

Support to development of complex software systems with  
API function call recommendations

# Contents

1	Introduction	3
2	Related Works	3
2.1	MAPO . . . . .	4
2.2	UP-Miner . . . . .	4
2.3	CLAMS . . . . .	5
2.4	APIrec . . . . .	7
2.5	Buse-Weimer algorithm . . . . .	8
2.6	APIMiner . . . . .	9
2.7	JSS approach . . . . .	10
2.8	Jira extension . . . . .	11
2.9	CodeBroker . . . . .	12
2.10	PAM . . . . .	13
3	Problem statement	14
3.1	Concept of recommendation . . . . .	14
3.2	Concept of API . . . . .	16
3.3	Concept of Pattern . . . . .	16
4	Code clone with Simian tool	16
4.1	About code cloning . . . . .	16
4.2	Simian overview . . . . .	18
5	Proposed approach	21
5.1	Overview . . . . .	21
5.2	Preprocessing . . . . .	21
5.3	Simian within Eclipse platform . . . . .	22
5.4	CLAMS adaptation . . . . .	23
5.5	API recommendations . . . . .	24
6	Validation	25
6.1	Evaluation framework . . . . .	25
6.2	Comparing results . . . . .	25
6.3	Perfomances . . . . .	25
7	Conclusion	25
	References	26
8	References	26

# 1 Introduction

In recent years, the problem of API recommendations rise more and more in relevance in the complex software systems. When a developer writes some code to implement a specific features, he should want some suggestions about useful functions in this context. An API (Application programming interface) is a set of procedures, protocols and objects that gives to the developers the necessary building blocks to implement a specific functionality in easy and understandable way. The API depends on the language that the developer use and it usually composed by libraries with objects and methods useful for the task. The definition is very general and it may change based on the context of the developed project; so, the developer may get confusing about what kind of method or class to use or how to use them in a proper way.

The issues is how to perform a good enough recommendation in this context, balancing possible bias and put the proper hints for the developer. Moreover, the form of the recommendation is also important because in general there are variety of possible suggestion such as code snippet, patterns for the methods, enhance documentation and all things that make a recommendation really usable for the current project.

# 2 Related Works

Considering the state of the art, there are several approaches that face the problem described in previous section. The key point is to find, on one side, a simple and effective way to extract API functions call and build the recommendation system and, on other hand, is necessary to use formal and proper algorithm and techniques that produce very suitable results. In the literature, there are main approaches is first to analyze the code that you are interested in, adopting some similarity metric, optionally followed by clustering techniques and then how to represent the obtained results in form of recommendation that should be clear and effective. In next pages, we present a set of approaches that follow the same main skeleton to solve the problem; what is different among them is the techniques, algorithms and produced results but they are all useful and give the flavour of the API call function problem. Table 1 gives an initial overview on the approaches, with the proposing tool, how they analyze the source code, the similarity metric and algorithm, the eventually clustering technique, the dataset considered to validate the results and the given output. Next section describe 9 different approaches to face this problem and the related framework, challenges and the enhancement given by them to the literature.

Table 1: Summary of relevant paper

Proposed tool	Parser for AST	Similarity	Clustering	Dataset	Output
MAPO	JDT	Class name, Method name API call sequences	Data-driven	Java projects	Eclipse plugin that shows possible API pattern
UP-Miner	Roselyn AST parser	SeqSim technique similar sequences	Two clustering on frequent sequences with BIDE	C#	Probabilistic graph of API that can be queried
CLAMS	JDT parser	Distance matrix	LCS, HDBSCAN techniques AP-TED for top rank snippet	6 popular Java libraries	Patterns for API
APIRec extension	GumTree AST parser	association-based model with fine-grained code changes	association-based change inference model	50 Java project randomly from Github	Most frequent API call
Buse-Weimer algorithm	Symbolic execution for path enumeration	Distance matrix	K-medoids algorithm	Java SDK	Human readable documentation
APIMiner extension	see APIMiner	Structural similarity among API	FP-growth with WEKA tool	Android projects	Enhance documentation
JSS approach	No info	Object usages social network with co-existence relation	Co-existence relation with Modularity index method call similarity with Gamma index	Android project	API usage pattern
JIRA extension	JDT	Cosine similarity history and descriptor recommender	Integrator component based on previous weights	Apache projects	Top ranked methods
CodeBroker	Back-end search engine	Latent semantic analysis	Discourse model user model	Java core libraries	Three layerd information relevant tasks, signature and JavaDoc
PAM	JDT	Structured EM algorithm	Probabilistic model with inference and training set	Java core libraries	List of API patterns

