DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics

Machine Learning of Bug Pattern Detection Rules

Nils-Jakob Kunze

DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics

Machine Learning of Bug Pattern Detection Rules

Maschinelles Lernen von Analyseregeln zur Erkennung von Fehlermustern

Author:	Nils-Jakob Kunze
Supervisor:	Prof. Dr. Dr. h.c. Manfred Broy
Advisor:	Dr. Elmar Jürgens, Dr. Benjamin Hummel, Dr. Lars Heinemann
Submission Date:	May 15, 2018

I confirm that this master's thesis is my own work and I have documented all sources and material used.

Munich, May 15, 2018   Nils-Jakob Kunze

Acknowledgments

Abstract

Software reuse is a central pillar of software development, enabling developers to rely on existing frameworks and libraries to do most of the heavy lifting. Frameworks and libraries attempt to be as easy to use as possible, but often they provide so much functionality that their application programming interface (API) cannot remain trivial. This complexity and the fact that the documentation is often not kept up to date can lead to errors when developers invoke these APIs.

In the context of object-oriented programming, one of the errors related to API usages are missing method calls. They occur when a developer correctly instantiates some object but forgets to call one or more of the necessary methods. We study a new method for detecting missing method calls in Java applications which was proposed by Monperrus et al. [12]. The main goal is to understand how good this method is at detecting code smells and true bugs when applied to a large software system.

To investigate this, we analyze more than 600 open source android applications. We manually evaluate the proposed findings of X randomly selected apps and also use an automatic benchmark which relies on artificially degrading existing code. We, also, compare the originally proposed technique to some slight variations of it and find [. . .]. The results are kind of mixed, but point in the direction of this method being marginally useful.

still not a fan of this paragraph - is it even necessary? / shorten it?

include any more information on the actual method? aka "relies on the majority rule / the insight that. . ." oä?

more concrete research question

describe results + concrete numbers

iv

Contents

1 Introduction

Developers often outsource a significant amount of work to existing libraries or software frameworks, which expose their functionality through an application programming interface (API). These APIs would ideally be simple and easy to use, but there is always a trade-off between usability and flexibility. Especially more powerful frameworks cannot provide a trivial API without limiting developers ability to use their functionality in the way most appropriate for the use case at hand. As a result, invoking an API correctly often requires a lot of knowledge about it. It can be of constraints or requirements which are not clear from the outset, but which have to be heeded to avoid serious bugs or complications, or about an involved interplay between different parts of the library. In the worst case, this knowledge might not even be contained in the API documentation.

Even if the documentation is of high quality, the complexity of some libraries makes it inevitable that there will be erroneous invocations of their APIs. The mistakes can relate to parameter choice, method order, or many other factors (e.g., some methods must be invoked in an extra thread, some specific precondition has to be satisfied, or some setup work must be performed). Examples, which come to mind in Java, are a programmer calling `next()` on an iterator without first checking with `hasNext()` if it even contains another object, or a class which overrides `equals()` without ensuring that an invocation of `hashCode()` always produces the same output for two equal objects.

Despite the fact that there might be numerous appropriate ways to use an API, there are underlying patterns which the correct invocations have in common. These patterns can take a lot of different forms and shapes, for example "call method `foo` before calling method `bar`", "if an object of type `Foobar` is used as a parameter to method `baz`, condition X has to be fulfilled", "never call method `qux` in the GUI thread", etc. For the Java examples mentioned above they could take the form of "an object which implements `equals()` must also implement `hashCode()`", "if object A equals object B, their `hashCode()` must also be equal" or "a call to `next()` on an iterator should be preceded by a call to `hasNext()` to ensure that it actually contains another object".

Patterns like these are also called API usage patterns [18] and they can aid in detecting potential defects. If some code in a software project deviates (too much) from the usual patterns when using an API, this can hint at a bug or is at least a code smell which should probably be corrected. Because of this, it is interesting to detect these unusual instances.

1.1 Motivation

consider the structure suggested by Williams: prelude, shared context, problem, solution, not yet 100 percent happy

As already mentioned above, there are many subtle mistakes a developer can make when invoking an API. In this work, we focus on one specific type of API usage problem, namely missing method calls, which can occur in the context of Object Oriented Programming (OOP). In OOP software, an object of a specific type is usually used by invoking some of its methods. Types will then have underlying patterns such as: "when methods `A` and `B` of type T_1 are invoked, then method `C` is called as well" or "methods `X` and `Y` of type T_2 are always used together". Given an object of this type T_1 on which only methods `A` and `B` are called, we can say that a call to `C` is missing. Respectively if we have an object of type T_2 where only `X` or only `Y` is invoked, we can say that a call to the other is missing.

Developers will fail to make important method calls when they are using classes with which they are not very familiar. As this can happen whether they are working with a new library, an extensive framework or even a whole platform (e.g., Android), we will use these words interchangeably through this work. These unfamiliar classes can be from an external source, but this is not a necessity. In large software projects it is common that one developer does not know the whole codebase, and thus, they will often encounter classes which they do not know how to use correctly. Altogether it seems simple to make an error related to missing method calls.

We can also confirm this intuition with hard data. In an informal review, Monperrus et al. found bug reports[1] and problems[2] related to missing method calls in many newsgroups, bug trackers, and forums. The issues range from runtime exceptions[3] to problems in some limit cases, but generally reveal at least a code smell if not worse. In addition to the informal review, Monperrus et al. [13] also did an extensive analysis of the Eclipse bug repository. First, they searched for syntactic patterns which they deemed related to missing method calls, such as: "should call", "does not call", "is not called" or "should be called". Manual inspection then confirmed 117 of the 211 (55%) obtained bug reports as indeed related to a missing method call.

These results show that even mature code bases can contain many bugs related to missing method calls, especially considering that this number is a lower bound on the total number of related bugs in the repository. After all, they might have missed some syntactic patterns and bugs which have not even have been discovered yet. Together, this makes it highly desirable to be able to automatically detect missing method calls in production code, not only to save developer time but also to make maintenance cheaper

---

[1]`https://bugs.eclipse.org/bugs/show_bug.cgi?id=222305`
[2]`https://www.thecodingforums.com/threads/customvalidator-for-checkboxes.111943/`
[3]`https://issues.apache.org/jira/browse/TORQUE-42`

and more manageable.

1.1.1 Detection - but how?

A simple and straightforward approach for detecting missing method calls would be to build a set of hard-coded rules, such as:

- "always call `setControl()` after instantiating a `TextView`"

- "in Method `onCreate()` of classes extending `AppCompatActivity` always call `setContentView()`"

- "when calling `foo()` also call `bar()`"

- "when calling `next()` on an Iterator always call `hasNext()` (before)"

Well-crafted and thought-out rules along these lines could facilitate a very high precision in detecting missing method calls and contribute to better, more bug-free code. However, creating and maintaining a list of rules like this would require a tremendous investment of time and money, especially in a world where software is changing continually. While the necessary effort might be justified for large and important libraries, it multiplies with the size of the library until it becomes completely infeasible.

To circumvent this problem, we would like to automatically detect locations in a code base where a method call is potentially missing using only the code itself as input. Such an approach would adapt to changes without requiring additional work from a developer and could also be applied to proprietary code which is closed to the public. While the discovered locations will probably not be as accurate as those discovered by a hand-crafted list of rules, they could then be examined by an expert who would determine the severity of the finding and issue a fix if necessary.

1.2 Contribution

In this thesis, we present a re-evaluation of the type usage characterization first introduced by Monperrus et al. [12] and further refined in a follow-up publication [13]. A type usage is the list of method calls which are invoked on an object of some type and occur in the body of some method. The general idea behind the technique by Monperrus et al. is to check for outliers among the type usages by using the majority rule: If a type is used in one particular way many, many times (that is, in the majority of cases) and differently only one (or a few) times, this probably indicates a bug.

We evaluate this concept by applying it to a data set of more than 600 open source android applications and performing a manual evaluation of the results. We further

experiment with small changes to their initial idea and put them under the scrutiny of an automated benchmark. Additionally, we compare the results of the manual evaluation against those of the automated one and consider what this means for future research.

actually include this?

mention some results!

In Chapter 2 we present previous work which goes in a similar direction. Chapter 3 explains the theoretical background of the method proposed by Monperrus et al. and some variations which we will compare in the evaluation. We explain the inner workings of our implementation in Chapter 4, before summarizing the details and results of the evaluation in Chapter 5. The last Chapter 6 concludes this work and gives an outlook for future research.

