

More Accurate Recommendations for Method-Level Changes

Georg Dotzler, Marius Kamp, Patrick Kreutzer, Michael Philippsen

Programming Systems Group, Friedrich-Alexander University Erlangen-Nürnberg (FAU), Germany

{georg.dotzler,marius.kamp,patrick.kreutzer,michael.philippsen}@fau.de

ABSTRACT

During the life span of large software projects, developers often apply the same code changes to different code locations in slight variations. Since the application of these changes to all locations is time-consuming and error-prone, tools exist that learn change patterns from input examples, search for possible pattern applications, and generate corresponding recommendations. In many cases, the generated recommendations are syntactically or semantically wrong due to code movements in the input examples. Thus, they are of low accuracy and developers cannot directly copy them into their projects without adjustments.

We present the Accurate REcommendation System (ARES) that achieves a higher accuracy than other tools because its algorithms take care of code movements when creating patterns and recommendations. On average, the recommendations by ARES have an accuracy of 96% with respect to code changes that developers have manually performed in commits of source code archives. At the same time ARES achieves precision and recall values that are on par with other tools.

CCS CONCEPTS

• Information systems → Recommender systems; • Software and its engineering → Software maintenance tools;

KEYWORDS

Program transformation, refactoring, recommendation system

ACM Reference Format:

Georg Dotzler, Marius Kamp, Patrick Kreutzer, Michael Philippsen. 2017. More Accurate Recommendations for Method-Level Changes. In *Proceedings of 2017 11th Joint Meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering, Paderborn, Germany, September 4-8, 2017 (ESEC/FSE'17)*, 11 pages. <https://doi.org/10.1145/3106237.3106276>

1 INTRODUCTION

Developers often perform the error-prone and repetitive task of applying the same *systematic edits* to many locations in their code base. The reasons for such systematic changes vary. They include bug fixes, the adaption of call sites to new APIs, etc., and constitute the majority of structural edits in software projects [22, 31].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ESEC/FSE'17, September 4-8, 2017, Paderborn, Germany

© 2017 Copyright held by the owner/author(s). Publication rights licensed to Association for Computing Machinery.

ACM ISBN 978-1-4503-5105-8/17/09...\$15.00

<https://doi.org/10.1145/3106237.3106276>

Developers usually perform systematic edits manually because often the code is slightly different in each location (with respect to names, contexts, etc.) so that they cannot use a plain textual search-and-replace. Refactoring wizards of IDEs also just offer a set of predefined transformations and thus only provide help for a subset of all systematic changes. Furthermore, they do not support code changes that lead to different semantics, which is necessary in many situations (e.g., to fix errors).

Current research tools [29, 35] learn systematic edits from one or more manually provided training examples and build generalized patterns from them. Then the tools search the code base for locations to apply these patterns to and present the code of the applied patterns as recommendations to developers. Since the tools construct the suggested code (to a large extent) purely syntactically, the code is often wrong (uses undefined variables, calls non-existing functions, misses statements, etc.). Thus, the recommended code is often inaccurate, i.e., it is not what developers would have written if they had done the systematic edits by hand. Although such inaccurate recommendations are helpful as they identify locations for code changes, developers still need to manually adjust them before they can insert them into their projects.

Fig. 1 illustrates this problem. The two recommendation systems LASE [29] and ARES learn a pattern from the code changes given in Fig. 1(a+b) (in a classic *diff* representation). Both tools find the matching code location in Fig. 1(c). LASE suggests the code change in Fig. 1(d). ARES produces the more accurate one in Fig. 1(e) that is closer to what a developer would have written. There are two reasons for the differences. First, since change (b) explicitly removes

```
init();
- foo.someMethod(42);
- print(foo);
+ if (foo != null) {
+   foo.someMethod(42);
+   print(foo);
+ }

init();
- assert (foo != null);
- foo.someMethod(42);
- foo.print();
+ if (foo != null) {
+   foo.someMethod(42);
+   foo.print();
+ }
```

(a) Example change.

(b) Example change.

```
init();
assert (foo != null);
foo.someMethod(42);
foo.run();
```

(c) Matching code location.

```
init();
assert(foo != null);
if (foo != null) {
  foo.someMethod(42);
}
foo.run();

init();
if (foo != null) {
  foo.someMethod(42);
  foo.run();
}
```

(d) Inaccurate recommendation by LASE.

(e) More accurate recommendation by ARES.

Figure 1: Example for code recommendations.

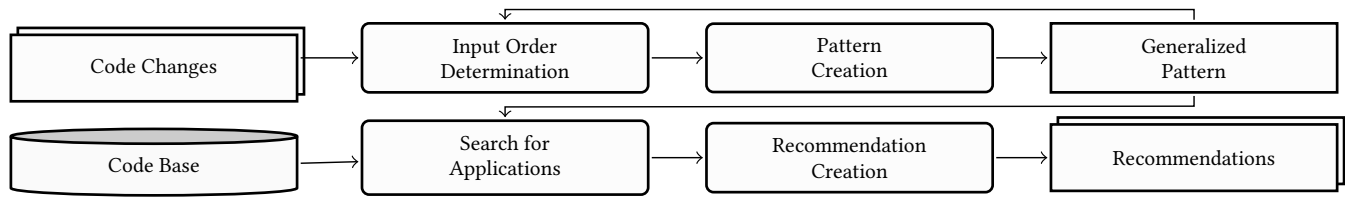


Figure 2: Workflow of ARES for one input set of code changes.

the *assert* statement, it should not be part of the recommendation. However, as LASE only applies the common subset of the code transformations that are present in the training examples, the *assert* remains untouched. Second, LASE leaves *foo.run()* in place. This is wrong because, from a developer’s point of view, *someMethod* and the code in the following line are both *moved* into the new *if*-statement. However, many approaches (like LASE) either do not express code movements accurately or cannot handle them due to the type of patterns and algorithms they use [35, 36]. Since the latter only support *delete*, *insert*, and *update* operations on the code, they can only learn to insert *print(foo)* and/or *foo.print()*, but not the more accurate code movement.

To avoid these two sources of inaccuracy, the pattern representation of ARES can both express variations in the input code changes and code movements. ARES also uses algorithms that can generate more accurate recommendations based on these patterns.

Fig. 2 shows the workflow of ARES. The loop in the upper row derives and refines a generalized pattern that represents all the code changes in a training set. The loop starts with two code changes from the input set and generates a pattern for them. Subsequent iterations refine this pattern by considering the next examples successively. With a *Generalized Pattern* ARES browses a given code base for locations where the pattern is applicable (*Search for Applications*). It then applies the transformation encoded in the pattern to a copy of each found location (*Recommendation Creation*) and presents the transformed copy as recommendation to the user.

Sec. 2 explains the pattern design and how this helps in creating more accurate recommendations. Secs. 3-6 describe the pattern creation, the search for locations to apply them, and the generation of the recommendations. We then quantitatively compare ARES with LASE in Sec. 7, and discuss related work before we conclude.

2 PATTERN DESIGN

ARES uses a pattern representation that is close to source code. Fig. 3 holds an example pattern (based on our previous work on this topic [15]) that expresses (and generalizes) code changes applied to a code base. The set of code changes is the input of ARES. There are two plain Java code blocks in the pattern, one original code block that represents the input examples before their transformation and one modified code block that represents the examples after their transformation. To express variations in the input examples, the patterns use a set of annotations, added as Java comments.

The example pattern starts with the *match* annotation that simply declares the beginning of the pattern. The tags *original* and *modified* allow an easy distinction between the two code parts of the pattern for a human reader. The original part of the *match* annotation also contains the letter *k*. This *k* stands for an identifier that occurs in the pattern body and means that the identifier name of *k* is not fixed and

any variable name in a code location is an acceptable replacement for *k*. In contrast, any location to which the pattern is applicable has to use the identifier name *foo* (line 12). The list of identifiers in the *match* annotation is one mechanism of ARES to express a generalization. When ARES creates a recommendation it replaces *k* with the actual variable name at the respective code location. This increases the accuracy of the generated recommendations.

Another generalization mechanism is the *wildcard* annotation. It matches arbitrary code during the search for suitable code locations. There are two different versions. First, wildcards tagged with *stmt* accept none or arbitrary statements at the code location (see lines 8 and 11). This design provides a solution to the problem of the deleted *assert* in the introductory example (Fig. 1) as it accepts the *assert* if it is present and has no effect otherwise. Thus, *stmt*-wildcards handle variations in the training examples and increase the accuracy of the generated recommendations.

