊴	㊵	㊶	㊷
強制終了	○	×	○	○	×	×	○	○	×	○	×
互換性	○	×	○	○	×	×	○	×	×	○	○
機能削除	×	○	×	×	○	○	×	○	×	×	×
機能要求	○	○	○	×	○	○	○	○	×	×	○
機能エラー	○	○	○	○	×	○	○	○	×	○	×
隠されたコスト	×	○	×	○	○	○	○	×	○	×	×
インターフェース設計	○	○	○	○	○	○	○	○	×	○	×
ネットワーク問題	○	○	○	○	×	○	×	×	×	×	○
プライバシーと倫理	×	×	×	×	○	○	○	○	×	×	×
アプリが応答しない	○	○	○	○	×	×	○	○	×	○	×
魅力のない内容	○	○	○	○	○	○	×	×	○	×	○
重いリソース	×	○	×	×	○	○	○	×	○	○	×

4 評価実験

4.1 実験の内容と評価項目

実際にレビューを入手し，提案手法により品質モデル上での分類を行う．以下の評価項目を定め，品質モデル上での分類の傾向や Khalid らの分類項目での分類との違いなどを調査することで，提案手法の有効性を確認する．

- レビューは品質モデル上でどのように分類されるか？
  品質モデル上での分類結果からどのように分類結果が表現されるかを確認する．品質モデル上で評価が難しく，新たに追加すべき項目について考察する．
- Khalid らの分類項目は品質モデル上でどのように分類されるか？
  「機能要求」「機能エラー」などが多いという特徴が品

質モデル上でどのように表されるかを調査する．

- 品質モデル上のどの項目を考慮すればよいか？
  項目間の関連や分類の特徴から，どのような項目に注目した分析が必要になるかを考察する．

4.2 分類結果

本研究では，PINTEREST と Super Mario Run で日本と北米のレビューを分類した．要旨では日本における PINTEREST での分類結果を紹介する．実験におけるレビューの分類手順を以下に示す．

1. App store において公表されているアプリケーションのユーザーレビューを「iTunes Store Web Service Search API」を通じて一定区間取得する．
2. 各ユーザーレビューの id，レビュータイトル，星評価，コメントを抽出する．
3. 抽出したレビューから Khalid らの区分で分類する．
4. 表 4 のような対応表を参考に，「利用時の品質モデル」，「製品品質モデル」の順で分類する．

最初に，Khalid らの分類区分での分類結果を表 5 に示す．次に，表 4 の項目間の関連に基づいて，「利用時の品質モデル」上での分類を行った．表 6 は分類結果である．表内の数字は苦情の出現件数であり，[] 内の数字は称賛レビューの出現件数である．ユーザーがアプリに過剰な期待をしたことで「過大評価」を行っていると考えられる苦情が見られ，品質モデル上での分類が困難であった．そこで，新たな分類項目として「㊸機能面」「㊹性能面」「㊺コスト面」「㊻品質面」で「過大評価」といった項目を追加した．「有効性」「実用性」「信用性」「利用状況完全性」について言及されることが多い．「利用時の品質モデル」上での分類結果から，表 7 のように「製品品質モデル」上での分類を行った．「機能正確性」「成熟性」「可用性」に関するレビューが多く，「強制終了」「機能エラー」が該当した．

表 5 Khalid らにおける PINTEREST の分類結果 (日)

JP	星1	星2	星3	星4	星5	合計
強制終了	40	17	19	21	19	116
機能削除	6	1				7
機能要求	5	4	7	5	4	25
機能エラー	21	9	6	4	3	43
インターフェース設計	3	4	5	1		13
ネットワーク問題	1		1		1	3
プライバシーと倫理	3	1	1		1	6
アプリが応答しない	3	2	2	1	1	9
魅力のない内容	4	3	2	2		11
重いリソース			1			1
特定できない	3	3	3	3	20	32
合計	89	44	47	37	49	266