2 Related Work

generally in this section: think about the tense I wanna use, present / past and if it works well! + stick to it! -> right now vote for present!

new ideas: mention code smells and their origin?, less focus on "bugs" per se (see aDoctor paper) -> while the thesis title is related to bugs, I think I can / want to use smell as well Reference die belegt, dass es eine Korrelation zwischen bugs und smells gibt look in fontana2016comparing, has references to this in introduction + adoctor as well

It is difficult and time-consuming to fix bugs in software. This is because the maintainer has to realize that there is a bug, identify its cause and understand the steps necessary to remove it. Especially for larger software systems automatically detecting low-quality code and improving it can contribute significantly to the maintainability of the code and prevent bugs in the future. Because of this, there has been a lot of interest in approaches for automatically detecting bugs, or even just smells in code.

In this chapter, we present a number of different approaches for smell and bug detection. We start with some "conventional" methods which mostly rely on hard-coded rules and patterns. Since we are applying the method studied in this thesis to some Android apps in Chapter 5, we then give a quick overview of smell detection related explicitly to android applications. Finally, we look into techniques which attempt to learn rules and properties from the code they are analyzing instead of relying on rules which their developers defined a priori. These automated techniques need fewer changes when the system under analysis evolves, but they might not be as accurate.

2.1 Finding Code Smells with Static Analysis

Findbugs[1] is a static analysis tool that finds bugs and smells in Java code. It reports around 400 bug patterns and can be extended using a plugin architecture. Each bug pattern has a specific detector, which uses some special, sometimes quite elaborate detection mechanism. These detectors are mostly built starting from a real bug, first attempting to find the bug in question and then all similar ones automatically as well.

---

[1] `http://findbugs.sourceforge.net/`, in the meantime there has been a successor: `https://spotbugs. github.io/`

In their work Ayewah et al. [1] apply it to several large open source applications (Sun's JRE[2] and Glassfish J2EE server[3]) and portions of Google's Java codebase. Their premise is, that static analysis often finds true but trivial bugs, in the sense that these bugs do not actually cause a defect in the software. This can be because they are deliberate errors, occur at locations which are unreachable or in situations from which recovery is not possible. In their analysis Findbugs detects 379 medium and high priority warnings which they classify as follows:

- 5 are due to erroneous analysis by Findbugs

- 160 are impossible or have little to no functional impact

- 176 can potentially have some impact

- 38 are true defects which have substantial impact, i.e., the real behavior is clearly not as intended

Their main takeaway is that this kind of static analysis can find a lot of true bugs, but there will also be a lot of false positives among them, and it can be difficult to distinguish them.

To alleviate the problem of a high false positive rate among the findings reported by Findbugs, Shen et al. [19] propose a ranking method which attempts to rank true bugs before less important warnings. It is based on a principle they call "defect likelihood", which is essentially the probability that a finding is a real finding. They calculate this probability using the findings of a large project (in this case the JDK), each of which they manually flag as either a true or a false positive. It can not only be calculated for one specific bug pattern but also across categories and types with variance as a tiebreaker. The resulting ranking can be refined with user feedback when it is applied to a specific project. In their evaluation on three open source applications (Tomcat[4], AspectJ[5] and Axis[6]) they compare their ranking against the default severity ranking of Findbugs, which uses a hard-coded value for each bug type. With cutoffs at 10%, 20%, 30%, . . . of the total findings they achieve precision and recall systematically better than the default ranking. This especially holds true for cutoff values around 50%.

---

[2]Version 1.6.0 (different builds) `http://www.oracle.com/technetwork/java/javase/downloads/java-archive-downloads-javase6-419409.html`
[3]v2 `http://www.oracle.com/technetwork/java/javaee/downloads/java-archive-downloads%2Dglassfish-419424.html`
[4]Version 6.0.18 `http://tomcat.apache.org/`
[5]Version 1.6.4 `http://eclipse.org/aspectj/`
[6]Version 1.4 `http://jlint.sourceforge.net/`

Which other papers?! -> should be some which are NOT about findbugs -> present alterantives, maybe work in a slightly different manner, or just some sort of overview paper for static code smell detection oä!

Other tools for detecting code smells are Decor[11] and iPlasma[10].

some more info?

2.2 Android specific Smell Detection

smth more here?!

While not investigating Android specific smells, Verloop [20] used Java refactoring tools (e.g. PMD and JDEODORANT) to detect basic code smells like large class or long method. interesting finding: classes inheriting from the Android framework smells can appear at different frequencies example: long method detected 2times as much in these classes . . .-> some more info? + references for pmd and jdeodorant

Hecht et al. [8] present an approach they call PAPRIKA which operates on compiled Android applications but tries to infer smells on the source code level. From the byte-code analysis, it builds a graph model of the application and stores it in a graph database. The nodes of the graph are entities such as the app itself, individual classes, methods and even attributes and variables. These nodes are then further annotated with specific metrics relevant to the current abstraction level, e.g. "Number of Classes" for the app, "Depth of Inheritance" for a class or "Number of Parameters" for a method. Finally, they extract smells using hand-crafted rules written in the Cypher query language[7]. This enables them to recognize four Android specific smells and four general, OOP related smells.

To evaluate this approach, the authors use a witness application, which contains 62 known smells. Using the right kind of metrics, they achieve precision and recall of 1 on this application. Further, they apply their tool to a number of free Android apps in the playstore and make some assertions about the occurrence of antipatterns in publicly available applications. One finding to emphasize is that some antipatterns, especially those related to memory leaks, appear in up to 39% of applications, even "big" ones like Facebook or Skype.

After Reimann et al. [17] presented a catalog of 30 Android specific code smells, Palomba et al. [15] developed a tool to detect 15 of them. It works on the abstract syntax tree of the source code and uses specifically constructed rules to detect each of the smells. For the evaluation of their approach, they analyze 18 Android applications and compare the results of their tool against a manually built oracle. The oracle was created by having two Masters students study the code of the applications in depth and

---

[7]`https://neo4j.com/developer/cypher-query-language/`

manually flag offending locations. While they reach an average precision and recall of 98%, there are some cases in which their tool fails or yields false positives.

One especially interesting case is the smell of 'missing compression'. It advises to compress the data when making external requests in order to save bandwidth. To detect places in the code where a request is made, but compression is missing, their hard-coded rule check if the developer used one of two popular compression libraries. However, recently a competing compression library has been gaining in popularity. Their rule does not include it and, thus, falsely flags the usage of this library as the "missing compression" smell. This is an excellent example of a case where hard-coded rules fail because they are not kept up to date (new library becomes popular, interface changes, ...), do not consider a special case or conditions change in some other way.

2.3 Inferring Properties from the Code at Hand

> General: read the summary paper [18] again and check which approaches are there and which could / should be mentioned + check to read stuff... + for more information: check this summary paper / the 2017 one? / both?

As the results of the previous paper showed, it can be difficult and costly to keep hard-coded detection rules up to date and relevant. Because of this, there has been interest in approaches which adapt automatically to changing requirements by learning rules from the source code and looking for violations of those rules.

> -> not perfectly happy with this summary, but we shall see Mention [4] as probably the first paper which proposed the general idea behind DMMC -> learning from the code at hand, instead of using predefined static rules "finding what's incorrect without knowing what's correct" -> contradiction and common behavior
> static analysis to infer what programmers believe about the system state, then check them for contradictions they provide several "templates" for believes a programmer must or may hold use these to extract rules from the code these believes can be about a number of things (a pointer not being null, a lock being unlocked, etc), each template is cartered for a specific type of believe can also look for common behavior and exceptions to them
> results: evaluation on linux and openbsd finds hundereds of bugs related to the different bug types their templates can find

> [Why is this potentially better than other approaches (actually learning an api vs static rules comes to mind)] some more übergang... + what other papers here?

Pradel et al. [16] noticed that in statically typed languages the compiler helps the programmer with passing method arguments in the correct order. However, the compiler cannot ensure the correct order of equally typed arguments (such as `setEndPoints(int`

`low, int high)`. Confusing the order of these arguments can cause unforeseen and hard to detect errors. Their idea is to detect mixed up method arguments using only semantic information in the source code, namely variable and parameter names. They extract these and compare the names found at the call sites to the names from the implementation body using string similarity metrics. If reordering the names yields a significantly better similarity score, they report an anomaly. In their evaluation on 12 Java programs from the DaCapo Benchmark suite [3] (more than 1.5 Million lines of code), they reach a precision of 72% and recall of 38% on seeded (i.e. "faked") anomalies. In real code, they found 29 anomalies of which 22 revealed true problems (76%).

In a more general attempt Fontana et al. [6][5] compare 16 different machine learning techniques for code smell detection. As their training data, they manually evaluate a large body of code and tagged four basic smells (data class, large class, feature envy and long method). All of the different approaches show good performance, but J48 and random forest have the best results, while support vector machines fare the worst.

2.4 Focus on Object Usages

> title sucks?! -> api misuses?

> Mention MUBench?! -> useful classification of api misuses Dataset of bugs related to API misuses . . .