Second, wildcards tagged with *expr* always refer to the following statement. They specify which part of it can contain an arbitrary expression. It is possible to have several such wildcards referring to the same statement. For example, the *expr*-wildcard in line 2 specifies that at the first occurrence of *verbose* in line 3 the search algorithm of ARES can allow arbitrary expressions. This means that at a possible code location for the pattern, the call to *init* may have none or an arbitrary number of arguments. This handles variations in the training examples on an even finer level.

The *modified* part of the pattern does not contain *wildcard* but *use* annotations. During the creation of a recommendation ARES replaces the *use* annotation with the code that was matched by the corresponding wildcard. A *wildcard* and a *use* correspond to each other if they have the same name (e.g., *A1*). As the name can appear anywhere in the modified part, a pattern can express movements

```

1 11match (original, (k)) {           11match (modified) {
2 12wildcard expr(A1,verbose,1)      12use (A1,verbose,1)
3 13this.init(verbose);              13this.init(verbose);
4 14this.shutdown();                 14updateValue();
5 15updateValue();                   15for (k = 0; k < 10; k++) {
6 16k = 0;                            16use (A2);
7 17while (k < 10) {                  17}
8 18wildcard stmt(A2);                18if (foo != null) {
9 19k++;                               19foo.someMethod(42);
10 20}                                  20}
11 21wildcard stmt(A3);                21choice {
12 22foo.someMethod(42);                22case
13 23                                  23System.out.print(foo);
14 24                                  24case
15 25                                  25this.print(foo);
16 26                                  26}
17 27this.shutdown();                 27}
18 28}                                  28}

```

(a) Original part.

(b) Modified part.

Figure 3: Generalized pattern.

of arbitrary code. This solves the accuracy problem of the moved *print* methods in the introductory example.

The *modified* part of the pattern also contains a *choice* annotation. ARES creates this annotation if some training examples add different code. As the added statements have no connection to the original code, ARES (and also other tools) cannot decide which statements lead to the most accurate recommendation. ARES instead lets the developer choose among the variants of the same recommendation. Thus, the *choice* annotation allows ARES to handle additional variations in the training examples to increase accuracy.

The above discussion demonstrates that our pattern design can increase the accuracy of recommendations. How to generate such patterns from examples and how to use them to create accurate recommendation is covered in the next sections.

3 INPUT ORDER DETERMINATION

When constructing the generalized pattern, ARES considers the input code changes one after the other in the loop that is shown in the upper half of Fig. 2. The order in which ARES uses them influences the pattern. Two changes that are very different (i.e., that have a large edit distance) probably lead to a generalized pattern that makes excessive use of wildcards to hide away the differences. The resulting pattern is over-generalized and will match in many locations of the code base. The smaller the edit distance between two changes is, the smaller are the differing code fragments that ARES hides in wildcards. Hence, the key idea is that in every iteration of the loop, ARES first identifies a code change that is as close to the current working generalized pattern as possible. Initially, in the first iteration when there is no working generalized pattern ARES chooses the two code changes that are as close to all other changes in the input set as possible.

To determine the edit distance, ARES uses a tree differencing algorithm that extracts the differences between two abstract syntax trees (ASTs). In general, this is more precise than approaches based on strings or tokens. Thus, ARES uses a tree differencing algorithm to extract code differences throughout the whole process (unless stated otherwise). Tree differencing algorithms differ with respect to their precision in tracking code movements. A tree differencing algorithm that reliably detects code movements leads to an edit distance that better captures the closeness of two input examples, which is important for the input order determination. Thus, ARES uses MTDIFF [14], the currently most precise tree differencing algorithm that considers code movements.

Let us sketch this input ordering process by means of an example. Assume that there are four code changes $c1..c4$. Each of them consists of an original method block oi and a modified method block mi . To obtain the distance between two code changes ci and cj , ARES computes the number of edit operations required to transform oi into oj plus the number of edit operations to transform mi into mj . Table 1 holds some fictitious edit distances for this example. As initial pair, ARES selects the examples that represent the two columns with the lowest sum ($c2$ and $c3$ in the example). The first iteration then constructs a working *Generalized Pattern*. Sec. 4 below describes this step in detail. The next iteration of the loop then computes the edit distances from the working generalized pattern

Table 1: Edit distances between code changes $c1..c4$.

	$c1$	$c2$	$c3$	$c4$
$c1$	-	4	5	6
$c2$	4	-	2	2
$c3$	5	2	-	2
$c4$	6	2	2	-
Σ	15	8	9	10

to the remaining code changes ($c1$, $c4$) and uses the change with the smallest distance to it for the next iteration.

Picking input examples in an order that yields fewer wildcards increases the accuracy of the recommendations as wildcards can hide code transformations. For example, it is possible that the combination of the working pattern with an input example forces ARES to generalize the loops in Fig. 3 into an extra wildcard. The resulting pattern would still match relevant code locations in the search step, but ARES would no longer transform the *while* into a *for* loop. Thus, more wildcards lead to less accurate recommendations.

4 PATTERN CREATION

The input of the *Pattern Creation* step are two code changes ($c1$ and $c2$) with their respective original and modified method bodies. Fig. 4 holds the running example for this section, from which ARES creates the pattern in Fig. 3. As discussed in Sec. 2, this pattern solves the accuracy problems.

The *Pattern Creation* step adds *wildcard* annotations where the original method bodies $o1$ and $o2$ differ. Similarly, it adds *use* annotations where $m1$ and $m2$ differ. To do so, ARES uses the tree differencing algorithm MTDIFF which takes one original and one modified AST as input and then matches nodes from the original AST with nodes of the modified AST. Two nodes are a possible match if they have the same type, e.g., if both are identifiers. MTDIFF uses heuristics for this matching of nodes. Based on the node matches, MTDIFF generates a small *edit script*, i.e., a list of edit operations that transform the original AST into the modified AST. It uses four different types of edit operations on the granularity of AST nodes, namely *delete*, *insert*, *update*, and *move*. The *move* operation moves a complete subtree to a new location.

Any AST node that is part of such an edit operation identifies a change between the method bodies that the *Pattern Creation* handles with annotations. It also keeps all the remaining code unchanged in the pattern to increase the accuracy.

Below we show in detail how ARES uses the edit scripts $D(o1,o2)$ and $D(m1,m2)$ provided by MTDIFF to create the pattern. To determine the correct names in the *use* annotations ARES also requires $D(o1,m1)$ and $D(o2,m2)$. As this process is the most complex part of ARES, we present it in six steps.

4.1 Change Isolation

When ARES generates a pattern from the input examples, it has to make sure that the edit scripts only cover the sections of the code that contain the actual change. Surrounding code that has nothing to do with the change but still is different in the input examples should not be part of the pattern. If there was no change isolation, such surrounding code would cause an over-generalized pattern

<pre> 1 { 2 d = 1.0; 3 this.init(true); 4 this.shutdown(); 5 updateValue(); 6 j = 0; 7 while (j < 10) { 8 String tmp = "bar"; 9 this.init(verbose, tmp); 10 j++; 11 } 12 assert (foo != null); 13 foo.someMethod(23); 14 15 16 17 18 19 20 } </pre>	<pre> { d = 1.0; this.init(true); updateValue(); for (j = 0; j < 10; j++) { String tmp = "bar"; this.init(verbose, tmp); } if (foo != null) { foo.someMethod(23); } this.print(foo); this.shutdown(); } </pre>	<pre> 1 try { 2 this.i = 5; 3 this.init(verbose); 4 this.shutdown(); 5 updateValue(); 6 k = 0; 7 while (k < 10) { 8 updateValue(); 9 printValue("foo"); 10 k++; 11 } 12 13 foo.someMethod(42); 14 15 16 17 18 catch (Exception e) { 19 20 } </pre>	<pre> try { this.i = 5; this.init(verbose); updateValue(); for (k = 0; k < 10; k++) { updateValue(); printValue("foo"); } if (foo != null) { foo.someMethod(42); } System.out.print(foo); this.shutdown(); catch (Exception e) { this.shutdown(); } } </pre>
(a) Original method <i>o1</i> of <i>c1</i> .	(b) Modified method <i>m1</i> of <i>c1</i> .	(c) Original method <i>o2</i> of <i>c2</i> .	(d) Modified method <i>m2</i> of <i>c2</i> .

