

Be Careful of When:

An Empirical Study on Time-Related Misuse of Issue Tracking Data

Feifei Tu

Jiaxin Zhu

Qimu Zheng Minghui Zhou

ESEC/FSE '18, November 4–9, 2018, Lake Buena Vista, FL, USA

背景

- 使用者は意図した場面と一致しない問題追跡データ (ITD) を使っているかもしれない
 - 特にITDを使った予測モデルでは、問題属性が頻繁に変更されるためデータ漏洩が起きやすい[1]
- データ漏洩がITDベースの研究結果に対してどれほどの影響を与えるか不明である
 - 実験の正当性を脅かしたりや無関係な結果を導き出すかもしれない

[1]Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. 2012. Leakage in data mining: Formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data (TKDD) 6, 4 (2012), 15.

目的

- ITDを使う研究者にその使い方について勧告する
 - データ漏洩が研究の結果に及ぼす影響
 - 影響力の大きいデータ属性の特定

Issue Attribute

- 問題属性
 - Summary（概要）, Priority（優先順位）, Component（構成）などバグ管理システムの問題報告の詳細を説明する要素のこと
 - ライフサイクル中で変更される属性もある

Bugzillaの問題報告の例

Bug 1356733

▼

Wrong cursor shown over anchors without href attribute

Get help with this page

UNCONFIRMED Unassigned

▼ Status

Product: Core ▼

Component: CSS Parsing and Computation ▼

Type:  defect

Importance: P3 normal

Status: UNCONFIRMED

Reported: 2 years ago

Modified: 2 years ago

▼ People

Assignee: Unassigned

Reporter: [Andrew](#)

Triage Owner: [Sean Voisen \(:svoisen\)](#) [On PTO until May 6]

CC: 1 person

▼ Tracking

Version: 52 Branch

Target: ---

Points: ---

▼ Firefox Tracking Flags

This bug is not currently tracked.

▼ Details

Whiteboard: ---

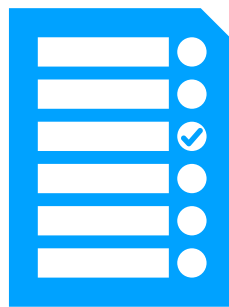
Votes: 0 votes

Data Leakage

- データ漏洩

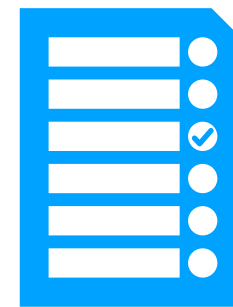
- 意図したデータを使っていない場合に起きるデータ漏れのこと
- 問題属性の値が頻繁に変更される場合に起きやすい

使いたいデータ



初期データ

実際に使ったデータ



修正データ



リサーチクエスチョン

- RQ1：既存の研究で、ITDの使用にデータ漏洩はあるか
- RQ2：問題報告がライフサイクル中で問題属性を変更する頻度と変更される可能性のある属性
- RQ3：データ漏洩による影響はあるか
- RQ4：RQ3でもしデータ漏洩の影響があった場合、最も影響力のある問題属性は何か



調査方法

RQ1:既存の研究でのITDの使用にデータ漏洩はあるか

- 1.2000~2017年のDBLPのデータベースから"bug report"と"issue report"で検索し， 関連があると判断した合計319件の論文を抽出[2]
- 2.その中から， 主要なソフトウェア/データ会議議事録およびジャーナルに掲載された110件の論文を選択
- 3.その中から， 研究に関係する43件の論文を抽出
- 4.それらの論文を参照し， さらに15件の関連論文を選択
 - 合計58件
- 5.それぞれのデータセットを調べて， データ漏洩があるか確認
 - データセットがない場合は， データ漏洩を避ける対策が講じられているか調査

[2]<https://dblp.uni-trier.de>

RQ2：問題報告がライフサイクル中で問題属性を変更する 頻度と変更される可能性のある属性

- 1.問題報告の変更率と変更頻度を定量化するために，MozillaとEclipseから関連プロジェクトのITDを収集
- 2.それぞれの問題属性の関連性を明らかにするために，それを使用している論文の数を調べる
- 3.属性が変更可能かどうかを判断し，各変更可能属性の変更率と変更頻度を計算する

$$\text{ChangeRate}(\text{attr}) = \frac{|I_{\text{attr}}|}{|I|} \quad \text{ChangeFrequency}(\text{attr}) = \frac{\sum_{i \in I_{\text{attr}}} T_{\text{attr}_i}}{|I_{\text{attr}}|}$$

I ：解決された全ての問題報告

I_{attr} ：ある属性が変更されて解決された全ての問題報告

T_{attr_i} ：ある問題報告にある属性のの変更数

RQ3：データ漏洩による影響はあるか

1.55件の論文の中から，最もよく研究されている3つのアプリケーションのそれぞれについて1つの論文を選択

- issue localization[3], 重複報告の検出[4],
問題修復時間の予測[5]

2.調査した論文を非リークデータとリークデータで再現し，結果からデータ漏洩の影響を調査

- Mozilla2016, Eclipse2010のデータダンプからITDを使用
- 修正データはMySQLダンプから抽出
- 初期データは問題の変更履歴を辿って修正データから復元

[3]Where should the bugs be fixed? more accurate information retrieval-based bug localization based on bug reports.