表 6「利用時の品質モデル」の分類結果 (PINTEREST, 日)

JP		星1	星2	星3	星4	星5	合計
㉜有効性		62	30	24	27	21[4]	164[4]
㉝効率性		7	3	8	3	2	23
満足性	㉞	49	22	22	24[1]	20[5]	137[6]
	㉟	62	27	25	27	22	163
	㊱	1				[4]	1[4]
	㊲	10	9	9	6	2[2]	36[2]
リスク回避性	㊳			2			2
	㊴	1	1			1	3
利用状況網羅性	㊶	44	17	20	23	19	123
	㊷	1				1	2
過大評価	㊸	3	2	4	1	1	11
	㊹			2			2
	㊻			1			1
㊼特定できない		5	4	1	1[1]	2[14]	13[15]
苦情レビュー数		245	115	118	112[2]	91[29]	681[31]

表 7 「製品品質モデル」の分類結果 (PINTEREST, 日)

JP		星1	星2	星3	星4	星5	合計
機能適合性	①	4	2	2	2	3[1]	13[1]
	②	62	28	25	27	22	164
	③	3	3[1]	6	4		16[1]
性能効率性	④	4	2	2	1	1	10
	⑤			1			1
互換性	⑦	1			1		2
使用性	⑨		1			[3]	1[3]
	⑪	2	4	6	2		14
	⑬	2	5	5	1	1	14
	⑭	2	1	1			4
信頼性	⑮	62	28	26	26	22	164
	⑯	55	21	21	23	20	140
	⑱					1	1
保守性	㉔	12	7	6	3	4	32
	㉖	12	7	6	3	4	32
	㉗	12	7	6	3	4	32
苦情レビュー数		233	116[1]	113	96	82[4]	640[5]

4.3 結果の要約

北米における PINTEREST のレビューの分類では，Khalid らの「機能エラー」に分類されていたレビューが品質モデル上では「有効性」「信用性」「機能正確性」「成熟性」に分類された．「機能エラー」でも「PINTEREST は好きだがアプリが正常に動作しない」というレビューは「実用性」にも分類され，「ピンの入れ替えが出来なくなった」というレビューであれば「効率性」にも分類されることがあった．北米の高評価帯で多く見られた「機能要求」では「効率性」「快適性」「機能面での過大評価」「機能完全性」「機能適切性」に分類された．「機能要求」でも「アップデートで使いづらくなった機能を戻してほしい」という

レビューは「効率性」「快適性」「機能完全性」「機能適切性」に分類され，明示されていない機能を要求しているレビューは「機能面での過大評価」に分類された．

Super Mario Run のレビューの分類では，「隠されたコスト」に分類されたレビューが「効率性」「機能適切性」に分類された．「隠されたコスト」でも課金自体を否定しているレビューは「コスト面での過大評価」「適切度認識性」にも分類された．次いで多く見られた「魅力のない内容」では「実用性」に多く分類され，低評価帯でアプリを楽しめていない様子が見られた．「機能エラー」では「有効性」「実用性」「信用性」「機能正確性」「成熟性」に分類された．北米の高評価帯で多く見られた「機能要求」では「新ステージ，アイテム，キャラが欲しい」というレビューが多く，「機能面での過大評価」に分類された．

5　考察

ほとんどのレビューが「利用時の品質モデル」上で分類されたが，「製品品質モデル」には分類されない項目も多い．特に「製品品質モデル」の「互換性」「セキュリティ」「保守性」「移植性」にはほとんど分類されない．品質モデルは Khalid らの分類項目と異なり度合いを表現するので，称賛レビューも集計した．結果，称賛レビューは「特定できない」に分類されることが多かったが，「有効性」「実用性」「快感性」「快適性」「適切度認識性」などに少なからず分類されたため，分類可能なものも存在する．レビューを分類する際に追加した「過大評価」の項目に分類されるレビューも一定数存在したが，「品質」や「コスト」に関する過大評価のレビューは全て他の項目で分類可能であったため，最終的には必要ないと考えられる．