— API methods are often complex and thus hard to use, additonalyl might be badly documented to allivieate this, the workflow of developers often includes searching for examples of how to use an particular api [22] built primarily a SEARch tool for code snippets explaining how tu use a particular API but to search over API usages, one has to extract them! Their tool MAPO extracts API usage patterns which describe how in a certain scenario some API methods are frequently called together and if the usage follows any sequential rules -> this tool doesn't look for any anomalies or code smells, but it is important as it extracts information similar to the one we are interested in. uses clustering to group the extracted api usages into patterns (+ frequent subsequences miner) since they focus more on the code search aspect their evaluation is a user study investigating the quality of produced code (with / w/o their tool), but this is intresting anyways. . .

[12][13] There have been some works concerned with finding patterns in object usages and using those to find potential bugs. what are object usages?! explain somewhere? Most relevant for this work and its primary inspiration are two papers by Monperrus et al. who propose technique for detecting missing method calls it is based on the intuition that types are often used in particular ways If a type usage has a large number of very similar type usages but a small number of exactly similar ones, it is likely that it is

missing a method call. They call this the "majority rule", it is explained in detail in Chapter 3.

first, Evaluation based on a simulation which creates defects by degrading real software that is it removes singular method calls from the code and sees if their method detects these known missing calls this on several software packages: eclipse userinterface + two other packages, derby and tomcat -> result: system can answer 54 to 76 percent of the queries of the queries it answers, 73 - 89 percent of the answers contain the actually missing call (not necessary as the first recommendation) all in all it reaches precision between 59 and 83 and recall of 41 to 68 however, questionable how relevant these numbers are b/c its not clear that real bugs will share characteristics with the artificially created cases

for this reason: additionally qualitative eval: throw the developed tool onto eclipse, derby and tomcat and manually analyze 30 very anomalous cases reported by it they find that there can be different reasons for a TU to be anomalous (did i explain tu so far? / use oject usage?) ideal case: method call is actually missing (they reported 17 issues and got 9 patches accepted!) it's just a smell ie situation could be rewritten or refactored for more clarity sometimes it's just related to software aging: outdated usage related to old version of API for example however, also negative results, where the anomaly is just that: an outlier, which is correct, even if a majority of other usages are different or simply related to dead code which isn't used anymore anyways —

Before this, Wasylkowski et al. [21] introduced a method to locate anomalies in the order of method calls. First, they extract usage models from Java code by building a finite state automaton for each method. The automata can be imagined similarly to the control flow graph of the method with instructions as transitions in the graph. From these, they mine temporal properties, which describe if a method A can appear before another method B. This process determines if there exists a path through the automata on which A appears before B, which in turn implies that a call to A can happen before one to B. Finally, they are using frequent itemset mining [7] to combine the temporal properties into patterns.

In this work, an anomaly also occurs when many methods respect a pattern, and only a few (a single one) break it. In their experiments, they find 790 violations when analyzing several open source programs. Manual evaluation classifies these into 2 real defects, 5 smells and 84 "hints" (readability or maintainability could be improved). This adds up to a false positive rate of 87.8%, but with the help of a ranking method, they can obtain the 2 defects and 3 out of 5 smells within the top 10 results.

In a related work, Nguyen et al. [14] use a graph-based representation of object usages to detect temporal dependencies. This method stands out because it enables detecting dependencies between multiple objects and not just one. Here the object usages are represented as a labeled directed graph in which the nodes are field accesses, constructor

or method calls and branching occurs because of control structures. Thus, the edges of the graph represent the temporal usage order of methods and the dependencies between them. They mine patterns using a frequent induced subgraph detection algorithm which builds larger patterns from small patterns from the ground up, similar to the way merge sort operates. Here an anomaly is also classified as a "rare" violation of a pattern, i.e., in relation to its size it does not appear frequently in the dataset. In an evaluation case study the authors find 64 defects in 9 open source software systems which they classify as 5 true defects, 8 smells and 11 hints, equalling a false positive rate of 62.5%. Using a ranking method the top 10 results contain 3 defects, 2 smells and 1 hint.

is this merge sort thing true?!

mention more related work from detecting mmc paper?

2.5 Other Approaches

change title?

As already mentioned, incorrect API documentation can be a big problem. It often occurs when an API developer changes the code but forgets to adapt the documentation accordingly. If a user of the API then relies on the documentation, she can run into unforeseen bugs, and what is worse, the documentation might even hinder her. To alleviate this problem, Zhou et al. [23] propose an automated approach for detecting "defects" in the API documentation. They consider two situations: either the documentation is incomplete in describing the constraints at hand, or the description exists, but it does not match the realities of the code. They make one crucial assumption: the code is correct because contrary to the documentation has been tested extensively.

Their approach analyzes the code statically and uses pattern-based natural language processing to build a first order logic formula of the constraints which are present. For the API documentation, they use heuristics and a part-of-speech tagger to find the restrictions and constraints which are mentioned, before synthesizing them into another first order logic formula. Finally, the resulting formulas are fed into an SMT (satisfiability modulo theories) solver [2] to detect inconsistencies between them. With this approach, can detect four classes of documentation defects. They apply their prototype implementation to the `java.awt` and `java.swing` packages and find 1419 defects, of which 81.6% are true positives. Using the heuristics extracted from these packages they also apply the prototype to some other packages and find 1188 defects of which 659 are true positives (a precision of 55.5%). Overall it is worrying (if not surprising), that there are so many inconsistencies in the JDK documentation, which even has a reputation for being of rather high quality. Less well-known APIs will probably have inferior documentation and thus even further increase the risk of API usage relates bugs.

too detailed?

3 Detecting Missing Method Calls

analogy: remove with waiter irgendwo erläutern PART 1 ist background (Foundation: Majority rule) 3.2 = mein teil

better intro. . .

In the introduction, we explored how easy it is to make an error related to missing a method call. Especially when a developer is working with complicated frameworks and libraries, it is easy to make a mistake. We would like to automatically detect instances of missed method calls to make maintenance easier and cheaper. The main idea is to extract patterns from the code at hand and then look for violations of these patterns. — . . . — As mapped out in Section 1 bugs related to missing method calls occur in real software systems. These bugs are often related to complicated APIs and frameworks, and we would like to detect them. The idea is to detect patterns in correct usages which will enable us to single out the faulty ones.

In this chapter we give a proper explanation of the method proposed by Monperrus et al. [12][13] which is based on the majority rule. This includes a proper qualification of the notions of type usage, as well as their similarity and almost similarity. We will then explain how the strangeness score, which describes the degree to which a particular type usage is different from its relevant neighbors, is calculated, before going into detail how the potentially missing calls are determined.

In the second section of this chapter, we explain some minor tweaks to the type usage notion, whose ability to detect missing method calls we explore further in Chapter 5.

more details? -> when proper contents are decided + do not mention chapter 5

3.1 Foundation: The Majority Rule

make sure that chapter intro + this section intro work well together

The intuition behind the majority rule is that a piece of code is likely to contain a defect if there are none or very few instances which are 'equal', but a lot of slightly different instances. -> what does this similarity mean?! conceptually! It then seems likely that the code which appears a selected few times only should be adapted in a way to match the majority of slightly different instances.

> use methods example, maybe potentially the one with comic sans

`->` in the context of method calls, we are mostly interested in the concrete method calls which are invoked `->` ka? `->` remove?! / use with methods! An analogy which explains the idea well goes as follows. Imagine being the waiter preparing the tables in a restaurant. There are 100 seats, and each of them is equipped with a plate, a knife on the right and a fork on the left. However, of these seats, only 99 have a spoon above the plate, and there is one seat s which is missing the spoon. Then it is highly likely, that this one exception to the rule "each seat should have a plate, a knife, a fork and a spoon" is a mistake, and you should add the spoon to the single seat s where it is missing. In this context, each piece of cutlery (knife, spoon, fork) can be imagined as a method call, i.e., we extend the concept of the majority rule to types and method calls.

The remainder of this section formalizes the different notions required to implement a system detecting such places.

3.1.1 Type Usages

Generally speaking, type usages are an abstraction over the analyzed code. They ignore the order of method calls and are only concerned with the list of method calls invoked on some object. The critical information which is associated with a type usage is the type of the object, the list of methods invoked on it and the context in which the object is used. We define the context as the name of the method in whose body the type usage appears together with the type of its parameters.

Formalized, given some code the type usages we extract as follows: For each variable x, note its type $T(x)$ and the context $C(x)$ in which it occurs. Additionally, save the list of methods which are invoked on x within $C(x)$: $M(x) = \{m_1, m_2, \ldots, m_n\}$. If there are two variables of the same type, this results in two type usages being extracted (unless they refer to the same object, more on this in Section 4.2).

As an example, consider the code in Figure 3.1. It contains two type usages, one of type `Button` and another one of type `Text`. The context for both is `createButton()`. The methods invoked on the `Button` object are initialization (`new`), `setText` and `setColor`. The `Text` object, on the other hand, is only initialized. Given the two variables b and t as input, the corresponding values of $T(x)$, $C(x)$ and $M(x)$ are shown on the right hand side.