Figure 4: Code changes *c1* and *c2*.

and thus inaccurate recommendations. In Fig. 4 change *c1* does not have a *try* statement whereas *c2* does. Without a preceding isolation of the main change, MTDIFF would create an *insert* operation for the *try* statement that in turn results in a *wildcard* annotation. This would lead to an over-generalized pattern with a single wildcard. To avoid this, ARES isolates the code parts that actually contain the relevant change, e.g., the body of the *try* node.

ARES implements the necessary *Change Isolation* in two phases. The first phase identifies the lowest nodes in the ASTs that encapsulate the changes, the second phase applies several heuristics if the first phase still leads to an over-generalization. In the first phase, ARES works with the ASTs of the original method o_i and the corresponding modified method m_i of a change. The edit script $D(o_i, m_i)$ helps to identify the lowest root nodes in the two ASTs that are affected by all the edit operations. Each such change-root has the maximal distance from the root node of its AST and still encapsulates all the differences between o and m . In the example, the change-root of $o1/m1$ is the complete method block from line 1 to 20. The change-root of $o2/m2$ is the *try* node (due to the change in the *catch*). As the roots are of different types, selecting these two nodes would still lead to an over-generalization. This and similar input examples are addressed by phase two. In all easier cases, ARES uses the change-roots for isolation and pattern generation to exclude all surrounding code.

If the first phase cannot prevent an over-generalization, the second phase applies the following three heuristics in the described order: (a) Search for similar statements (i.e., statements paired together by MTDIFF) in the children of the change-root of $o1$ and the children of the change-root of $o2$. For $m1$ resp. $m2$, there have to be matching nodes in $D(o1, m1)$ resp. $D(o2, m2)$. (b) Reverse the roles of $c1$ and $c2$ and then try (a) again. (c) Performs (a) and (b) again but this time with the grandchildren instead of the children. If all three heuristics fail, the current implementation of ARES stops without a generated pattern to reduce the execution time.

On the example, phase two identifies similar statements in the block of $c1$ and in the body of the *try* statement. These nodes isolate the actual change, avoid over-generalization in the resulting pattern, and thus reduce the irrelevant and inaccurate recommendations.

4.2 Edit Script Adjustment

While *Change Isolation* avoids over-generalization, *Edit Script Adjustment* finds a balance between over-generalization and over-fitting (i.e., the creation of patterns that only match in very specific situations, e.g., the training set). To keep a balance, ARES uses a rule-based system with over 50 rules (see <https://github.com/FAU-Inf2/ARES>)—far too many to present individually within the space restrictions of this paper. Hence, we can only discuss the two main issues that influence the balance and apply the relevant rules to the running example.

The first issue is that in most cases the tree differencing is too fine-grained. It often identifies many small changes on the level of expressions but only few changes on the level of statements, especially if the training examples are quite similar. The resulting patterns then have *wildcards* and *uses* for fine-grained expressions that often only fit the training examples. The accuracy of the recommendations will be high, but the recall will be low.

To avoid this, ARES looks for matching statements that vary in many of their sub-expressions. In those situations, ARES adjusts the MTDIFF-generated edit script to use a single edit operation that covers the whole statement (instead of many edit operations for all the differing sub-expressions). For example, rule #48 (see Fig. 5) adjusts the edit script for a declaration statement. It applies to line 2 in Fig. 4. Although both the left and the right hand side of the assignment differ in $o1$ and $o2$, MTDIFF keeps the assignment as this keeps the size of the edit script $D(o1, o2)$ small. To avoid over-fitting, ARES adjusts the edit script and replaces all *delete* and *insert* operations for both the left and the right hand side of the assignments (lines 8–9 in Fig. 5) with one *insert* operation for a full statement (line 10 in Fig. 5). Adding a *delete* operation for the assignment is unnecessary as this would only lead to a wildcard at the same position.

Since finding an optimal list of edit operations that includes code movements is NP-hard [10, 11], all move-aware tree differencing algorithms rely on heuristics and there are cases in which the edit script is not optimal. Any non-optimal list leads to unwanted

```

1: function RULE48(editOps, mapping)
2:   declMappings ← getDeclarationMappings(mapping)
3:   for (do1, do2) ∈ declMappings do
4:     lo1 ← leftSide(do1); lo2 ← leftSide(do2)
5:     ro1 ← rightSide(do1); ro2 ← rightSide(do2)
6:     if isDeleted(lo1, editOps) ∧ isDeleted(ro1, editOps) then
7:       if isInserted(lo2, editOps) ∧ isInserted(ro2, editOps) then
8:         removeOpsForTree(lo1, editOps); removeOpsForTree(ro1, editOps);
9:         removeOpsForTree(lo2, editOps); removeOpsForTree(ro2, editOps);
10:        addInsertForNode(do2, editOps)

```

Figure 5: Edit script adjustment Rule #48.

wildcards and thus increases the generality unnecessarily. This is the second balancing issue that the edit script adjustment addresses.

For instance, Rule #31 examines moves across nested code blocks. It applies to line 3 in Fig. 4. MTDIFF determines that the number of edit operations is minimized if the code of line 9 in *o1* is moved to line 3 in *o2*. As a consequence, MTDIFF generates a *move* operation. Due to this *move* operation, ARES would insert a wildcard in line 3. This is too general. Instead, the adjustment step replaces the *move* with a *delete* operations for the call in line 9 of *o1*. Then it adds the *insert* operation for *verbose* in line 3 of *o2*.

Rules #13–19 handle movements of identical statements. For example, there exist two identical statements in *o2* (lines 5 and 8) for the call in line 5 of *o1*. Although this is not a problem for the running example, the adjustment step has to take care of wrong pairs of two identical statements inside the tree differencing results. Otherwise, there are unnecessary *moves* in the list of edit operations which can lead to an over-generalized pattern.

Rules #42–46 handle changes that are already covered by changes of the parent statement. These rules apply to the lines 8 and 9 of *o1* that are replaced by new statements in *o2*. MTDIFF generates *delete* and *insert* operations for each AST node in both lines. For example, MTDIFF generates a *delete* operation for *String*, *tmp*, etc. For the *Pattern Creation* only the *insert* for the invocations of *updateValue* and *printValue* in *o2* are relevant. Thus, the rules remove all edit operations on nodes that are part of the statements. For the *delete* of the *assert* in line 12, ARES also replaces the deleted expressions with a single *delete* of the complete statement.

4.3 Match Insertion

This step inserts the *match* annotation into the generalized pattern (see Fig. 3). The main purpose of this annotation is to provide a list of identifier names that differ between *o1* and *o2*. Using wildcards for them is too verbose and would unnecessarily enlarge the patterns. Instead, ARES uses the matched node pairs from $D(o1, o2)$ to find identifiers in the same match pair but with different names. For the running example, this is only the pair (*j*, *k*) of the loop variables. By mentioning the names in the annotation, ARES can remove them from all *update* operations in both $D(o1, o2)$ and $D(m1, m2)$.

4.4 Wildcard and Use Insertion

This step adds the *wildcard* and *use* annotations to the pattern according to the edit operations that are left after the previous steps. ARES replaces each statement that an *insert* adds in *o2* with a wildcard. If the *insert* operation adds an expression, ARES adds a *wildcard* annotation in front of the statement. Similarly, for each *insert* that affects expressions or statements in *m2* ARES adds *use*

annotations. For each *delete* operation that removes a node from *o1* ARES adds a *wildcard* annotation at the corresponding spot in *o2* (determined with $D(o1, o2)$ and the heuristics). In the same way, ARES replaces each *delete* in *m1* with a *use* annotation in *m2*. As *move* operations basically delete a node in *o1/m1* and add it in *o2/m2*, ARES adds wildcards for the *delete* and *insert* operation expressed by the *move*. Note that at this point there are no longer *update* operations as the edit script adjustment either removed them or replaced them with *insert* operations. ARES also memorizes which annotation belongs to which edit operation in order to facilitate the following name assignment step.