## 2.1 MAPO

In [1] the authors propose MAPO that perform methods extraction using data mining techniques, clustering them to have a more representative data and build a recommender through GEF tool. So, they divided the process into three main steps: analyzing the source code, the API miner and API recommender. First of all, MAPO parse the source code coming from Java Github projects that use Eclipse Graphical Editor Framework (GEF), using JDT parser utilities that allow to analyze in deep a Java file and extract classes, interfaces, method invocations and method declarations. For the recommendations, MAPO considers as API method suitable for the recommender only those belong to third-parts libraries, considering the GEF framework as external, class cast and creation associated to external class and the method call that belongs to external class; basically, the authors ignore the libraries and so the internal API of the JDK. Then MAPO collects all source code by using @ as initial mark to identify a method call and # as separator between method and related class. About the conditional statements, such as if, while and so on, the authors not considered the possible relations among multiple conditions and in practice the represent them as flat code in which MAPO considers all branches. As claim by authors, this simplification is necessary otherwise the mining phase is infeasible and the API recommender is not effected by them. Once the method calls are selected, there is the problem of method overweight and common-path overweight that involves respectively several method calls and sequence in common in the source code that can introduce bias in the result; for example, in fragment of code MAPO can select a method call only because it is replicated and not for its real effectiveness. To avoid this problem, the authors select the longest sequence in common that covers also the smaller sequence of method call and, in this way, the reduce the bias. Moreover, MAPO selects only third-parties libraries using inline methodology, that consist to explore the parse tree of classes and identify the ancestor and so excluding the JDK methods.

Once MAPO have all method call, the next step is mining the sequence to produce recommendation. To do this, is not enough to consider all method sequences because this can bring incorrect results. So, MAPO using first similarity metric, based on the name and the usage of methods to identify similar sequence and then clustered them by using data-driven hierarchical clustering. In particular, the authors consider three level of similarity given by class names, method names and called API method and obtain in this way the similar sequence. Then, by using matlab tool, MAPO using classical hierarchical clustering to obtain the ranked list of representative sequences, transform them into transaction and put them into a database. Finally, the API recommender is a Eclipse plugin that allow developer to click on the method of interest, execute a query on DB to show possible uses by using sample code and developer can also see details on the proper window tab in which the API method are highlight. To validate this approach, the authors use a dataset composed by 20 projects that used Eclipse GEF, run the tool and show the effectiveness of their approach by a quantitatively comparison. To reinforce the validity, they also conduct an empirical study on by considering a set of tasks to do and select a group of Java developers that is involved to solve those tasks.

## 2.2 UP-Miner

A similar approach is performed by [?] in which the authors create UP-Miner tool that improve MAPO in term of accuracy. Going in deep, the aim of the authors is to achieve the succinctness but also the effectiveness of mined pattern that represent an enhancement of previous approach. Once the authors define API usage mining, that is the optimal number of patterns under a given threshold, they propose a clustering technique based on BIDE algorithm. As first step, UP-Miner extracts API pattern sequences using Roslyn AST parser from the projects that compose the dataset. Then, the apply the SeqSim n-gram technique that takes two sequence computes the