Khalid らの「機能エラー」に分類されたレビューは，主に「有効性」「実用性」「信用性」「機能正確性」「成熟性」に分類されることが多く，Khalid らの「機能要求」に分類されたレビューは，主に「効率性」「快適性」「機能面での過大評価」「機能完全性」「機能適切性」に分類された．これらの品質モデルでの分類結果は，アプリや国ごとで同じ組み合わせで出現することが多く，品質モデルの項目間で関連性があると考えられる．

どのアプリケーションでも「有効性」「信用性」「機能正確性」「成熟性」が一定数入っていることから，レビューからはシステムの正確性が読み取れることが分かった．ユーザーはアプリケーションの機能面や性能面に対して多くの提言を持つ反面，「セキュリティ」に対してのレビューがほとんどないことから，セキュリティ面に関する問題はユーザーからの事例がない限り，レビューを読み取ることが難しい．また，「移植性」に当てはまらないことから，他のハードウェアや Web サービスをアプリケーションに移植した場合に多く出現するものだと考え，別のアプリケーションについて調査するべきだと考えられる．

提案手法の妥当性について考察する．日本の PINTEREST で Khalid らの分類方法に基づいて分類した場合，苦情レビュー 1 件あたりの該当する項目は 1 項目であるものが 204 件と多数で，2 項目であるものが 23 件，3 項目であるものが 5 件と，複数の項目にわたるレビューはほとんど存在しなかった．一方，品質モデル上で分類した場合，苦情レビュー 1 件あたり 6 項目に分類されるレビューがほとんどであり，該当した項目数の平均は 5.37 項目と多数であったため，分類の漏れを抑えるための今回のようなガイド手法が必要であると考えられる．また，対応表による絞り込みのフェーズを自動化し，最終的に分類される項目は人の目で判断するといったように，品質モデル上でのレビューの分類を半自動化できるとも考える．

6　まとめと今後の課題

本研究では，日本と北米のスマートフォンアプリケーションのレビューを品質モデル上の区分に合わせて分類した．過去の「機能要求」に分類されたレビューが，品質モデル上で主に「効率性」「快適性」「機能面の過大評価」「機能完全性」「機能適切性」に細分化されるといったように，レビューにおける話題を，品質モデルの項目として説明できるようになった．今後は，他のプロジェクトに対してもレビューを分類し，手法の精練を行うことや，結果に対して主項目ごとの出現数をレーダーチャートで表現するなどの分析方法の提案を課題と考えている．

参考文献

[1] Leonard Hoon，Rajesh Vasa，Jean-Guy Schneider，Kon Mouzakis：“A Preliminary Analysis of Vocabulary in Mobile App User Reviews”，Swinburne University of Technology Faculty of Information and-Communication Technologies，pp.245-248，2012.

[2] Hammad Khalid，Emad Shihab，Meiyappan Nagappan，Ahmed E.　Hassan：“What Do Mobile App Users Complain About?”，In IEEE Software，Vol.32，No.3，pp.70-77，2015.

[3] 安部寛生，波多野雅信，小林佑汰：“日本のスマートフォンアプリケーションにおける評価の低いユーザーレビューでの苦情内容の分析”，南山大学理工学部 2017 年度卒業論文.

[4] 平井賢人，稲垣絢也：“スマートフォンアプリケーションにおけるユーザーレビューの内容の分析ー低評価レビューと高評価レビューの傾向の違いについてー”，南山大学理工学部 2018 年度卒業論文.

[5] 松永夏季：”スマートフォンアプリケーションのレビューにおける苦情の分析—地域による傾向の違いの調査—” 南山大学理工学部 2019 年度卒業論文.

[6] 小井土文哉，吉田稜：“スマートフォンアプリケーションのレビューにおける苦情の分析—同一アプリを対象とした場合の地域による傾向の違いに対しての考察—”，南山大学理工学部 2020 年度卒業論文.