Since in the example the *insert* of *verbose* is left in $D(o1, o2)$, ARES adds a *wildcard* annotation before the statement and tags it with *expr*. Additionally, it specifies which expression corresponds to the wildcard (*verbose*). As it is possible that an expression occurs several times on the same statement, the wildcard in line 3 of the final pattern in Fig. 3(a) also specifies the number of the occurrence (1 for the example). Fig. 3(b) shows the corresponding *use*.

Another operation is the *insert* of the assignment in line 2 of *c2*. ARES replaces this assignment with a *wildcard* annotation in *o2*. As it is a replacement for a complete statement, the tag *stmt* is added to the wildcard. Similarly, ARES replaces the assignment in *m2* with a *use* annotation. For the inserted statements in lines 8 and 9 ARES also inserts *wildcard* and *use* annotations. After the insertion of the annotations in line 9, ARES immediately combines both adjacent annotations into a single one.

ARES proceeds with the remaining edit operations in this fashion and finally creates the result in Fig. 3.

4.5 Wildcard Name Assignment

This step assigns the names that link the *wildcard* annotations in the original part to the *use* annotations in the modified part. Since these names can occur in different spots on both sides of a pattern, they encode code movement. Hence, they are crucial for the accuracy of the recommendations.

To identify the correct *wildcard/use* pairs, this step examines the statements that were replaced by wildcards (memorized by the previous step). In the running example, a wildcard replaced *updateValue* in line 8 of *o2*. Then ARES uses $D(o2, m2)$ to find a node in *m2* that is matched to the moved *updateValue* from *o2*. In the example, the matched node is the call of *updateValue* in *m2*. As *updateValue* in *m2* was also replaced by a *use*, ARES links the *wildcard* and *use* of the call together and gives them the same name (*A2* in the example). Similarly, ARES assigns the other names to create Fig. 3.

The previous steps may create a pattern that starts with a *stmt* wildcard. However, when searching for applications it is unclear which sequence of statements this wildcard should match. Therefore, ARES enforces that patterns begin with a specific statement instead of a wildcard. In the running example, ARES removes the *wildcard* that replaced the assignment in line 2 of *o2*. If there is a corresponding *use* with the same name and this *use* is also the first statement in the pattern, ARES also removes it. The same applies to *stmt* wildcards at the end of the pattern as they match the complete remaining function and thus the recommendation would be unnecessarily large. The evaluation shows that this has no negative

impact on precision and recall compared to other recommendation systems. If the assigned *use* of a removed wildcard is not at the top or bottom of the pattern, ARES keeps the *use* annotation and only removes the assigned name. This is necessary since the corresponding *wildcard* is no longer present.

4.6 Choice Insertion

The final step handles differences between the modified parts of the code changes that do not correspond to code in the original parts. To increase the accuracy of the recommendations it is necessary to handle those differences explicitly. After the previous steps, the differences are visible as they correspond to *use* annotations without assigned names. This step replaces each such *use* with a *choice* annotation because the input examples provide insufficient information to determine the right recommendation based on a purely syntactical approach.

In the running example, the *use* that replaced the *assert* and *print* call has no assigned name and is thus changed into a *choice* annotation. Based on *c1* and *c2* it is impossible to determine which *print* invocation should be part of the recommendation. Hence, ARES generates several recommendations, one for each variant. It is up to the developer to decide which of the recommendations is the most appropriate. Similar to the name assignment process, ARES only has to examine the statement that was replaced by the *use* annotation to identify the code for the *case* annotations in Fig. 3(b).

5 SEARCH FOR APPLICATIONS

This section explains how ARES finds code locations where a given pattern is applicable. For a high accuracy, it is important that the algorithm identifies the correct matches between the AST nodes in the pattern (including wildcards) and AST nodes at the code location. To make the algorithm as fast as possible, this section no longer relies on a tree structure but uses a serialized list of AST nodes. However, this is no limitation to the movement support as the *Recommendation Creation* (Sec. 6) handles them.

As example we use the pattern in Fig. 3 and browse the code in Fig. 6(a). ARES first searches for suitable starting points in the code and then executes the AST-node matching from there in parallel. Such a starting point is an AST node that has the same type as the first AST node in the original part of the pattern and that is not part of an annotation. The first node in Fig. 3 is the method call in line 3. Suitable starting points in Fig. 6(a) are the calls in lines 2, 3, 4, 7, and 10. ARES uses each of these calls as a *startNode* when executing the algorithm given in Fig. 7. The other input arguments are the AST of the method that contains *startNode* and the AST of the original part of the pattern.

In lines 2 and 3 of Fig. 7, ARES generates a list of AST nodes for the code location (starting at *startNode*) and a list of AST nodes for the pattern (starting from the first code node in the body of the *match* annotation). With the loop in lines 6–29 the algorithm compares both lists of AST nodes to determine whether they are identical (with respect to the wildcards), in which case the pattern is applicable to the code location that starts at *startNode*.

In the running example, both lists start with a *call* node. As the node of the pattern is not a wildcard, *w* (line 13) is *null* and the algorithm uses *isMatch* to compare *n_{cl}* of the code location with

```

1 Foo foo = Library.getObject();   Foo foo = Library.getObject();
2 this.init(getVerbose());         this.init(getVerbose());
3 this.shutdown();                 updateValue();
4 updateValue();                   for (c = 0; c < 10; c++) {
5 c = 0;                            System.out.println(c);
6 while (c < 10) {                  }
7 System.out.println(c);           if (foo != null) {
8 c++;                               foo.someMethod(99);
9 }                                  }
10 foo.someMethod(99);              this.shutdown();
11 return foo;                       return foo;

```

(a) Matching code location. (b) Recommendation.

Figure 6: Recommendation Creation.

n_p of the pattern. The function *isMatch* returns true if the types of these AST nodes are identical (e.g., if both are *call* nodes). There are only two exceptions. If both AST nodes are identifiers, *isMatch* is true if *n_p* is in the list of identifiers of the *match* annotation or if the identifier names are equal. The second exception concerns boolean constants as they make a large difference in the pattern due to their limited value range. In this case, *isMatch* is only true if both are identical. ARES does not compare the values of other constants. This increases the generalization of a pattern and makes it applicable to more code locations because different literals do not prevent a pattern from being applicable.

If the comparison with *isMatch* is successful, the algorithm adds the pair (*n_{cl}*, *n_p*) to the set of matched nodes (*matches*). The recommendation algorithm later uses this set to replace the identifiers in the pattern (e.g., *k*) with identifiers of the code location (e.g., *c*) and also to replace constants in the pattern (e.g., 42) with constants in the code location (e.g., 99). If the comparison with *isMatch* is unsuccessful, the algorithm looks for another valid position to backtrack

```

1: function SEARCH(startNode, methodBody, originalPatternPart)
2:   NLcl ← getNodes(methodBody, startNode)
3:   NLp ← getPatternNodes(originalPatternPart)
4:   poscl ← 0; posp ← 0
5:   resets ← ∅; matches ← ∅; visited ← ∅
6:   while poscl < |NLcl| ∧ posp < |NLp| do
7:     ncl ← getNode(NLcl, poscl); np ← getNode(NLp, posp)
8:     w ← null
9:     if isWildcard(np) then
10:      w ← np
11:     else if hasAssociatedWildcard(np) then
12:      w ← getAssociatedWildcard(np)
13:     if w = null then
14:       if isMatch(ncl, np, originalPatternPart) then
15:         matches ← (ncl, np)
16:         poscl ← poscl + 1; posp ← posp + 1
17:         continue
18:     else
19:       if w ∉ visited then
20:         addLast(resets, (poscl, posp + 1, matches, visited))
21:         visited ← visited ∪ w
22:       if allowedNode(w, ncl) then
23:         matches ← (ncl, w)
24:         if allowedReset(w, ncl) then
25:           addLast(resets, (poscl + 1, posp + 1, matches, visited))
26:         poscl ← poscl + 1
27:         continue
28:       if resets = ∅ then return null
29:       (poscl, posp, matches, visited) ← removeLast(resets)
30:   if posp = |NLp| then
31:     return matches
32:   else
33:     return null

```

Figure 7: Algorithm to search for suitable code locations.

(lines 28–29). If there is a valid position, ARES resets the variables to the saved values and continues from the last valid position.