similarity that is based on the shared items between them in term of objects and classes used and on the longer and consecutive sub-sequences rather than the shorter ones. This phase produces weighted results that are used for clustering that are conservative on, in sense that the maximum distance between two cluster is the maximum distance between two elements of those cluster. Then, UP-Miner use BIDE algorithm, that have the key concept of frequent sequence: a general sequence became frequent if its super-sequences (namely the sequences that contains the considered sequence) is greater or less of the given threshold. Consider this, BIDE algorithm can extract the longest common sequence that are useful to divided the results into different clusters, but it is not sufficient because at this point there may be some redundant cluster. So, it is necessary to apply once again the previous phase considering the cluster as usage pattern and this two-step clustering grants an improvement in term of redundancy considering also two different threshold (one for pre-BIDE and another for post-BIDE application). Until this, UP-Miner address the problem of coverage but the aim of the authors is to reach also the succinctness of patterns. To do this, UP-Miner use a dissimilarity metric that measure the diversity of a usage pattern from another and an utility functions to maximize in order to obtain the better results, decreasing the threshold at each step of the algorithm implemented for this task. Once UP-Miner computes the correct and more succinctness as possible usage API pattern, the authors show them to the developer by building a probabilistic graph in which each node is a usage pattern and the edge is weighted with certain probability. To produce this prototype and make some experiments that involves the developers in a similar way as we see in MAPO, the authors use a very large C# dataset as input files and compare their results with MAPO ones.

## 2.3 CLAMS

Go further with the examination of previous, we have also CLAMS tool proposed in [3] that follows a quite similar approach but it differs from the point of view of the results. In fact, CLAMS produce as API recommendations snippet of code that represent a pattern for a certain library. As usual, the preprocessing phase is done by analyze the AST of the source code (in case of CLAMS, projects related to 5 popular Java libraries) with a dept-first search using JDT as previous example. This phase produces snippet of code that bring all information about API implemented in the code. The similarity technique is based on Longest Common Subsequence (LCS) and is more effective rather than an analysis at source code level. By using this technique, CLAMS creates a distance matrix that is used as input by the clustering module of CLAMS, that implements the hierarchical version of DBSCAN algorithm, called HDBSCAN, plus a post-clustering processing to eliminate the sequences that are identical to snippets for each cluster obtained. The aim of HDBSCAN is to isolate the less representative methods and, more in general, points into a distribution that are really far from the rest of the dataset. To do this, the algorithm uses a distance called core distance to draw a circle on the points to exclude and then computes the mutual reachability distance that reduce the presence of sparse node in the dataset. Then, by applying Prim's algorithm, HDBSCAN calculates the minimum spanning tree among the closer nodes and finally sort them by a hierarchical clustering (this phase is not present on the original DBSCAN algorithm), as well explained in the related work. There is also a useful Python library used in CLAMS that provide utility functions to make all this step in a very understandable way. The core module of CLAMS is represented by the snippet generator, that performs six main steps in order to obtain the final snippets that represent patterns. First of all, CLAMS replaces all literals present in the code with their abstract type by using srcML, a tool that produce XML file starting from a source code, and removes all comments. At the end of this step, we have code with the same structure of the original source file but more abstract. Then, CLAMS identifies API call in that code and what are not related to them and creates two lists.

In the next step, the authors identify all variable in the scope of the API sequence. To finish the process, the non-API statements are removed and they put on top of snippet variable declaration related to API, plus of course the snippet of code related to those API. So, the snippet of code that is produced in this way is composed by a sequence of variables related to API class and a possible use of them, with considering also the statements present in the original code (CLAMS retained also the structure of the code). The authors put a comment Do Something in the section of code in which the founded variables may be used.

Once CLAMS has these results, the snippet selector module find for each cluster the most representative snippets by giving them a score. To do this, the authors use another algorithm that works on AST, called AP-TED that creates a distance matrix between two clusters and from this, calculate the similarity. Finally, CLAMS ranks all representative snippet using the definition of snippet support define as follow: a snippet is supported if there is a file that contains a supersequence of it. An example of extracted pattern is depicted below and it is related to Paho, a MQTT client for Eclipse platform.