3.1.2 Exact and Almost Similarity

To detect type usages which are anomalous by the majority rule, we now need a notion what it means for two type usages to be exactly similar or only almost similar.

> explain better what a SIMILAR method is (if so mention what it is) + SAME method

```
class A extends Page {
    Button b;

    Button createButton() {
        b = new Button();
        b.setText("hello");
        b.setColor(GREEN);
        ... (other code)
        Text t = new Text();
        return b;
    }
}
```

T(b) = 'Button'
C(b) = 'Page.createButton()'
M(b) = {<init>, setText, setColor}

T(t) = 'Text'
C(t) = 'Page.createButton()'
M(t) = {<init>}

Figure 3.1: Example illustrating the Extraction of Type Usages (taken from [13])

Informally, we say that two type usages are exactly similar, iff they have the same type, the type usage appears in a similar method (i.e., the context is identical) and if their list of method calls is identical. Recall that the context is the name of the containing method together with the type of its parameters. The reason we call these type usages "similar" instead of "equal" because several things are not the same: variable names, surrounding or interspersed code, method order and the actual location (class) are all disregarded and might indeed differ.

The notion of almost similarity is analogous. We say that a type usage y is almost similar to a given type usage x if they share the same type and context and the list of method calls of y is the same as the method calls of x plus one additional call.

Consider the example code snippets in Figure 3.2. The type usages of `Button` in classes `A` (top left) and `B` (top right) are exactly similar, because they not only appear in the similar method `createButton`, but also invoke the same methods `setText` and `setColor`. Observe that variable names and method parameter have no impact on this. Additionally, it is not influenced by the order of methods or any interspersed code which does not call a method on the `Button` objects. The type usage in class `C` (bottom left) is almost similar to those in `A` and `B`, because besides the two methods `setText` and `setColor` it also invokes one additional method: `setLink`. Finally, the type usage in class `D` (bottom right) invokes **two** additional methods in comparison to the ones in `A` or `B`, meaning that it is not almost similar to them. However, it is in fact almost similar to the type usage in class `C`, since it only invokes one additional method in comparison: `setTooltipText`.

To formalize these notions, we define two binary relationships over type usages. The relationship for exact similarity is called E' and we say that two type usages x and y are

some better explanation WHY we need this / add it in section intro?

```
class A extends Page {
    Button createButton() {
        Button b = new Button();
        ... (interlaced code)
        b.setText("hello");
        ... (interlaced code)
        b.setColor(BLUE);
        return b;
    }
}
```

```
class B extends Page {
    Button createButton() {
        ... (code before)
        Button aBut = new Button();
        ... (interlaced code)
        aBut.setColor(RED);
        aBut.setText("goodbye");
        return b;
    }
}
```

```
class C extends Page {
    Button myBut;
    Button createButton() {
        Button myBut = new Button();
        myBut.setColor(ORANGE);
        myBut.setText("Great Company!");
        myBut.setLink("https://www.cqse.eu");
        ... (code after)
        return b;
    }
}
```

```
class D extends Page {
    Button createButton() {
        Button tmp = new Button();
        tmp.setLink("https://slashdot.org/");
        ... (interlaced code)
        tmp.setText("All the news");
        tmp.setColor(GREEN);
        tmp.setTooltipText("Click me!");
        return b;
    }
}
```

Figure 3.2: Examples of similar and almost similar Type Usages (also inspired by [13])

exactly similar if and only if:

$$
\begin{aligned}
xE'y \iff T(x) &= T(y) \ \wedge \\
C(x) &= C(y) \ \wedge \\
M(x) &= M(y)
\end{aligned}
$$

Given this, the set of exactly similar type usages in relation to x is defined as:

$$E(x) = \{y \mid xE'y\}$$

Observe that this relation holds true for the identity, i.e., $xE'x$ is always true and $\forall x : x \in E(x)$ (and thus $|E(x)| \geq 1$).

For almost similarity, we define the relation A' which holds if two type usages are almost similar. A type usage y is almost similar to a type usage x iff:

$$
\begin{aligned}
xA'y \iff T(x) &= T(y) \ \wedge \\
C(x) &= C(y) \ \wedge \\
M(x) &\subset M(y) \ \wedge \\
|M(y)| &= |M(x)| + 1
\end{aligned}
$$

Similarly to $E(x)$ we define $A(x)$ as the set of type usages which are almost similar to x:

$$A(x) = \{y \mid xA'y\}$$

In contrast to $E(x)$, $A(x)$ can be empty and $|A(x)| \geq 0$. Consider also that A' is not symmetrical, i.e., $xA'y \not\Longleftrightarrow yA'x$ since one of the two type usages has to have fewer methods. This also means that $y \in A(x) \implies x \notin A(y)$.

We can further refine the definition of almost similarity. Instead of looking for type usages which have the same method calls plus one additional one, it is also possible to consider those type usages which have the same method calls plus k additional ones. Then, in the definition of A the last line changes to: $|M(y)| = |M(x)| + k, k \geq 1$ This means that we would look for type usages missing k calls instead of just one.

Note that given a codebase with n type usages the sets $E(x)$ and $A(x)$ for one given type usage x can be computed in linear time $O(n)$ by iterating once through all type usages and checking for similarity or almost similarity.

3.1.3 The Strangeness Score

Recall the assumption behind the majority rule: A type usage is abnormal if a small number of type usages is exactly similar, but a significant number are almost similar. Informally, this means a few places use this type in exactly the same way, but a significant majority use it in a way which is slightly different (by exactly one method call). Assuming the majority is correct, the type usage under scrutiny is deviant and potentially erroneous, because it is missing a method call.

To capture concretely how anomalous an object is, we need a measure of strangeness. This will be the strangeness score. It allows to order type usages by how much of an outlier they are and thus to identify the "strangest" type usages, which are most interesting for evaluation by a human expert.

Formally, we define the S(trangeness)-Score as:

$$\text{S-score}(x) = 1 - \frac{|E(x)|}{|E(x)| + |A(x)|}$$

To see that this definition makes sense, consider the following extreme cases: Given one type usage a without any additional exactly similar or almost similar type usages, it will have $|E(a)| = 1$ and $|A(a)| = 0$. Then, we can calculate the strangeness score: S-score$(a) = 1 - \frac{1}{1} = 0$. So a unique type usage without any "neighbors" to consider is completely normal and not strange. On the other hand, given a type usage b with 99 almost similar and no other exactly similar type usages, we have $|E(b)| = 1$ and $|A(b)| = 99$. This results in a strangeness score of S-score$(b) = 1 - \frac{1}{1+99} = 0.99$. Intuitively such a type usage is very strange, and indeed, the S-score supports the intuition.

3.1.4 Which Calls are missing?

Only detecting the type usages which are outliers is not enough. We would also like to present the developer with some candidate suggestions, what might be the method call that is missing. The intuition for this is to look at all the almost similar type usages and the additional method call that each type usage invokes. We then take the frequency of the unique method calls among them and use this as the suggestion likelihood.

do not like the formulation, improve!

Formally, the calls we can recommend for a type usage x are the calls that are present in the type usages contained in $A(x)$ but are not in $M(x)$. The set of these methods shall be called $R(x)$ and is defined as follows:

$$R(x) = \bigcup_{z \in A(x)} M(z) \setminus M(x)$$

Now for each of these methods m in $R(x)$, we can calculate their likelihood as follows:

$$\phi(m, x) = \frac{|\{z \mid z \in A(x) \wedge m \in M(z)\}|}{|A(x)|}$$

This is identical to the share of type usages which use this method among all the type usages which are almost similar to x. Observe that this definition also works when $M(x)$ is empty or x is `this`.

$T(x) = \texttt{Button}$
$M(x) = \{\texttt{<init>}\}$
$A(x) = \{a, b, c, d, e\}$ 　　　　　　　 $R(x) = \{\texttt{setText}, \texttt{setFont}\}$
$M(a) = \{\texttt{<init>}, \texttt{setText}\}$
$M(b) = \{\texttt{<init>}, \texttt{setText}\}$ 　　　 $\phi(\texttt{setText}, x) = \dfrac{4}{5} = 0.8$
$M(c) = \{\texttt{<init>}, \texttt{setText}\}$
$M(d) = \{\texttt{<init>}, \texttt{setText}\}$ 　　　 $\phi(\texttt{setFont}, x) = \dfrac{1}{5} = 0.2$
$M(e) = \{\texttt{<init>}, \texttt{setFont}\}$

Figure 3.3: Example for computing the Likelihood of missing Calls [13]

again reference source of this example!!!

To illustrate these definitions, consider the example in Figure 3.3. The type usage under analysis is denoted with x, operates on a `Button` object and the only method that it invokes is the initialization. There are five almost similar usages denoted as a, b, c, d and e. Four of them (a through d) have a call to `setText` after the initialization.