Backtracking is necessary due to the use of wildcards. The pattern design (see Sec. 2) allows a wildcard to replace none or more statements/expressions. This leads to several valid end points of a wildcard match and creates the reset nodes. In the example, the node *verbose* is associated with a wildcard. Thus, *getAssociatedWildcard* identifies the wildcard in line 3 of the pattern. As this is the first appearance of this wildcard, the algorithm creates a valid reset position in line 20. This reset position covers the case that the wildcard is empty and does not match any code at this location. The reset point starts at the position after *verbose* ($\text{pos}_p + 1$) and thus continues without a match.

In our example, *init* has an argument and thus the wildcard has a node to match. The function *allowedNode* checks whether the wildcard can replace n_{cl} . A replacement is possible if n_{cl} is within the scope of the wildcard. For the wildcard in line 2, the scope consists of the arguments of *init*. Here this is the case and the algorithm adds the appropriate match (line 23). If n_{cl} is also the last node of one argument of *init*, the node is also a valid end point of the wildcard and the algorithm adds a new reset point that starts at the end of the argument (lines 24–25). Then the search continues at the next node in the list of the code location (lines 26–27). The algorithm handles the other wildcards (lines 8, 11) in a similar fashion. The difference is only that this time *allowedNode* accepts all nodes from the current code block and *allowedReset* accepts only complete statements.

After the loop terminates, the algorithm checks whether the process reached the end of the template and thus whether the code location matches all nodes of the template. In this case the search is successful and returns the set of matched nodes (lines 30–33).

6 RECOMMENDATION CREATION

The general idea of the creation step is to use MTDIFF to create the list of edit operations that change the *original* part of the pattern into the *modified* part (see Fig. 3) and to apply these edit operations to a copy of the code location (identified in the previous step). Note that ARES does not use the *modified* part of the pattern as the base for the recommendation. Using the list of edit operations instead has the advantage that ARES can preserve parts of the copied code, e.g., the identifier *c* and the number *99* (see Fig. 6(a)). This increases the accuracy of the recommendation.

To be able to apply the list of edit operations, the copied code has to look like the *original* part of the pattern. For that purpose, ARES removes all AST nodes from the copy that were matched with a wildcard during the search. Then ARES inserts the wildcard annotations into the copy. After these changes, the copy looks like the *original* part of the pattern in Fig. 3(a) except for the preserved code parts. This allows ARES to apply the edit operations to create the *modified* part of the pattern (with respect to the preserved code parts). Then ARES replaces the *use* annotations with the code that is paired with the corresponding wildcard, i.e., the wildcard with the same name.

Finally, ARES works on the *choice* annotations. For a single *choice* ARES could create one copy of the recommendation per *case*. If there are more *choices*, then conceptually there is the potential for

exponential growth. Hence, ARES limits the number of copies. If there are at most *max* case statements in all *choices* (2 in the example), ARES creates *max* + 1 copies of the current recommendation code. The *n*-th copy contains the code in the *n*-th *case* annotation of each *choice*. If a *choice* annotation has fewer than *n* cases, we omit this choice completely. This implies that the last copy does not contain any code from *choice* annotations. After this step, ARES presents the *max* + 1 copies as variants of one recommendation to the developer.

7 EVALUATION

We first evaluate the accuracy of ARES and compare it with that of LASE [29], before we show that the improved accuracy does not gravely impact precision, recall, or execution times. For all measurements we use a workstation with 128 GB of RAM and a 3.6 GHz Intel Xeon CPU, OpenJDK 8 and Ubuntu 16.10. We also discuss the current limitations of ARES in this section.

7.1 Accuracy

Our benchmarks comprise several real-world source code archives that vary with respect to the number and size of the sets of training examples (similar code changes). We always check how close the recommended code changes get to changes that can actually be found in the commits.

Eclipse–Two Input Code Changes. In this part, we use 2 groups of similar code changes (Bugzilla Ids 77644, 82429) from the *Eclipse JDT* project [1] and 21 groups from the *Eclipse SWT* project [2]. Meng et al. [29] used the same *manually* collected 23 groups for their evaluation of LASE. All these code changes are present in commits of the respective repositories.

Table 2 lists the *Bugzilla Ids* of the changes; *m* is the size of the group of similar code changes that exist in the repositories for a bug. In most cases, *m* is equal to the evaluation provided by Meng et al. The only exceptions are the rows with Ids 20 and 21 for which we identified a different number of changes in the repository.

For the next two segments of Table 2 (*LASE–Two Input Code Changes*, *ARES–Two Input Code Changes*) we use the same *two* changes of a group for the creation of the patterns. We carefully selected the two changes to reproduce the same precision and recall values that Meng et al. [29] list in their evaluation.

For each of the 23 bug fixes, the Δ -columns list the number of code locations for which LASE and ARES create a recommendation. The columns marked with \checkmark show how many of the *m* manually identified locations the tools find. For each of them column A_T and A_C give the accuracy, i.e., the closeness of the recommendation to the code that the original developers of the change wrote. Thus, a perfect accuracy means that the recommendation has the same statements, moved code and identifiers that are present in the repository. For developers this is important because a higher accuracy means less work to actually apply a recommendation to a project. To measure the accuracy we compute the Levenshtein distance [24] (LD) between the method body of the recommendation and the method body of the changed code from the repository. We use two variants. The first uses AST tokens (T), the second actual characters (C). A high token accuracy A_T shows that the recommendation is accurate with respect to the syntax (e.g., uses the same number

Table 2: Comparison with LASE on 23 code changes from Eclipse JDT and Eclipse SWT.

Id	Bugzilla Id	m	LASE—Two Input Code Changes						ARES—Two Input Code Changes					ARES—All Code changes						
			Δ	✓	A _T %	A _C %	P%	R%	Δ	✓	A _T %	A _C %	P%	R%	Δ	✓	A _T %	A _C %	P%	R%
1	77644	4	2	2	52	52	100	50	2	2	84	83	100	50	4	4	90	89	100	100
2	82429	16	14	14	66	40	100	88	21	14	68/81	76/84	67	88	25	16	54	61	64	100
3	114007	4	4	4	94	81	100	100	4	4	100	100	100	100	4	4	100	100	100	100
4	139329 ₁	6	2	2	99	86	100	33	2	2	100	100	100	33	35	6	80	82	17	100
5	142947 ₁	12	12	12	98	89	100	100	12	12	100	100	100	100	12	12	100	100	100	100
6	91937	3	3	3	50	37	100	100	3	3	100	100	100	100	3	3	100	100	100	100
7	103863	7	7	7	32	43	100	100	7	7	100	100	100	100	7	7	100	100	100	100
8	129314	4	4	4	96	80	100	100	3	2	94/100	94/100	67	50	5	4	95/100	95/100	80	100
9	134091	4	4	4	79	56	100	100	4	4	99	99	100	100	4	4	99	99	100	100
10	139329 ₂	3	4	3	63	64	75	100	3	3	95	94	100	100	3	3	57/100	56/100	100	100
11	139329 ₃	3	3	3	92	75	100	100	3	3	100	100	100	100	3	3	100	100	100	100
12	142947 ₂	9	12	9	74	61	75	100	7	6	78/89	74/87	86	67	12	9	78/90	75/89	75	100
13	76182	6	6	6	56	58	100	100	6	6	92	92	100	100	6	6	92	92	100	100
14	77194	3	3	3	79	58	100	100	2	2	98	98	100	67	3	3	96	97	100	100
15	86079 ₁	3	3	3	92	69	100	100	2	2	100	100	100	67	3	3	93	92	100	100
16	95409	9	8	8	49	46	100	89	4	4	42/85	43/84	100	44	18	9	45/68	47/69	50	100
17	97981	4	3	3	71	56	100	75	3	3	100	100	100	75	4	4	100	100	100	100
18	76391	6	3	3	97	86	100	50	3	3	100	100	100	50	9	6	100	100	67	100
19	89785	5	5	5	91	75	100	100	5	5	97	94	100	100	5	5	84/100	79/99	100	100
20	79107	10	26	10	96	75	38	100	12	4	99	99	33	40	27	10	98	97	37	100
21	86079 ₂	3	2	2	99	73	100	67	2	2	100	100	100	67	3	3	73/100	75/100	100	100
22	95116	5	4	4	97	83	100	80	4	4	100	100	100	80	5	5	100	100	100	100
23	98198	15	67	12	61	51	18	80	38	10	71/93	75/94	26	67	356	15	76/95	79/95	4	100
Avg.		6	9	5	78	65	92	87	7	5	92/96	92/96	90	76	24	6	87/94	88/94	82	100

m: Available Locations; Δ: Generated Recommendations; ✓ Correct Recommendations;
A_T %: Token Accuracy; A_C %: Character Accuracy; P%: Precision; R%: Recall;

of *if* statements). We use $A_T = 1 - LD_T / \max(|r_T|, |m_T|)$, where $|r_T|$ is the number of tokens in the recommendation and $|m_T|$ is the number of tokens in the method body from the repository. As comments are not part of any token, they are ignored for this measurement. We also define $A_C = 1 - LD_C / \max(|r_C|, |m_C|)$, where $|r_C|$ is the number of characters in the recommendation and $|m_C|$ is the number of characters in the method body from the repository. This includes both comments and whitespace. Each cell in the accuracy columns contains the mean of the correctly identified recommendations (✓).