---

```
{
    final int keepAlive;
    char[] password;
    String message;
    ConnectionManipulationProxyServer proxy;
    String clientId;
    final int[] res;
    final MqttDefaultFilePersistence DATA_STORE;
    String username;
    MqttConnectOptions options = new MqttConnectOptions();
    options.setCleanSession(boolean);
    options.setUserName(username);
    options.setPassword(password);
    options.setKeepAliveInterval(keepAlive);

    MqttClient client = new MqttClient("a string" + proxy.getLocalPort(), clientId,
        DATA_STORE);
    client.setCallback(this);
    proxy.enableProxy();
    client.connect(options);

    for (int i=0; i<0; i++) {
        client.publish(username+"a string"+clientId+"a string", message.getBytes(), 0,
            boolean);
    }
    proxy.disableProxy();

    while (client.isConnected() && res[0] < 0*keepAlive) {
        try {
            client.publish(username+"a string"+clientId+"a string", message.getBytes(),
                0, boolean);
        } catch (MqttException e) {
            // Do something
        }
    }

    Assert.assertFalse("a string", client.isConnected());
}
```

```

    if (client.isConnected()) {
        client.disconnect(0);
    }
    client.close();
}

```

---

Moreover, for human readability reason, CLAMS beautifies snippet using A-style tool, that removes useless spacing and fixes the indentation of the final snippets. For the evaluation task, the authors use 5 popular Java libraries coming from Github projects and they are Apache Camel, Drools, Restlet framework, Twitter4j, Project Wonder and Apache Wicket. In the evaluation section of their paper, the authors taking into account different clustering algorithm and make experiments with HDBSCAN, already mentioned, k-medoids that is similar to the previous one but achieve more coverage but less precision. This algorithm is based on find a central point that has the equal distance from the rest of the point, called medoid. In our case, the medoid is a particular method API call that is the most representative for the library. K-medoid algorithm uses also n-gram based technique and similarity distance matrix as HDBSCAN. To answer to the research question, the authors of CLAMS preform also a comparison with NaiveSum and NaiveNoSum approaches, that are clustering techniques less precise with respect to the HDBSCAN and K-medoids. In fact, the better results are coming out from the latter algorithms. Furthermore, the evaluation is composed by a user survey about the utility and real support given by CLAMS patterns. All this information are available at the CLAMS official site, that contains also the Github project, the original dataset and the instruction to set up the environment to launch CLAMS as standalone platform on Linux machine. As claimed by authors, this work is really interesting and flexible because it is not dependent on a specific program language. The only constrain regarding the srcML tool to produce the XML files related to API patterns, but it is not a big issue from the adaptation point of view, as we can see later.

## 2.4 APIrec

APIrec tool [4], instead, uses statistical techniques in order to keep trace of the context in which a developer writes its own code, by analyzing the co-occurrences and fine-grained code changes. Differently from the previous approaches, the authors keep trace of the context in which the API method call may be useful. As usual, to extract the source code from the 50 Java projects randomly selected from Github, APIrec navigates the AST of the source code using GumTree tool. Before going in deep with the implementation, the authors give several definitions that are useful to understand the behaviour and the contributions to the state of the art given by APIrec. First of all, the term API for the authors indicates both external and internal method calls and APIrec performs recommendations only if the context of API is correct for a certain situation. Regarding the AST, we have the atomic change that represent a new element on AST composed by kind of operation, AST node and label. A collection of atomic changes is called transaction and is stored in a bag, a particular data structure that we commonly find in the AST definitions. An important concept that is a key definition for the APIrec implementation is the code context, represented by code tokens related to the API that the developer is writing. Taking into account the tokens, the authors consider both the distance and the order of this, as an API method calls have a specific call order and they are really effectiveness only if there are called in the proper way. To remark this feature, APIrec gives also a weight based on how near the token by is considering the distance matrix. Once they define this set of metric and concept, the authors show the inference model based on likelihood scores taking into account both change of the context and code.