[4]Towards more accurate retrieval of duplicate bug reports.

[5]Filtering bug reports for fix-time analysis.

RQ4：最も影響力のある問題属性は何か

- 問題属性の変数を制御することで影響力の違いを明らかにする
 - 値が変更可能な属性の中から一つの属性をリークデータに設定し，残りの属性を非リークデータに設定して研究を再現し精度の差を計算
 - Perf_{mod}：修正データの精度（リークデータ）
 - Perf_{init}：初期データの精度（非リークデータ）
- 2.3件それぞれの論文で，全ての可変な属性に対して再現する

$$\frac{Perf_{mod}}{Perf_{init}} - 1$$



調查結果

RQ1:既存の研究でのITDの使用にデータ漏洩はあるか

- 抽出した論文の58件中3件は問題属性の可変性について、55件はITDのアプリケーションについて書かれていた
- 55件から11件の論文にデータ漏洩があることが確認できた

Table 2: Distribution of Definitely & Potentially Problematic Studies on Different Applications

Application	Definitely problematic papers	# of definitely problematic papers	Potentially problematic papers	# of potentially problematic papers
Issue Localization	[88] [28] [82] [53] [72] [77] [30] [84]	8	[39] [45] [15] [83]	4
Duplicate Detection	[59] [46]	2	[73] [60] [58] [14] [25] [69] [37] [38] [62] [35] [1] [19] [36] [81]	14
Issue Assignment	[23]	1	[7] [6] [9] [24] [40] [61] [49] [8] [80] [55] [56] [86] [70]	13
Issue-Fix Time Prediction	–	0	[75] [31] [4] [10]	4
Priority Prediction	–	0	[67] [68]	2
Severity Prediction	–	0	[33] [66]	2
API Recommendation	–	0	[65]	1
Classes Recommendation	–	0	[3]	1
Packages Recommendation	–	0	[22]	1
Configuration Bugs Prediction	–	0	[76]	1
Blocking Bugs Prediction	–	0	[71]	1
Total	–	11	–	44

RQ2：問題報告がライフサイクル中で問題属性を変更する 頻度と変更される可能性のある属性

- 18のうち15種類の問題属性が変更されることがわかった
- 変更頻度は全ての可変属性で1より大きく，頻繁に変更されている可能性がある
 - 研究結果がデータ漏洩によって影響を受けている可能性がある

Table 3: 18 Attributes in 55 Investigated Studies

Attributes	# of occurrences	Could it be changed?	Change rate%		Change frequency		Interpretation
			Mozilla	Eclipse	Mozilla	Eclipse	
Description	48	F	–	–	–	–	Detailed outline of the issue, such as what is the issue and how it happens.
Summary	46	T	16.49	20.16	1.2912	1.2088	A one-sentence summary of the problem, displayed in the header next to the issue number.
Component	15	T	22.83	17.45	1.2811	1.2310	Issues are divided up by <i>product</i> and <i>component</i> , with a <i>product</i> having one or more <i>components</i> in it.
Product	13	T	30.59	6.62	1.2050	1.1317	
Priority	12	T	9.70	13.42	1.1889	1.2323	The <i>priority</i> attribute is used to prioritize issues, either by the <i>assignee</i> , or someone else with authority to direct their time such as a project manager. The default values are P1 to P5.
Severity	9	T	7.10	9.06	1.1432	1.1102	The <i>severity</i> attribute indicates how severe the problem is – from blocker ("application unusable") to trivial ("minor cosmetic issue"). You can also use this field to indicate whether an issue is an enhancement request.
Version	9	T	10.49	6.95	1.1404	1.1820	The <i>version</i> attribute usually contains the numbers or names of released versions of the product. It is used to indicate the version(s) affected by the issue report.
Comment	9	T	99.06	99.95	8.4384	4.7831	Discussion about the issue.
Reporter	6	F	–	–	–	–	The person who reported the issue.
Report Time	6	F	–	–	–	–	The data and time when the issue is reported.
OS	5	T	9.55	2.70	1.0516	1.0262	These indicate the computing environment where the issue was found.
Platform	4	T	8.05	1.39	1.0348	1.0171	
Type ⁷	4	T	–	–	–	–	The type of the report, i.e., defect, task, feature.
Assignee	2	T	39.36	58.72	1.3519	1.4778	The person responsible for fixing the issue.
Keywords	2	T	19.14	6.35	1.9612	1.1674	The administrator can define keywords which you can use to tag and categories issues, e.g., crash or regression.
Attachments	1	T	36.00	28.00	2.6574	1.8663	Attached files (e.g. test cases or patches) to issues.
CC List	1	T	76.72	46.85	3.4213	2.0520	A list of people who get mail when the issue changes, in addition to the Reporter, Assignee and QA Contact.
Target Milestone	1	T	24.61	56.22	1.4210	1.5009	A future version by which the issue is to be fixed.