Across all groups of code changes, ARES can produce more accurate recommendations compared to LASE and even achieves a perfect accuracy (100%) for 11 of the 23 groups. For these groups, the recommendations are identical to the changes that were actually performed by a human developer. A closer inspection of the recommendations by LASE shows that LASE ignores code transformations that occur only in one of the input examples which reduces A_T . For some cases, the ARES accuracy column contains two values. In these cases, the pattern contains a choice annotation and ARES generates several recommendation variants. A_T and A_C give the minimal and the maximal accuracy values for the corresponding groups. In 3 of the 5 cases (Ids: 2, 12, 23), even the minimal values for A_T of ARES outperform LASE. Often (for 11 groups) ARES achieves perfect accuracy values (100%) in A_C . Hence, ARES recommends exactly the code a developer has written, including comments, coding style, and whitespace.

Eclipse—All Code Changes. In some cases (e.g., patterns for critical bugs) a high recall is more important than a high precision (and accuracy). For these cases, ARES supports more than two input changes. Each additional change in the training set can increase the number of wildcards and hence the recall. However, each additional wildcard can also decrease precision and accuracy. To examine the effects of additional input changes we added a third set of evaluation results in Table 2 (ARES—All Code Changes). For these results, ARES

uses all input changes to create a pattern. Thus, we maximize the recall and minimize precision and accuracy for each of the groups.

For this set, the accuracy of ARES is still at 94% (87% in the minimal choice case) on average. In most cases, the accuracy of ARES in this configuration is also higher than the accuracy of LASE with only two input changes. This means that even with the most general pattern for a group and thus the pattern with the least accuracy, ARES still achieves higher accuracy values than LASE. Also, ARES still has a perfect accuracy for 8 groups.

JUnit—All Code Changes. The above analysis shows that ARES achieves a higher accuracy for the 23 groups of code changes from Eclipse. A threat to the validity of these results is that it is possible to optimize a system for such a small dataset. To examine whether ARES also has higher accuracy values for a larger dataset, we use 3,904 groups of code changes from JUnit [3] taken from the results of C3 [23]. C3 is a tool that identifies groups of similar code changes in code repositories with the help of clustering algorithms. The groups by C3 are suitable inputs for ARES and LASE. This time we use all input examples (not only two) to train ARES and LASE.

Whereas it is again possible to measure the accuracy (as the manual code changes are in the repository) there cannot be numbers for precision and recall as m is unknown for this dataset.

Table 3 shows the accuracy results for JUnit. For 482 sets of input changes both LASE and ARES produce recommendations and generate a recommendation for the input changes. We excluded input changes for which one of the tools produces an error or does not recommend a change. Similar to the evaluation above, ARES reaches a high accuracy of 91% to 100%. Again for most groups, the accuracy of ARES is higher than that achieved by LASE. This is also true for the minimal accuracy values.

7.2 Precision, Recall, and Time Measurements

Precision & Recall. Let us now compare the precision $P(\frac{\checkmark}{\Delta})$, and recall $R(\frac{\checkmark}{m})$ values. The higher the precision and recall, the better

is the tool. Overall, the precision of ARES is similar to that of LASE. For 20 of the 23 groups of code changes the precision of ARES is identical to or higher than that of LASE. The average recall of ARES is still at a high level (76%) but below that of LASE (87%). However, Christakis and Bird [13] found in their study that the recall is less important to developers compared to a high precision. Hence, with respect to precision and recall there are (almost) no costs for the improved accuracy of ARES.

A closer look reveals that ARES is more precise than LASE for three change groups (Ids: 10, 12, 23) because of two reasons. First, ARES can handle variations in the input changes with appropriate wildcards. Instead, LASE uses the *Maximum Common Embedded Subtree Extraction* algorithm to identify the common AST that all training examples share and which is more general in some cases. Second, ARES keeps common code, even if it is unrelated to the changes, whereas LASE identifies code that has no dependencies to the code change and excludes it from the pattern.

While it can be beneficial to keep some unrelated code in the pattern, keeping too much of it causes a loss of precision (Id 8) or recall. There is room for more research concerning this issue. In contrast to LASE, ARES is also currently limited to the method body and does not include the method signature in the patterns. This lack of a signature in the pattern lowers the precision in two cases (Ids: 2, 20).

The groups 20 and 23 are outliers as the precision of both ARES and LASE is considerably lower compared to the other groups. The reason is that the input examples for both groups only add the same statements and otherwise have very little code in common. This leads to short and very general patterns.

Time Measurements. Here we compare the times that ARES and LASE take to create the patterns and to use them when they browse the projects for possible pattern applications (Eclipse SWT: ~2,000 Java files, ~375,000 lines of code; Eclipse JDT: ~5,750 Java files, ~1,372,000 LOC). Fig. 8 shows the measurements. Each code change (row in Table 2) corresponds to one measurement point. The lines in the boxes are the medians of the time measurements. The boxes of the plot define the 25% and 75% quartiles, the whiskers show the minimum and maximum.

ARES creates patterns faster than LASE (*Pattern Creation* in Fig. 8) because the ChangeDistiller [17] tree differencing algorithm that LASE uses has a higher runtime than MTDIFF [14]. However, the search for pattern applications in ARES (*Pattern Use* in Fig. 8) is slower because the patterns of ARES are currently limited to method bodies and do not contain method signatures. LASE can use the method signature to filter out methods and thus has to inspect fewer method bodies. Still, even for such large repositories ARES completes the search for one pattern within a minute in most cases.

Table 3: Accuracy on JUnit.

	LASE	ARES (Min/Max)
Groups of Code Changes		3,904
Shared Recommendations		482
Shared Recommendations A_T (Mean) %	90	91/97
Shared Recommendations A_T (Median) %	100	100/100
Shared Recommendations A_C (Mean) %	76	90/95
Shared Recommendations A_C (Median) %	82	100/100

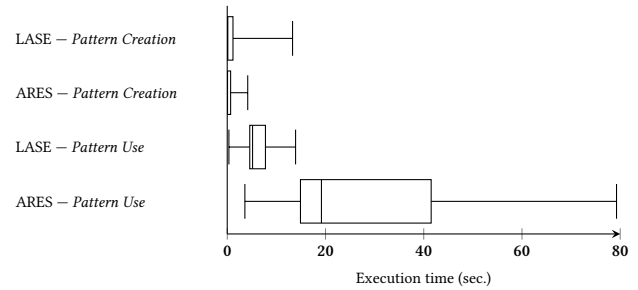


Figure 8: Time per change group for two input code changes (25%/75% quartiles, whiskers: minimum/maximum).

7.3 Limitations and Threats to Validity

The major limitation of ARES is the current restriction to method bodies. Due to this restriction we also had to exclude the group of code changes with Bugzilla Id 74139 that Meng et al. [29] use in their evaluation of LASE. This was necessary as the code changes in this group only have a common signature but no common statements.

Whereas both the search and recommendation creation of ARES already support larger changes, extending the pattern creation to work with full classes would need annotations that, for example, define when methods or fields can be reordered. The pattern creation would also require new rules and transformations for wildcards outside of code blocks.

Another limitation is that ARES (like LASE) can only search for one pattern at a time. Similar to their application in code clone detection tools, suffix trees [21] may accelerate the search for several patterns in parallel.

The threat to validity of the evaluation is the implementation of LASE that we built from a publicly available version [4]. We had to apply several bug fixes to be able to replicate the evaluation results of Meng et al. [29]. Despite our efforts, there are still some small differences left. We also changed the implementation to bypass the UI to enable script-based performance measurements. We argue that if we have introduced new errors they are probably small since we obtained most of the original results.