The entire model is based on the correlation between a code token and another, expressed as

we said in term of atomic changes in the AST and transactions. In particular, APIrec evaluates the probability that certain events, namely the transactions, occurs given another. This score is called association score and it is used to calculate the distance between two changes into the code as well as new methods that will take a part in the final recommendation. At the end of these computations, all possible scores and weights are calculated but it is not enough to perform the recommendations. In fact, it is necessary to apply machine learning techniques in order to train APIrec with the code changes and context. To do this, the authors use hill-climbing adaptive learning filled with three parameters: numbers of co-occurrence founded with fine-grained atomic changes, numbers of co-occurrences related to changes tokens and two weights. With this process, the necessary scores are calculated and APIrec is able to perform the recommendation by considering the most probable method calls for a certain API. The methods are also ranked by highest probability of usage but also distance, scope and dependency are considered in the ranking phase. To support their tool, the authors conducts several experiments over a very large dataset, including also analysis regarding code change context, user empirical studies, evaluation of accuracy and predictions.

## 2.5 Buse-Weimer algorithm

Paper from Buse and Weimer [5] is still about API call but is focused to produce automatically a documentation for the projects in a human readable format. Once they obtain data from mining phase, the proposed algorithm extract API pattern and rearrange them in a more readable and effectiveness format. The focus of this work is on Java documentation that provides useful hints and suggestions when a developer is implementing an API functionality. Although in general the Java doc helps in this kind of activity, it often lacks in something, as the examples are too general or it not able to give a concrete hint for the problem that the developer is trying to solve. So, the aim of the authors is to produce an enhanced version of documentation that is really useful for the developers, starting from retrieve the API patterns, defined as the sequence of function call for a certain API class, called target class. Once this first step is done, the algorithm tries to produce a human-written documentation for the API, taking into account some characteristics such as the lines of code, that are 11 on average but 5 for the median. Moreover, the authors consider also the abstract initialization, abstract usage and exception handling. All this information is extracted using JDK utilities and they are validated by the authors, putting more effort on the human aspect. To validate the results, they involve over 150 developer and collect their answers about the proposed results. By analyzing these statistics, a typical developer wants multiple uses for a certain class or API method and not only one, that may be not useful for his particular context. Another key point is the conciseness of the suggested snippet of code, as well as the readability and variables names that must be related to the context, also including temporary ones to improve the readability. So, from this survey, we can identify four key points to achieve the human-written documentation: size of the code, readability, representativeness and concreteness.

To reach these goals, let look to the algorithm implementation, that the authors divide into four main step, the path enumeration, predicate generation, clustering and finally the output documentation. Starting from the path identification, they scan the code and identify acyclic part of the code that represent a path for the target class. Notice that this approach can led some bias but it happens in practice and the authors choose to stay close with respect to real implementation. The So, they use intensively human users to perform test and to measure the accuracy of their proposed algorithm. As the previous approaches, they parse the code and cluster the significant result with a distance matrix in order to build the related documentation. As the purpose is different from previous works, they are adding some clustering also on predicate



and represent the abstract pattern as a graph. Then they compute symbolic execution over these paths in order to produce inter-procedural path predicates that are logical formulas used to represent the code and in particular, if a certain statement is reachable or not. From this abstraction, the algorithm computes use seeds that are local instantiations of fields, objects and whatever is related to the target class. Finally, from these it produces concrete uses, the real hints about the target class, that are stored in graph form in which edges keep trace about what happens before and so, in this way, we have the context as well as the chain of methods necessary to implement the features related to the target class. To have better results, however, it is not enough to produce usage examples, as we seen so far with the other related works. Clustering is necessary to avoid heavy computation, and the proposed algorithm exploits the well-known k-medoids algorithm with some modifications, as the original algorithm is not suitable to detect distance about objects. The distance matrix that is taking into account by the k-medoids algorithm is based on the happens before relation obtained from the graph. So, at the end of this computation, the algorithm obtains the cluster and summarize them into abstract uses, represented once again in graph form. The final step of the proposed algorithm is to produce the documentation. Starting from the abstract uses, it uses a topological approach to avoid the cycle and branches that appear in the code, as the final recommended documentation about the target class must be a flat file. Using this approach, the authors are able to retrieve a Java documentation related to the target class and respect also the Java syntax, so they avoid malformed documentation. Moreover, with this approach, the algorithm handles the exception treatment with try catch clauses, that are put always in the correct order. As said before, the focus of this approach is on readability of the recommendation from human's perspective, so the authors set up a very big evaluation framework composed by 47 SDK classes and dataset, the eXaDOC tool as concurrent approach and over 150 people as tester. The threats rise up from this evaluation are related to the validity of dataset (it may be not indicative and it doesn't represent all possible situation) and the background of the developers chosen for the evaluation (not expert in the field). However, this approach is useful to understand the concept of readability that is very useful for my approach.