The last one, e, has a call to `setFont` instead. Thus, we can recommend the methods `setText` and `setFont` with a likelihood of $4/5 = 80\%$ and $1/5 = 20\%$ respectively.

3.1.5 Ignoring the Context

> move this section down below extensions, but state explicitly, that they DID do this, only the rest is mine?

The reason for integrating the context into the definition of similarity and almost similarity is the idea that a type might be used in different ways depending on the situation it is used in. An example would be the class `FileInputStream` which is rarely opened and closed in the same method. Monperrus et al. [13] find in their evaluation that not considering the context adds a lot of noise (not relevant items in $E(x)$ and $A(x)$) and ultimately using it improves the precision of their experiments. However, one might question the general assumption behind using the context. The idea that types are used in different ways seems reasonable, but is the name of the function that they are used in a good separator between those use cases? Furthermore, using the context might work well for some types and some systems, but does it make sense everywhere? One problem with using the context is that it drastically reduces the number of type usages which are even considered for (almost) similarity. In smaller datasets, this makes it much more likely that "patterns" appear, which in truth are only artifacts of randomness.

All in all, it is not clear that using the context necessarily improves the performance of the system. Because of this, next to the standard variant $DMMC$ (Detecting Missing Method Calls), we also define a variant called $DMMC_{noContext}$, which does not take the context into account. It uses adapted measures for similarity and almost similarity, $E'_{noContext}$ and $A'_{noContext}$ which are defined in the same manner as E' and A', except that they do not require the context of the type usages to be identical. The definitions of E, A, S-score and ϕ stay the same, besides that they now make use of the newly defined relations.

3.2 Extensions

> more clearly state that this here is my work?

In the previous section, we summarized the work of Monperrus et al. [12][13], in this one we will explore some possible modifications to their method. For instance, it might be possible to leverage the majority rule not only for detecting missing method calls, but also to identify superfluous or flat-out wrong method invocations. Additionally, there are different ways of organizing the type usage data, which could yield better results.

In this section we will give the motivation and theoretical background for those modifications, in Chapter 5 we will explore further how they perform in comparison with the original.

3.2.1 Class-based merge

> basically the claass based merge is kinda the same as the inheritance relation right? (different split to dataset, either method/not methodlevel (context) or compelte class, or smth which inherits from X?)

The $DMMC_{noContext}$ variant seeks to answer the question if it makes sense to group the type usages based on the method they appear in. Following this line of thought, it is also possible to ask, if it is optimal to collect the type usages with method granularity. Even when ignoring the context, the type usages will still be extracted on a per-method basis, that is only the methods invoked in one method body will belong to one type usage. It is possible though, that grouping method calls with class granularity yields much better results. Often the methods invoked on one object within a specific class include many, if not all, of the methods invoked on it in its whole lifetime. Thus, a class-based type usage offers a much bigger window into the objects utilization and potentially this makes it easier to detect if some method is missing.

> question another assumption behind the std variant

We will denote this variant as $DMMC_{class}$. To obtain the class-based type usages, we extract the type usages at the method level, exactly as before. However, before calculating the strangeness score, we merge all those type usages which have the same type and originate from the same class. Note that this will merge them even if the original type usages came from different objects because tracing each objects life through the whole class would be too expensive. With these newly produced type usages we can then calculate the set of similar and almost similar type usages using $E'_{noContext}$ and $A'_{noContext}$, respectively.

To formalize $DMMC_{class}$, we need the new operator $\mathrm{Cls}(x)$ which denotes the class from which a type usage was extracted. We can then define the set of type usages which have the same type and class as x as:

$$TC(x) = \{x_i \mid T(x) = T(x_i) \wedge Cls(x) = Cls(x_i)\}$$

Then we can calculate the set of new type usages as follows:

$$\{x' \mid T(x') = T(x_0),$$
$$C(x') = \bigcup C(x_i),$$
$$M(x') = \bigcup M(x_i),$$
$$with \quad x_i \in TC(x_0)\}$$

idea: introduce a separation of ALL tus based on the type and class, then merge the ones in each set (-> need set of all tus, can iterate through all separations for x0 etc)

not super clean, what is x_0 :/, smth like rotate through all the tus until all of them have been used?

Recall that for exact and almost similarity we will use the version which ignores the context, so $C(x')$ is not that relevant.

mention some effects (such as much smaller dataset, blabla?) or somewhere else?

3.2.2 Investigating different Anomalies

While a *missing* method call might be the first error type that comes to mind and also potentially the most frequent one, the general technique of the majority rule can also be easily adapted to look for two different anomalies related to method calls. First off, it is possible to flip around the notion of almost similarity and investigate if there exists a superfluous method call somewhere. Additionally, one can envision a developer to invoke the *wrong* method, rather than forgetting a call or adding one too many.

some more? (example), just words...

Superfluous Method

The idea behind $DMMC_{superfluous}$ is the inverse of the method proposed by Monperrus et al.. Instead of checking if a lot of other type usages are calling an additional method and the current one is thus missing a call, here we check if a lot of other type usages are making *fewer* calls, and the current one is doing something extra which it should not. It is not intuitively clear what kind of errors this procedure will discover or indeed if there will be any. Furthermore, it is possible that it will only find outliers which are intentional outliers, be it to handle a special case, to work around a bug or for any other reason. Nonetheless, it seems reasonable to investigate this idea and only dismiss it if it does not yield any interesting results.

maybe best rationalization: it is there, before trying it we don't actually know if its bad, could be that this is actually a type of error which has some impact

include some example? (either for use case or for formalization?)

Adapting the normal system to detect this type of outlier is rather simple. We only need to adapt the definition of $A(x)$ to:

$$A_{superfluous}(x) = \{x \mid yA'x\}$$

Recall that the almost similar relation A' is not symmetric and observe that in comparison to $A(x)$, here x and y are flipped in the application of A'. Thus, this definition of almost similarity considers all those type usages which call one method less than the input type usage to be almost similar. With the definition of S-score and ϕ as above, the scores will now indicate which type usage is calling anomalously many methods and which method might be the one that is too much.

Wrong Method

Further extending the idea of the previous section and thinking consequently, developers might *miss* a method call, they might *add* a useless one but, of course, they could also simply choose the *wrong* one. Especially if the developer is not all too familiar with a given framework or there is some ambiguity in the method names or documentation, it is entirely possible that he chooses the wrong method for a given task. Of course, it is possible that this type of error results in noticeable bugs much faster than a missing call, but again we do not know this before investigating.

example? maybe some sort of conversion bug?

To formalize this new variant $DMMC_{wrong}$ looking for invocations of a wrong method, we again need only change the definition of almost similarity. This time we define a new relation A'_{wrong} as follows:

$$
\begin{aligned}
xA'_{wrong}y \iff & T(x) = T(y) \ \wedge \\
& C(x) = C(y) \ \wedge \\
& M(x) \neq M(y) \ \wedge \\
& |M(x)| = |M(y)| \ \wedge \\
& \exists S.|S| = |M(x)| - 1 \wedge S \subset M(x) \wedge S \subset M(y)
\end{aligned}
$$

For this relation to hold, again the type and the context of both type usages have to be the same. Furthermore, it requires that all but one method of $M(x)$ and $M(y)$ are identical, that is, that the sets of method calls they make differ only by exactly one method.

can I define this a bit better or not?

anything more here?, better description?

4 Implementation

After laying out the theoretical foundations of detecting missing method calls using the majority rule in the last chapter, this chapter focusses on the practical side and on the internal workings of the system we developed. It explains the steps from receiving the input program to outputting a list of anomalies which hint at potentially missing method calls. We give details about design decisions taken and the reasoning behind them, pitfalls that had to be overcome and the trade-offs that had to be accepted. Additionally, we explore some of the mistakes made and dead ends that we encountered.