8 RELATED WORK

The book *Recommendation Systems in Software Engineering* [34] and the study by Gasparic and Janes [18] provide an overview of the related work. We discuss the works closest to ARES.

Example-based recommendation. LASE [29] is closest to ARES. The main differences are that LASE does not handle moved code parts accurately and thus the accuracy of the recommendations is lower. RASE [27] relies on LASE to create a generalized pattern from examples. With this pattern it refactors the code without altering its semantics to replace all changed locations with a single unifying code fragment. Thus, it has the same low accuracy as LASE and also only supports refactorings. In contrast to ARES, SYDIT [28] generates an edit script from just one example which limits its generalization ability. Developers also must find pattern applications manually with SYDIT whereas ARES finds them automatically. Critics [43] addresses the review process. The developer provides a generalization of a single code change as input. Critics then finds matching spots in the code where a similar change may have been forgotten or where an inconsistent modification may

have occurred. For Critics, it is the task of the developer to provide the generalization and to take care of the design of the pattern to increase the accuracy. ARES creates the pattern automatically.

REFAZER [35] is another tool that learns code transformations from examples. It uses a domain specific language (DSL) to represent code change patterns. In contrast to ARES, the algorithms that generate this pattern do not support code movements and thus if a code pattern relies on movements it cannot be as accurate as ARES. Currently, a direct comparison is not possible as REFAZER only supports C# and Python whereas ARES uses Java.

Santos et al. [36] compare three different ways (structural, AST-based, information retrieval) to search for additional code change locations based on up to two examples. If a search finds such a location, their system applies a set of code changes. If this is successful, their system recommends the code change. Their AST-based approach uses the longest common subsequence (LCS) [9] and thus is less precise than ARES in regard to moved code. Also this approach is time consuming as it requires the execution of the LCS on many different ASTs. ARES is faster as its backtracking algorithm can abort the search for possible pattern applications earlier. Their information retrieval approach relies on the similarity of code parts (e.g., variable names) and is less precise than ARES that includes the code structure in the pattern. Their structural approach searches for methods with similar signatures, packages, etc. The authors argue that this is useful in combination with the AST-based and information retrieval approaches. It is possible to include their structural approach in ARES to increase precision.

Tate et al. [37] use a proof-checker to learn transformations from examples. They can only find provably semantic-preserving transformations.

Padioleau et al. [32] introduce semantic patches to manually generalize patches obtained from standard diff tools. Andersen et al. [5, 6] extend this idea and introduce a tool that creates generalized (semantic) patches from hand-picked diffs. ARES is more accurate as it supports code movements whereas diffs are limited to *insert* and *delete* operations.

Code completion. Cookbook [19] uses generalizations from examples to suggest code completions while the developer is typing. In contrast to ARES it is a line based approach and does not support code movements and thus has a lower accuracy if code movements are necessary. MAPO [44] and Precise [42] search for similar API usages in code repositories. MAPO recommends frequently used call sequences. Precise extracts API calls including their argument values and corresponding declarations from repositories, extracts groups with similar arguments using a k-nearest neighbor algorithm, and generates API usage recommendations, including possible parameters. Bruch et al. [12] also use a k-nearest neighbor algorithm and find code fragments that are similar to code that is currently being developed. Bajracharya et al. [7] use structural semantic indexing for this purpose. All focus on code completions and thus small coding suggestions. In contrast, ARES strives to suggest larger transformations including changes of complete methods.

Specialized recommendation. In addition to example-based approaches, there are also tools for specific tasks. CFix [20] uses predefined patterns to automatically fix concurrency bugs. Weimer et al. [41] and AutoFix-E [40] generate bug fixes automatically, but are

limited to testable fixes and specifications. Robbes and Lanza [33] learn code transformations by examining edit operations in the IDE. Their tool cannot automatically find locations. Stratego/XT [38] and DMS [8] offer DSLs to specify AST transformations. Manniesing et al. [25] focus on loop transformations. Other works solely locate patterns in source code, e.g., the Dependency Query Language [39] or the Program Query Language [26]. In contrast to ARES, developers must specify patterns manually. They also do not get recommendations. Miller et al. [30] let developers select multiple code fragments and then they apply a change to all of them in the same fashion. In contrast to ARES this is limited to identical changes and developers have to find the code locations themselves. Thung et al. [16] introduce a recommendation system that supports developers in backporting Linux drivers. Their framework is specifically tailored to the backporting and is not intended for general code change recommendations.

9 CONCLUSION

We presented the novel Accurate REcommendation System (ARES) that specializes on code movements to increase the accuracy of code recommendations. A higher accuracy means that ARES generates code recommendations that better reflect what a developer would have written. ARES achieves the higher accuracy results with a pattern design that expresses code movements more accurately compared to the state-of-the-art. Similar to other tools, ARES generates these patterns from source code training examples. The generated patterns of ARES contain only plain Java code with a set of annotations and therefore are not a black box for developers. Thus, developers can read and manually adapt the patterns.

We also presented in detail how ARES generates, generalizes, and applies these patterns. With these techniques ARES achieves an average recommendation accuracy of 96% in our evaluation and outperforms LASE. Precision and recall are on par. The execution time is a bit higher, but still below two minutes for large real-world source code archives.

For reproducibility and to kindle further research, we open-source ARES, the rules for the edit script adjustment, all the evaluation inputs and results, including the human-readable patterns generated by ARES (<https://github.com/FAU-Inf2/ARES>).

REFERENCES

- [1] 2017. Eclipse JDT Repository. (2017). [git://git.eclipse.org/gitroot/jdt/eclipse.jdt.core.git](https://git.eclipse.org/gitroot/jdt/eclipse.jdt.core.git).
- [2] 2017. Eclipse SWT Repository. (2017). <http://git.eclipse.org/c/platform/eclipse.platform.swt.git/>.
- [3] 2017. JUnit Repository. (2017). <https://github.com/junit-team/junit4.git>.
- [4] 2017. LASE. (2017). <https://www.cs.utexas.edu/%7Emengna09/projects.html>.
- [5] Jesper Andersen and Julia L. Lawall. 2008. Generic Patch Inference. In *ASE'08: Intl. Conf. Automated Softw. Eng.* L'Aquila, Italy, 337–346.
- [6] Jesper Andersen, Anh Cuong Nguyen, David Lo, Julia L. Lawall, and Siau-Cheng Khoo. 2012. Semantic Patch Inference. In *ASE'12: Intl. Conf. Automated Softw. Eng.* Essen, Germany, 382–385.
- [7] Sushil K. Bajracharya, Joel Ossher, and Cristina V. Lopes. 2010. Leveraging Usage Similarity for Effective Retrieval of Examples in Code Repositories. In *FSE'10: Intl. Symp. Foundations of Softw. Eng.* Santa Fe, NM, 157–166.
- [8] Ira D. Baxter, Christopher Pidgeon, and Michael Mehlich. 2004. DMS: Program Transformations for Practical Scalable Software Evolution. In *ICSE'04: Intl. Conf. Softw. Eng.* Edinburgh, Scotland, 625–634.
- [9] Lasse Bergroth, Harri Hakonen, and Timo Raita. 2000. A Survey of Longest Common Subsequence Algorithms. In *SPIRE'00: String Processing and Inf. Retrieval Symp.* A Coruna, Spain, 39–48.