## 2.6 APIMiner

In [6], the authors extend APIMiner tool with API pattern considering Android projects from Github as dataset. In particular, they develop a module that perform the recommendation based on mining. The original API miner tool [apiminer] retrieves information about the documented API method in the Android interface similar to JavaDoc view. This kind of recommendation coming from concrete source code extracted from private repositories. The overall architecture is composed by the source repositories, a pre-processing module based on slicing algorithm, the ranker module that classifies the summarized methods considering all the lines of code (source code metric), the number of commits in the original repositories (called process metric) and the number of download (called usage metric). Moreover, the authors use also Java Weaver, a tool that builds and retrieves automatically the Java documentation related to the extracted methods calls. About the slicing algorithm, that represent the core of the system, it works as follows: first, it takes as inputs the API method call that the developer wants to analyze and all the body statement in which this method was found. Of course, this kind of analysis is performed off-line to avoid loss of time. At each iteration, the algorithm looks for similarity by analysing the list of variables present in the body statement of each method. If it finds some similarities, it put the method in a list that represent the final recommendation, as in this list there are the most relevant methods. The slicing is performed both backward and forward; the first looks the writing variables while the second analyze the the reading ones. The dataset is composed only

by Android projects because this system provides an API to validate APIMiner approach. All the projects are under open source license and are compilable, otherwise the slicing algorithm doesn't work (it visits the AST to retrieve information about method calls).

Concerning instead the extension mentioned before, the authors perform the API extraction by considering FP-Growth association as main index of similarity and run it using Weka tool that consider also the relation between two API call, defined as sequence of methods that implements specific functions. This approach relies completely on Weka implementation. During the mining, the tool discharges the call with single call because is not relevant. The results are evaluated by define two main metrics: support, defined as the number of patterns that include the method, and confidence, the probability of the method in the antecedent transaction. The final recommendation is displayed in the JavaDoc window in Eclipse and Android Studio, although the results is related to Android API functions. With respect to original APIMiner work, the authors also extend the graphic interface in which the recommendations are showed; in particular, the tool shows the complete chain of method calls related to the client method. An important drawback of this approach is that it works only for Android projects and it is no tested at the moment for other kind of APIs.

## 2.7 JSS approach

Then, we have a look on [7], that perform clustering by using graph format. The authors, starting from a graph representation of the so called object usage, build a social network based on the co-existing relation among nodes. The aim of this work is to cover the less frequent API pattern in Android context. Before going in deep to the proposed implementation, the authors point out some definitions that are used in the approach. First of all, take a general code snippet, an object usage is a list of methods that belongs to an API class that are used in the part of code that we analyze; in general, a fragment of code can contain more objects usage and this feature is represented by a co-existence relation among objects usage. So, one object usage became a node in the graph and, if there is a co-existence relation, we put an edge between them and the weight are the number of co-occurrences. An usage pattern, instead, is the sequence of object usages that belong to different API class. Once they define this to key concept, they underline the challenges underlying the API mining and propose an approach quite different that we have seen so far. In facts, they represent the object usage, and more in general usage patterns, with a graph as we said; then, they define the co-existence relation and method call similarity, that are the baseline to define a similarity score among API call. Regarding the co-existence relation, it is represented as a weight in the graph and represent the number of occurrences of the object usage in an API class; basically, objects usage that are included in a particular API class are connected by a co-existence relation. As the authors want to reach quite good coverage of API call and avoid the redundancy, they need some clustering technique, as we know from the other related works.