4.1 Overview

Our system is (primarily) a system for detecting missing method calls. It works by statically analyzing a given software application, extracting the type usages which it contains and finally determining if any of them are anomalous by the Majority rule as described in Section 3.1. The end result should be a list of locations which are potentially missing a method call. Given a software system, which shall be tested for anomalies, conceptually the following steps have to be realized:

1. Extract all type usages from the software

2. For each type usage x among them:
   a) Search for type usages which are exactly similar to x (i.e. calculate $E(x)$)
   b) Search for type usages which are almost similar to x (that is determine $A(x)$)
   c) Calculate the strangeness score of x
   d) Extract the list of potentially missing calls and their likelihood ϕ

3. Output a list of anomalous type usages, sorted by their S-score, together with the calls they are potentially missing
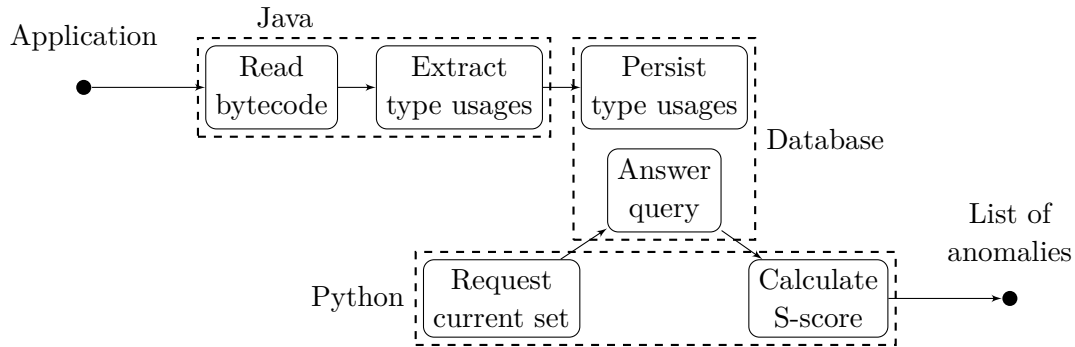
Figure 4.1: System overview

However, this simple outline does not represent the actual realities of the system. Instead of a singular process with [flat / not branching / ...] flow, it is split into three different parts which are laid out in Figure 4.1. First, a java application reads the byte code of the program under analysis and iterates over all method definitions to extract the type usages which are present. The extracted type usages are then persisted in a database to ensure flexibility and high performance in the following analysis phase. For the actual anomaly detection we use a relatively simple python program to enable fast prototyping and again flexibility. It goes through all sets of type usages (more on those sets in Section 4.6), requests the type usages in the set and calculates their strangeness score. After iterating through all the sets of the application it finally outputs the results.

partitions instead of sets?

The following sections give a more in depth explanation of the separate steps.

smth more?

4.2 Bytecode Analysis

maybe rename section?

We use Soot [1] for the bytecode analysis. Soot was originally a Java optimization framework but has since evolved to support a wide range of usages. One can use it to analyse, instrument, optimize and visualize Java and Android applications. For this purpose it provides call graph construction, points-to analysis, def or use chains, inter and intra-procedural data-flow analysis and taint analysis. We do not need most of it's functionality, but simply use it to statically extract type usages from the input application.

It would be possible to extract type usages from the source code, but extracting them from compiled code has some advantages. The JVM byte code is very standardized and

---

[1] `https://sable.github.io/soot/`

Option	Parameter	Explanation
`-app`	-	Run in application mode
`-keep-line-number`	-	Keep line number tables
`-output-format`	none	Set output format for Soot
`-allow-phantom-refs`	-	Allow unresolved classes
`-src-prec`	apk-class-jimple	Sets source precedence to format files
`-process-multiple-dex`	-	Process all DEX files found in APK
`-android-jars`	[path]	The path for finding the android.jar file

Figure 4.2: The parameters passed to Soot

so is the Dalvik byte code found in Android applications, which facilitates the analysis. While Soot supports source code analysis in theory, it only works up to Java 7 and byte code analysis is the recommended approach. Besides that, using compiled applications simplifies our experimental setup and as a marginally useful side-effect, it also helps with analyzing obfuscated applications. Finally, there are no real disadvantages to this approach, if the program is only available as source code we can simply compile it before analyzing it.

Soot operates by transforming the given program into intermediate representations which can then be optimized or analysed. It provides four intermediate representations with different use cases, we use Jimple, Soot's primary representation. Jimple is a typed 3-address representation, which is especially suited for optimization, but also suffices for our purpose.

Soot is a very powerful framework and one can customize it to carter quite specialized requirements. We will mostly analyze Android applications and the settings we are using reflect this. In Figure 4.2 we have listed the most important settings we are using to configure Soot.

> more detailed description of settings? i think not

4.3 Extracting Type Usages

> some more on extraction! + the algorithm! some information on what kind of further analysis is attempted (local must alias, bla) general / not sure where yet: explain the changes I made to their code ie proper refactoring, etc

The Soot execution is divided into a number of 'phases' each of which has a specific task like for example the aggregation of local variables. The actual work is done by so called transformations, which are able to modify the input they receive but are not

required to do so. The phases are further grouped into 'packs' and the registered packs are applied successively during the execution. We register a custom piece of code in the Jimple transformation pack, which is applied to every method in the analysed program.

This custom transformation receives as input a Jimple representation of the method body it is currently analyzing. To extract the type usages, it iterates through all the statements in the method and marks those which are invoking a method. In the end it has extracted a list of all method calls that are happening in this method. It now iterates through this list of method calls and groups them into type usages. To do so, it first checks on which variable or object the current call was invoked and checks if there is already a type usage for this object. If there is, it adds the call to the type usage and advances to the next method call in the list. If there is no type usage for the object so far, it creates a new one. After completing this process for all method invocations in the method body the analysis is complete and it returns the list of type usages it extracted.

One important step of this analysis is the 'LocalMustAliasAnalysis', which attempts to determine if two local variables (at potentially different program points) must point to the same object. This is necessary to decide if the calls made on these locals should be grouped into one type usage. The underlying abstraction is based on global variable numbering and follows the ideas presented by Lapowsky et al. [9] with some minor adaptions. This analysis is a Soot feature and in test runs on big applications we noticed that it can be very expensive. Because of this we made it optional, however for our benchmarks and evaluation we were able to use it.

One additional detail is that we are excluding some classes from the Android analysis for performance reasons. The packages we exclude in this manner are `java.*`, `android.*`, `soot.*` and `javax.*`. These are all framework classes and we are primarily interested in the type usages occurring in the application itself not in the framework. This means that we do not store type usages that occur *inside* of those classes, we do, however, still record usages of them.

> explain a bit better that these classes are kind of exactly what we are most interested in as far as type usages go, but we don't want too step ITNO them + make sure this works good together with coresponding paragraph in the next section

> some changes that had to be made to the analysis framework for android analysis?

4.4 Storing Type Usages

After the transformer has analyzed the current method body and extracted a list of type usages from it, we persist them into a database. For this, we are using HSQLDB[2], a SQL relational database written in Java, which provides a multithreaded and transactional

---

[2]`http://hsqldb.org/`

database engine with in memory or disk based tables. We chose HSQLDB because it is small and has good performance as well as an easy setup. Additional arguments where its permissive license and the large number of features it supports.

In their work, Monperrus et al. save all data in a text based format, but this means they need to parse everything again for the analysis, which can be slow and require a lot of memory for large inputs. Additionally, the data takes up a lot of storage space when persisted to disk. In contrast, using a database has many positive ramifications. First off, retrieving data from a database is fast and we can access the data selectively based on changing criteria. Additionally, we gain a lot of flexibility and it becomes easier to extend both, the stored data and the analysis itself. (This enables us to quickly build several prototypes of the analysis using Python.))

its already useful to analyze one application in separation but we really care about the framework / library classes which are used by more than one application analyze a lot of applications using the same framework/library on these classes we can build a large dataset of type usages -> have much more information to determine if a tu is an anomaly or not -> having a database really helps with that (can hold a lot of data + can selectively query it are the main points i think?) of course it's useful to ALSO analyse tus over classes which are private to my application (if there is enough data) but the framework thing is just a) easier to get the info and b) kind the point of this thingy?

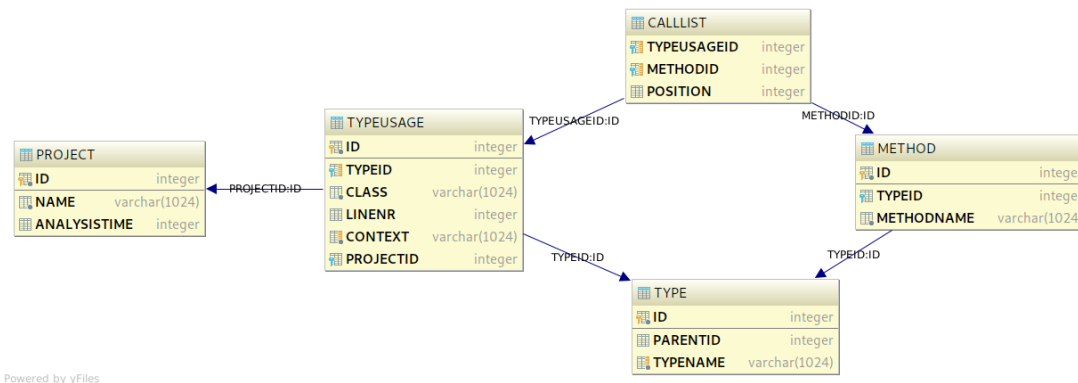Figure 4.3: Overview of Database Tables

REMOVE the xfiles watermark

+ describe obviously (+ what was made how why) Figure ref depicts the database layout (how in depth? explain every field? I'm not so sure... maybe for now a rough overview) the central piece is of course the type usage table, holds information related to one particular type usage the class it was recorded in (fully qualified pck name), if

possible the line of code in which the object was first encountered and its context, that is the name of the method in which it was found + the method parameters. additionally it holds references to the "project" it belongs to, which in our setup is the Android application it was taken from, and the type it belongs to, that is what type the object of the tu is an instance of (rephrase...)