RQ3：データ漏洩による影響はあるか

- issue localizationに関する論文では・・・
 - ファイルと問題報告のsummaryとdescription属性との類似性の計算
 - MRR(mean reciprocal rank)とMAP(mean average precision)でモデルを評価, 再現率を計算 (Table5)
- Initial data(非リークデータ)よりもModified data(リークデータ)の方が評価が高いことがわかった
 - 重複レポートの検出, 問題修復時間の予測に関する論文でも結果に影響が出ることがわかった

Table 5: Performance (MRR and MAP) for Issue Localization [88] (Eclipse)

Eclipse	MRR	MAP
Initial data	0.39893	0.24767
Modified data	0.42458	0.26303

RQ4：最も影響力のある問題属性は何か

- Issue localizationを調査した論文
 - 可変属性がsummaryしかなかったため，Table5よりsummaryが大きな影響を与えていることがわかる
- 重複報告の検出を調査した論文：Table8

Table 8: Impact of Attributes (MAP) on Duplicate Detection [59] (Mozilla and Eclipse)

MAP	Modified data	Initial data			
		<i>summary</i>	<i>product</i>	<i>component</i>	<i>priority</i>
Mozilla	0.247306	0.230845	0.24589	0.240918	0.24742
$(\frac{Perf_{mod}}{Perf_{init}} - 1)\%$		<u>7.130759</u>	0.575867	<u>2.651525</u>	-0.04608
Eclipse	0.323553	0.321243	0.324447	0.316071	0.32418
$(\frac{Perf_{mod}}{Perf_{init}} - 1)\%$		<u>0.7191</u>	-0.2756	<u>2.3672</u>	-0.1934

RQ4：最も影響力のある問題属性は何か

- 問題修復時間の予測を調査した論文
 - 影響力：Assignee > Component, Priority, Severity > OS, Platform

Table 9: Impact of Attributes (AUC) on Issue-Fix Time Prediction [59] (Mozilla and Eclipse)

Project	Modified data	Initial data (AUC / ($\frac{Perf_{mod}}{Perf_{init}} - 1$)%)					
		Assignee	Component	OS	Platform	Priority	Severity
Mozilla Core	0.744	0.711/ <u>4.6414</u>	0.752/ <u>-1.0638</u>	0.743/0.1346	0.744/0	0.741/0.4049	0.743/0.1346
Mozilla Bugzilla	0.958	0.938/ <u>2.1322</u>	0.957/0.1045	0.96/-0.2083	0.96/-0.2083	0.957/0.1045	0.958/0
Mozilla Firefox	0.701	0.677/3.5451	0.675/ <u>3.8519</u>	0.704/-0.4261	0.704/-0.4261	0.701/0	0.701/0
Mozilla Thunderbird	0.666	0.619/ <u>7.5929</u>	0.668/-0.2994	0.666/0	0.668/-0.2994	0.665/0.1504	0.664/0.3012
Mozilla Seamonkey	0.898	0.887/1.2401	0.898/0	0.898/0	0.898/0	0.898/0	0.898/0
Eclipse Platform	0.691	0.723/ <u>-4.4260</u>	0.691/0	0.693/-0.2886	0.69/0.1449	0.673/ <u>2.6746</u>	0.682/ <u>1.3196</u>
Eclipse PDE	0.659	0.671/-1.7884	0.658/0.1519	0.659/0	0.659/0	0.652/ <u>1.0736</u>	0.653/0.9188
Eclipse JDT	0.672	0.685/-1.8978	0.67/0.2985	0.675/-0.4444	0.672/0	0.659/ <u>1.9727</u>	0.655/ <u>2.5954</u>
Eclipse CDT	0.776	0.806/ <u>-3.7221</u>	0.765/ <u>1.4379</u>	0.776/0	0.776/0	0.775/0.1290	0.773/0.3881
Eclipse GEF	0.666	0.662/0.6042	0.667/-0.1499	0.659/1.0622	0.67/-0.5970	0.664/0.3012	0.645/ <u>3.2558</u>

まとめと所感

- ITDを使ったアプリケーションにデータ漏洩があり，それが研究結果に影響を与えることがわかった
- 問題属性の変化によって起こるデータ漏洩のうち，Summary, Component, Assigneeの変更が結果に最も大きな影響を与えることがわかった
- データ漏洩を防ぐために，より精密な問題追跡システムが必要だと思った
- 問題追跡データから遅延相関分析をする時の参考になればいいと思った

おしまい



シャルル