To do this, they exploit the previous definitions and propose two level of clustering, one related to co-existence relation and the other one based on method calls similarity. For the first level, they propose a modularity index applied to community structures, defined as subnets of node densely connected. So, exploiting the graph format, they apply a greedy algorithm that calculates time by time the modularity, with the function goal that try to achieve the maximum one. By running this algorithm, they get the optimum cluster and perform the first level of clustering. For the second level, they focus on the method call similarity and propose the Gamma index, based on consistent and inconsistent comparison. A comparison is consistent if the distance of two object usage that belongs to different clusters is smaller than another pair that belong to the same cluster. By applying these two techniques, they obtain a good coverage of usage patterns and

avoid the redundancy by using abundance metric that describes how many times an object usage appears in the corpus. Through this metric, the authors retrieve the most popular object usage but it is not enough to perform the recommendations. The last step, in facts, is to map the objects usage into usage patterns in form of real code snippet that support the developer during the API implementation. For the testing phase, the authors provide a large corpus of 11,520 Android projects coming from Github, focusing on the Android Application Package (APK) because it contains the compiled code and it is not possible to analyze the Google Play original source code. The next step is the setup of a golden set that is a set of queries suitable to the author's purpose. This set is extracted in order to reach the high coverage as possible. The tool is fed with a single query and following the described process, the authors retrieve the expected pattern for that query. To validate the overall process, they also set up an user evaluation with questions about the readability and understandability of the suggested code.

## 2.8 Jira extension

In [8], the authors propose a tool for mining API and make recommendation within the JIRA platform, that is an issue management project based on summary, description and component related to a particular issue. The main idea of this work is to analyze the pre-change and post changed files and, in this way, find recommendations starting from a textual description of the input. The first step is the preprocessing of the input, necessary to clean the code and to give it a proper representation for the algorithm used in next steps. This textual preprocessing is done taking into account two issue: tokenization and stemming. The first one involves the process of break into smaller piece of code the entire document using delimiters as frontier and put it in a word token structures (also called bag of words). The stemming, instead, is related to the root of the word and transform it in stem word: in this way, the authors summarize multiple words to avoid bias during the analysis. Once this preprocessing is done, the algorithm uses a term frequency indicator to count the number of times that a word appears in the document and so, obtain the most popular token. A similar measure is calculated also for the document and after a formula showed in the paper, the authors retrieve weights and put them in a vector; in this way, each bag of word is associated to a weight that measure its relevance for the recommendation. The framework is composed by three main part: history based recommender, descriptor based recommender and the integrator that put all together. For the history recommender, the algorithm compares the indicator on the Jira platform using similarity distance matrix, starting from Jira fields, named This module consider summary and description as key value of comparison and store them in its knowledge base called Historical Feature Request Database. Once the similarity scores are obtained, the algorithm perform aggregation of these scores and perform the final comparison between the historical scores (calculated in this step) and the new feature request that is coming. To do this, they create a top-k request looking at the history and choose the recommendation with the highest value among them. Description based component, instead, compares the new feature request with the Javadoc of the method, to have a more detailed recommendation. Here there is a preprocessing phase regarding the API doc which consists in extract method call taking into account the @param and @return annotation plus the discharge of HTML tags and Java comments. As similarity measure, in this phase they use cosine similarity between the current feature and the preprocessed API. The last component is the integration, that merge the historical part with the description part, apply Gibbs sampling and try to calculate the best results at each iteration. In particular, they pick first the no-zero historical recommendations and then compare them with the results of descriptor module; from these results, then algorithm creates a top rank recommendation related to the developer's features. Regarding the evaluation of these results, the authors select 5 most famous Apache projects (Hadoop, CXF, AxisJava,

Hbase and Struts 2) and looking for Github projects that implements these libraries. They filter these projects considering the presence or not of the pom.xml file and, after this preprocessing, they retrieve 207 projects as corpus of the tool. To select the golden set, the authors also considering the status of the file that belongs to Jira platform: in particular, they take into account is the new files are added or are changed with respect to original while they not include the deleted files to the golden set.