Each type holds a reference to it's parent (if one exists) to enable rebuilding the inheritance hierarchy if so desired additionally the methods that can be invoked on a type are stored in the "method" table (basically just their (full) name).

Finally the calllist table holds the information which of the method calls that COULD be invoked on a type are actually called on the typeusage in question

+ some of the considerations made (x. normalenform, sowas?) -> ask torben? -> is there a problem with class and context (potentially duplicate information!, should actually belong to a class table i guess?)

4.5 Improvements and Dead Ends

should this be a section or a subsection? if this is before anomaly detection we can kind of explain why we are using the python + it's kinda the process from one to the other? db -> too slow -> . . .

explaining all the work i did + what was already there why some solutions where discarded (eg pure database / could be revisited if it turns out to be the best anyways - performance)

Monperrus et al. released their code and data + link to understand it better we refactored it extensively and implementd the database backend connection (mention or rather not?) during the refactoring process discovered small discrepancies between what's described in the paper and the code -> namely calculation of almost similarity does a double intersection i think? (recheck this!) which means that cases where the method calls are as follows + example where it gives the wrong results was there some other bug? - i think not + however we fixed this and reworked their analysis, the results didn't change in a significant way

idea: use database to calculate strangeness scores, cause the data is already there hindsight: not the best idea std query: view with merge over many tables -> quite slow even on small test dataset lots of attempted improvements to the query with better indices, smart cached subtables etc but all in all futile, it remained slow or memory intensive or both (on a application of medium size -> mention LOC of teamcale, resulting in one dataset with about X typeusages) also tried postgresql which is supposed to be faster for some applications, but results disappointing evaluation of other db system + why it didn't work (basically pure db stupidity) We also invested some time into exploring

the possibility of using PostgreSQL[3], but as explained in Section 4.5 it turned out that its performance was inferior. (change, this is the deadends section xD)

not sure if this here or in next section?

biggest problem: nxn comparison of so many TUS (-> N2 laufzeit) it seems db not able to unrevel this with my unsuited instruction oä however thinking about what happens: for each tu go through all tus, first check if type and context are the same (/only type for dmmc noContext) only if that's the case go and compare the methods this gives rise to natural subsections / partitions of the dataset, namely those sets of tus with the same type(+context) only those have to be compared to each other partition of dataset -> makes sense anyways and db is perfectly prepared for that with that realization go on to next step of the analysis: actually detecting anomalies

4.6 Anomaly Detection

-> look at python code and thinks about what needs to be said + suitable structure

need another picture? or fine without? i'm not sure i need that much detail... maybe another img or flow chart for python ablauf (or too much detail?)o

somewhere explain the "sets of tus" + peformance issues before (ref db dead end) depending on the variant we partition by types or by combinations of types+context outer loop: iterate through all the partitions in the dataset + belonging to the current partition lvl -> observe: using the database it is very simple to obtain the possible partitons (+query?) and additionally it is super easy to get the corresponding TUs (+query) -> here the index on [column] really helps -> this is exactly what a db is made for, it is fast and seamless then iterating through all the tus in the current partition comparing the methods being invoked (sets -> does not count number of invocations) comparison each with each only once -> upper triangle matrix, saves nearly half again -> for each tu calculate the score + save the list of results for further analysis / evaluation by expert / hand

this only part affected by the different variants presented in chapter 3 already talked about context vs no context the class merge variant is performed by simply inserting some preprocessing after loading the data with type partition -> merge those tus which belong to the same class -> done retrieving relevant type usages for the current variant + potentially modifying them + throwing them into the detector

-> detection of superfluous or wrong method can use exactly the same setup, only need to change the implementation of the almost equal check mention in the end that i tried a Better anomaly detection (is the anomality rule actually a GOOD measure for this kind of anomalies, or should we use something totally different?)) clustering detector try (+

---

[3]`https://www.postgresql.org/`

hypersphere idea?) but all in all it didnt work out / not enough time + vermutungen warum es nicht funktioniert -> dataset exploration + size of method and typeusage lists per partition!

4.7 Other Ideas / Problems

not sure i need this section + what should be included

static functions evaluation! something to fix the dotchaining problems -> for both basically I'm just saying that i looked into it and it was too much work / impossible?

5 Evaluation

roter faden: results of manual analysis vs benchmark (ie. is context better than no context) take their method + analyze it on a qualitative + quantitative basis on a big android dataset

The XXX section revolves around the benchmark system, the basic idea behind it and how it is designed to ensure maximum flexibility.

does this order make sense or rather split by research question?

5.1 Methodology

5.1.1 Qualitative Evaluation / Android Case Study

android apps, blabla manual review of top 50 findings for each "method" -> which are the different methods?

downloading + automatically analyzing android apps (in evaluation section?)

mention that monperrus hold 0.9 as treshhold for smth being an anomaly and that this will be our cutoff oä

5.1.2 Automated Benchmark

building benchmarking infrastructure flexibility with python class setup degrading type usages + checking if they will be detected some more subsections? (different ways for degrading type usages?)

sometimes in even further degraded settings (degrade more than one)

general idea behind evaluation method and what we are trying to detect actually it's a sort of simulation (see recommender system book p. 301f) - micro vs macro evaluation, . . . imitation of the real system of software development using a much simpler system, namely dropping method calls obvious question: how similar are the such created mmcs to mmcs in the wild?

Metrics explanation - Precision, Recall additional metrics from recommender systems book p.245f (robustness, learning rate - p. 261) performance -> training time + performance when working

5.2 Results

Which questions are we trying to answer with the evaluation? quick overview + method to do this

qualitative evaluation: are the findings useful in practice? quantitative evaluation: given some assumptions, how useful can we expect this technique to be? robustness? Qualitative vs automatic same results? make sure I will be able to a) answer those questions and b) the answers are "interesting"

große frage: bringt das Verfahren etwas?

for each question: explanation of question, then results + interprettation

5.2.1 Dataset Overview

rausziehen (extra section)

what kind of dataset do we have? (how many apps, how big, how many TUs, how are the TUs split between the apps -> size of applications) number of partitions (types+context / types) and their sizes! what are typeusages in Android apps like? (length of method list, percentage on Android framework, percentage on other stuff, etc)

general data about avg tus related to different settings etc -> what does our dataset even consist of?

look into and understand the type of data i have -> how often small inputs, how often big, ... -> paint a bunch of graphs of the pkl results! (histogram of tu list sizes, etc) (correctly bin histogram!) tu method list sizes - also check their paper, q mas?

indea behind using android: do real software systems actually behave in this way ie: is it "necessary" to use classes in the same way probably: most likely present in GUI systems common way of using some classes -> investigate via Android (+ additional advantages: rich open source community, java, blabla)

mention somewhere: the monperrus2013 p11 highlight? (at least equivalent for my dataset)

5.2.2 RQ1: Does the Strangeness Score behave as expected on Android Apps?

-> ist keine forschungsfrage, gehört zur vorherigen section! (the behavior that I'm expecting is basically a mathematical necessity!)

do the general assumptions hold? (most tus have a low score, most apps have few findings, etc) -> answer with graphs related to general score verteilung -> most tus are "normal", a few are "abnormal" what do those assumptions mean? -> they expect some kind of uniformity to the type usages, probably mostly present in GUI etc frameworks verteilung of strangeness scores, histogram of it

Frage: Ist das ganze überhaupt schnell genug? -> JA! (auslesen kosten +)

5.2.3 RQ2: How "useful" is this technique?

wie relevant sind diese obtained results in der praxis (as determined by fp vs tp)

change title. . .

qualitative evaluation findings: how many true findings in relation to false positives does it find + average number of findings per app (remember that the findings now are for all 626 apps)

pick a couple of high scoring findings and explain the failure modes (eg doesn't take branching into account, etc) + of course the REAL bug + some real smells

and of course regarding the "results" always qualify -> this only for android, open source apps, manual evaluation, blabla, so we never actually know for sure

think if i should totally change the manual eval: take random sample of Apps and there look at all the anomalous (>.9) findings (kind sounds a bit smarter?)

RQ2.1: Is there a noticeable difference between the different variants?

each for the slightly adapted forms (no context, class merged, etc)

RQ2.2: Given a known missing method call bug, can this approach detect it?

only if time!

reverse test: find actual missing call in bug database -> can I find it using the approach and is it among the top findings?