- [10] Philip Bille. 2005. A survey on tree edit distance and related problems. *Theoretical Computer Science* 337, 1 (June 2005), 217–239.
- [11] Utsav Boobna and Michel de Rougemont. [n. d.]. Correctors for XML Data. In *Database and XML Technologies*. Lecture Notes in Computer Science, Vol. 3186. 97–111.
- [12] Marcel Bruch, Martin Monperrus, and Mira Mezini. 2009. Learning from Examples to Improve Code Completion Systems. In *ESEC/FSE'09: Europ. Softw. Eng. Conf. and Symp. Foundations Softw. Eng.* Amsterdam, The Netherlands, 213–222.
- [13] Maria Christakis and Christian Bird. 2016. What Developers Want and Need from Program Analysis: An Empirical Study. In *ASE'16: Intl. Conf. Automated Softw. Eng.* Singapore, Singapore, 332–343.
- [14] Georg Dotzler and Michael Philippsen. 2016. Move-optimized Source Code Tree Differencing. In *ASE'16: Intl. Conf. Automated Softw. Eng.* Singapore, Singapore, 660–671.
- [15] Georg Dotzler, Ronald Veldema, and Michael Philippsen. 2012. Annotation Support for Generic Patches. In *RSSE'12: Proc. Intl. Workshop Recommendation Systems for Softw. Eng.* Zurich, Switzerland, 6–10.
- [16] David Lo Ferdian Thung, Le Dinh Xuan Bach and Julia Lawall. 2016. Recommending Code Changes for Automatic Backporting of Linux Device Drivers. In *ICSME'16: Intl. Conf. Softw. Maintenance and Evolution*. Raleigh, NC, 222–232.
- [17] Beat Fluri, Michael Wuersch, Martin Pinzger, and Harald Gall. 2007. Change Distilling: Tree Differencing for Fine-Grained Source Code Change Extraction. *IEEE Transactions on Software Engineering* 33, 11 (Nov. 2007), 725–743.
- [18] Marko Gasparic and Andrea Janes. 2016. What Recommendation Systems for Software Engineering Recommend: A Systematic Literature Review. *Journal of Software and Software* 113 (2016), 101–113.
- [19] John Jacobellis, Na Meng, and Miryung Kim. 2014. Cookbook: In Situ Code Completion Using Edit Recipes Learned from Examples. In *ICSE'14: Intl. Conf. Softw. Eng.* Hyderabad, India, 584–587.
- [20] Guoliang Jin, Wei Zhang, Dongdong Deng, Ben Liblit, and Shan Lu. 2012. Automated Concurrency-Bug Fixing. In *OSDI'12: USENIX Symp. Operating Systems Design & Impl.* Hollywood, CA, 221–236.
- [21] Toshihiro Kamiya, Shinji Kusumoto, and Katsuro Inoue. 2002. CCFinder: A Multilingual Token-based Code Clone Detection System for Large Scale Source Code. *IEEE Trans. on Softw. Eng.* 28, 7 (July 2002), 654–670.
- [22] Miryung Kim and David Notkin. 2009. Discovering and Representing Systematic Code Changes. In *ICSE'09: Intl. Conf. Softw. Eng.* Vancouver, Canada, 309–319.
- [23] Patrick Kreutzer, Georg Dotzler, Matthias Ring, Bjoern M. Eskofier, and Michael Philippsen. 2016. Automatic Clustering of Code Changes. In *MSR'16: Conf. Mining Softw. Repositories*. Austin, TX, 61–72.
- [24] Vladimir Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics-Doklady* 10, 8 (1966), 707–710.
- [25] Rashindra Manniesing, Ireneusz Karkowski, and Henk Corporaal. 2000. Automatic SIMD Parallelization of Embedded Applications Based on Pattern Recognition. In *EuroPar'00: Europ. Conf. Parallel Computing*, Vol. 1900. Munich, Germany, 349–356.
- [26] Michael Martin, Benjamin Livshits, and Monica S. Lam. 2005. Finding Application Errors and Security Flaws Using PQL: a Program Query Language. In *OOPSLA'05: Conf. Object-Oriented Progr., Systems, Languages & Appl.* San Diego, CA, 365–383.
- [27] Na Meng, Lisa Hua, Miryung Kim, and Kathryn S. McKinley. 2015. Does Automated Refactoring Obviate Systematic Editing?. In *ICSE'15: Intl. Conf. Softw. Eng.* Florence, Italy, 392–402.
- [28] Na Meng, Miryung Kim, and Kathryn S. McKinley. 2011. Systematic Editing: Generating Program Transformations from an Example. In *PLDI'11: Intl. Conf. Progr. Lang. Design & Impl.* San Jose, CA, 329–342.
- [29] Na Meng, Miryung Kim, and Kathryn S. McKinley. 2013. LASE: Locating and Applying Systematic Edits by Learning from Examples. In *ICSE'13: Intl. Conf. Softw. Eng.* San Francisco, CA, 502–511.
- [30] Robert C. Miller and Brad A. Myers. 2001. Interactive Simultaneous Editing of Multiple Text Regions. In *ATC'13: USENIX Annual Techn. Conf.* Berkeley, CA, 161–174.
- [31] Tung Thanh Nguyen, Hoan Anh Nguyen, Nam H. Pham, Jafar Al-Kofahi, and Tien N. Nguyen. 2010. Recurring Bug Fixes in Object-Oriented Programs. In *ICSE'10: Intl. Conf. Softw. Eng.* Cape Town, South Africa, 315–324.
- [32] Yoann Padioleau, Julia Lawall, René Rydhof Hansen, and Gilles Muller. 2008. Documenting and Automating Collateral Evolutions in Linux Device Drivers. In *Eurosys'08: Europ. Conf. Comp. Sys.* Glasgow, Scotland, UK, 247–260.
- [33] Romain Robbes and Michele Lanza. 2008. Example-Based Program Transformation. In *MoDELS'08: Intl. Conf. Model Driven Eng. Languages and Systems*. Toulouse, France, 174–188.
- [34] Martin P. Robillard, Walid Maalej, Robert J. Walker, and Thomas Zimmermann (Eds.). 2014. *Recommendation Systems in Software Engineering*. Springer-Verlag, Heidelberg, Germany.
- [35] Reudismam Rolim, Gustavo Soares, Loris DăĂŽantoni, Oleksandr Polozov, Sumit Gulwani, Rohit Gheyi, Ryo Suzuki, and Bjorn Hartmann. 2017. Learning Syntactic Program Transformations from Example. In *ICSE'17: Intl. Conf. Softw. Eng.* Buenos Aires, Argentina, 404–415.
- [36] Gustavo Santos, Klérison Paixão, Nicolas Anquetil, Anne Etien, Marcelo Maia, and Stéphane Ducasse. 2017. Recommending Source Code Locations for System Specific Transformations. In *SANER'17: Intl. Conf. Softw. Analysis, Evolution and Reengineering*. Klagenfurt, Austria, 160–170.
- [37] Ross Tate, Michael Stepp, and Sorin Lerner. 2010. Generating Compiler Optimizations from Proofs. In *POPL'10: Intl. Symp. Principles of Progr. Languages*. Madrid, Spain, 389–402.
- [38] Elco Visser. 2001. Stratego: A Language for Program Transformation based on Rewriting Strategies. System Description of Stratego 0.5. In *RTA'01: Rewriting Techniques and Applications (Lect. Notes Comp. Science)*, Vol. 2051. Utrecht, The Netherlands, 357–361.
- [39] Xiaoyin Wang, David Lo, Jiefeng Cheng, Lu Zhang, Hong Mei, and Jeffrey Xu Yu. 2010. Matching Dependence-Related Queries in the System Dependence Graph. In *ASE'10: Intl. Conf. Automated Softw. Eng.* Antwerp, Belgium, 457–466.
- [40] Yi Wei, Yu Pei, Carlo A. Furia, Lucas S. Silva, Stefan Buchholz, Bertrand Meyer, and Andreas Zeller. 2010. Automated Fixing of Programs with Contracts. In *ISSTA'10: Intl. Symp. Softw. Testing & Anal.* Trento, Italy, 61–72.
- [41] Westley Weimer, ThanhVu Nguyen, Claire Le Goues, and Stephanie Forrest. 2009. Automatically Finding Patches Using Genetic Programming. In *ICSE'09: Intl. Conf. Softw. Eng.* Washington, DC, 364–374.
- [42] Cheng Zhang, Juyuan Yang, Yi Zhang, Jing Fan, Xin Zhang, Jianjun Zhao, and Peizhao Ou. 2012. Automatic Parameter Recommendation for Practical API Usage. In *ICSE'12: Intl. Conf. Softw. Eng.* Piscataway, NJ, 826–836.
- [43] Tianyi Zhang, Myoungkyu Song, Joseph Pinedo, and Miryung Kim. 2015. Interactive Code Review for Systematic Changes. In *ICSE'15: Intl. Conf. Softw. Eng. - Volume 1*. Florence, Italy, 111–122.
- [44] Hao Zhong, Tao Xie, Lu Zhang, Jian Pei, and Hong Mei. 2009. MAPO: Mining and Recommending API Usage Patterns. In *ECOOP'09: Europ. Conf. Object-Oriented Progr.* Genova, Italy, 318–343.