5.2.4 RQ3: How meaningful are the benchmarking results?

"meaningful" not a good word, rather already describe what I'm doing (ie do the bnechmark results align with the results of the manual evaluation oä)

does it make sense to evaluate this method using the benchmarking as proposed by Monperrus et al.? -> comparison between manual evaluation of findings vs automatic benchmark over the different techniques used (context, no context, class merge, . . .)

first present benchmark results of different techniques (context, class merge, . . .) then relate to qualitative results

are the results better if we leave out the context / merge on a per class basis / FOR DIFFERENT Ks!. . .

5.2.5 RQ4: How robust is this technique in the face of erroneous input data?

I would add this, even if the results of comparison with manual evaluation are relatively negative / don't say anything worauf is robustness bezogen? -> results quality or smth else?

Robustness explanation wie viel brauche ich um sinnvoll viel zu lernen - hälfte der leute machen fehler -> wie robust brauchst du die daten / ist das verfahren ! what is needed for it to work (lines of code, feature richness, etc) problem: how to evaluate when it is "working" vs not working -> again simulate "missing methods", compare performance?

additional test: for true findings, try to throw away parts of the data -> when can we not find this anymore -> mathematical answer! / instead of throwing away: introducing random errors in the related tus

5.2.6 RQ5: What are the requirements for the input size?

benny didn't like "requirements"... ie: how big of a codebase do we need to be able to "learn". When can we apply this to a project which does not rely on some well known open source framework (Android) where it is easy to gather additional data, but on for example an in-house-closed-source library oä.

extremely hard to answer especially if the findings regarding benchmark validity are more or less negative. however, even if I cannot give experimental answers, I can at least give some lower bounds + thoughts (need at least 10 tus within the same "category" to even reach a score > 0.9, from practical results in qualitative analysis, that is probably not enough, blabla)

5.3 Discussion

5.4 Threats to Validity

explain problems with the automatic evaluation how related are the degraded TUs to missing method calls you can find in the wild? improving the metrics by dropping cases where we know we won't find an answer

also mention runtime! wie sehr sind die resultate aus der Android case study 1. Wahr (subjektive bewertet etc) 2. Übertragbar auf andere Anwendungsfälle(sind open-source programme of vergleichbar mit professionellen, Android eco System mit anderen, etc) ...

6 Conclusion

general takeaway about suitability of this method + reasons WHY I would say so

new method is amazing and way better in case xy but worse in zz robustness has shown this and that future research would be interesting in the direction of . . .

6.1 Contribution

-> even have this section?

Datenset generiert, zumindest was über software sturktur lernen, bla !

6.2 Future Research

apply this anomaly detection method to other features (implements, overrides, pairs of methods, ...) Reihenfolge von Methoden - hat gut funktioniert und wäre potentiell nützlich, aber kleines subset + lots of research already done order of method calls? (probably should wait for empirical results - how long are typical method lists (state explosion, ...) / latice solution?) efficient update (only update files touched by change + the scores for affected type usages) - maybe database view for scores will already do this automatically? higher precision by some means? (clustering, etc) -> better anomaly detection algorithm! (but difficult / constrained by the input size (few methods,. . .) -> refer to dataset analysis section) Performnance (shouldn't be a problem?) . or potentially later look into other features (like implements/override), pairs of methods, etc

Bibliography

[1] N. Ayewah, W. Pugh, J. D. Morgenthaler, J. Penix, and Y. Zhou. "Evaluating static analysis defect warnings on production software." In: *Proceedings of the 7th ACM SIGPLAN-SIGSOFT workshop on Program analysis for software tools and engineering.* ACM. 2007, pp. 1–8.

[2] C. W. Barrett, R. Sebastiani, S. A. Seshia, C. Tinelli, et al. "Satisfiability modulo theories." In: *Handbook of satisfiability* 185 (2009), pp. 825–885.

[3] S. M. Blackburn, R. Garner, C. Hoffmann, A. M. Khang, K. S. McKinley, R. Bentzur, A. Diwan, D. Feinberg, D. Frampton, S. Z. Guyer, et al. "The DaCapo benchmarks: Java benchmarking development and analysis." In: *ACM Sigplan Notices.* Vol. 41. 10. ACM. 2006, pp. 169–190.

[4] D. Engler, D. Y. Chen, S. Hallem, A. Chou, and B. Chelf. "Bugs as deviant behavior: A general approach to inferring errors in systems code." In: *ACM SIGOPS Operating Systems Review.* Vol. 35. 5. ACM. 2001, pp. 57–72.

[5] F. A. Fontana, M. V. Mäntylä, M. Zanoni, and A. Marino. "Comparing and experimenting machine learning techniques for code smell detection." In: *Empirical Software Engineering* 21.3 (2016), pp. 1143–1191.

[6] F. A. Fontana, M. Zanoni, A. Marino, and M. V. Mantyla. "Code smell detection: Towards a machine learning-based approach." In: *Software Maintenance (ICSM), 2013 29th IEEE International Conference on.* IEEE. 2013, pp. 396–399.

[7] J. Han and M. Kamber. *Data mining: concepts and techniques.* Second. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2006.

[8] G. Hecht, R. Rouvoy, N. Moha, and L. Duchien. "Detecting antipatterns in android apps." In: *Mobile Software Engineering and Systems (MOBILESoft), 2015 2nd ACM International Conference on.* IEEE. 2015, pp. 148–149.

[9] C. Lapkowski and L. J. Hendren. "Extended SSA mumbering: Introducing SSA properties to languages with multi-level pointers." In: *International Conference on Compiler Construction.* Springer. 1998, pp. 128–143.

[10] C. Marinescu, R. Marinescu, P. F. Mihancea, D. Ratiu, and R. Wettel. "iPlasma: An Integrated Platform for Quality Assessment of Object-Oriented Design." In: *ICSM.* 2005.

[11] N. Moha, Y.-G. Gueheneuc, L. Duchien, and A.-F. Le Meur. "Decor: A method for the specification and detection of code and design smells." In: *IEEE Transactions on Software Engineering* 36.1 (2010), pp. 20–36.

[12] M. Monperrus, M. Bruch, and M. Mezini. "Detecting missing method calls in object-oriented software." In: *European Conference on Object-Oriented Programming*. Springer. 2010, pp. 2–25.

[13] M. Monperrus and M. Mezini. "Detecting missing method calls as violations of the majority rule." In: *ACM Transactions on Software Engineering and Methodology (TOSEM)* 22.1 (2013), p. 7.

[14] T. T. Nguyen, H. A. Nguyen, N. H. Pham, J. M. Al-Kofahi, and T. N. Nguyen. "Graph-based mining of multiple object usage patterns." In: *Proceedings of the the 7th joint meeting of the European software engineering conference and the ACM SIGSOFT symposium on The foundations of software engineering*. ACM. 2009, pp. 383–392.

[15] F. Palomba, D. Di Nucci, A. Panichella, A. Zaidman, and A. De Lucia. "Lightweight detection of Android-specific code smells: The aDoctor project." In: *Software Analysis, Evolution and Reengineering (SANER), 2017 IEEE 24th International Conference on*. IEEE. 2017, pp. 487–491.

[16] M. Pradel and T. R. Gross. "Detecting anomalies in the order of equally-typed method arguments." In: *Proceedings of the 2011 International Symposium on Software Testing and Analysis*. ACM. 2011, pp. 232–242.

[17] J. Reimann, M. Brylski, and U. Aßmann. "A tool-supported quality smell catalogue for android developers." In: *Proc. of the conference Modellierung 2014 in the Workshop Modellbasierte und modellgetriebene Softwaremodernisierung–MMSM*. Vol. 2014. 2014.

[18] M. P. Robillard, E. Bodden, D. Kawrykow, M. Mezini, and T. Ratchford. "Automated API property inference techniques." In: *IEEE Transactions on Software Engineering* 39.5 (2013), pp. 613–637.

[19] H. Shen, J. Fang, and J. Zhao. "Efindbugs: Effective error ranking for findbugs." In: *Software Testing, Verification and Validation (ICST), 2011 IEEE Fourth International Conference on*. IEEE. 2011, pp. 299–308.

[20] D. Verloop. "Code smells in the mobile applications domain." PhD thesis. 2013.

[21] A. Wasylkowski, A. Zeller, and C. Lindig. "Detecting object usage anomalies." In: *Proceedings of the the 6th joint meeting of the European software engineering conference and the ACM SIGSOFT symposium on The foundations of software engineering*. ACM. 2007, pp. 35–44.

[22] H. Zhong, T. Xie, L. Zhang, J. Pei, and H. Mei. "MAPO: Mining and recommending API usage patterns." In: *European Conference on Object-Oriented Programming.* Springer. 2009, pp. 318–343.

[23] Y. Zhou, R. Gu, T. Chen, Z. Huang, S. Panichella, and H. Gall. "Analyzing APIs documentation and code to detect directive defects." In: *Proceedings of the 39th International Conference on Software Engineering.* IEEE Press. 2017, pp. 27–37.