## 2.9 CodeBroker

Now we analyze CodeBroker tool, proposed in [9] that use information retrieval techniques in order to make recommendations. The authors consider a lot of techniques for their tool, such as information delivery, retrieval by reformulation, knowledge augmentation and finding task with similarity metric. All these definitions compose the conceptual framework that is a baseline for the implementation and evaluation of the CodeBroker tool. It is a interface agent with back-end utilities that takes a query as input and return the component related to it. Notice that this recommendation is based on Javadoc generate from Java source files. The tool is based on two communication channels with the developer that is interested in API recommendation: one is an implicit where the system autonomously retrieve methods information and details from a given query. In this case it shows information organized in three layers, namely task relevance, signature details and full JavaDoc. Regarding the second channel, it is explicit because the developer can refine the query based on its current needs and the system can adapt itself to this new situation and this technique is called retrieval by reformulation. Furthermore, CodeBroker creates a discourse model to represent the projects and an user one to represent in some way the software knowledge and personal information regarding used method in the project. It uses LSA as similarity techniques to do comparison and Java core libraries as dataset to test the tool. About the implicit communication, this term describes a set of information that can be inferred by the system without taking in account the user's hints; for the explicit channel communication, instead, the tool considers the user's need by looking its model plus the discourse one that I mentioned before.

As first step, CodeBroker arrange the query taking into account the context of the developer, called constrain part, the program that is the concept of functionality and the code that is the embodiment of it. So, for the similarity analysis, the authors consider the context and so the conceptual similarity and also the constrain similarity, that involved between two different signatures of the methods. About this last concept, they reuse it to apply the Latent similarity analysis (LSA) as main technique to perform comparison between two different API methods. Going in deep, the tool performs the so called signature matching that outlines the similarity of two components based on their signature structure. Although this comparison should be not representative enough, the authors claim that is suitable for their purpose: so, the value of the comparison is in the range from 0.0 to 1.0, that represent the exactly match between two different method. However, this first analysis must be enhanced by retrieval by reformulation technique as the LSA cannot analysis in deep fragment with comments or task relevant information from the code. So, as mentioned before, the authors use explicit communication channel to allow the developer to formulate once again the initial query: the typical use case is that a user want to improve the initial query with other components and so he change the query in order to retrieve more information or very different components with respect to the initial one. It is true especially for the Github repository, that have very complex structure inside them and may a no expert developer want to know all these details. Now it is time to introduce the concept of module, namely the part of the code that implements the developer's main feature. This kind of activity is performed by building the discourse model of the developer, that represent the

sequence of tasks necessary to implement the all features and it is used to improve the final components recommendation. At the beginning, this model is empty as the developer is starting to develop and he doesn't know a priori what are the components that are useful for his task. During the query phase, this model is filled with respect to the developer's choice and all this information are retrieve on the RCI console used for the final recommendation. There is another model that CodeBroker takes into account during its analysis: the user model, that represent the developer's knowledge in abstract form. Based on this definition, it is very different with respect to the discourse model and it is partially filled based on the developer skills. This model is used to remove possible components that the user already known and so to avoid the redundancy problem. The authors define the knowledge as the number of implemented class by the developer and, in general, it is different from user to another. The final recommendations is performed through RCI-display already mentioned, with three layers: the first one shows the components related to the query, the second is linked to mouse movement (it displays signature information about the retrieved components) and, finally, a completed description in HTML external page, included the JavaDoc, belongs to the third layer recommendation. For the evaluation task, the authors use Java 1.1.8 core libraries and JGL library, with 663 classes and about 7000 methods to analyze.

## 2.10 PAM

Finally, we look at the [ref] that propose an approach similar to the others, but with the aim to reduce the redundancy of the mined patterns. The proposed tool, called PAM, is based on probabilistic techniques and wants to cover not the most common pattern but the most significant, which are also the most difficult to retrieve. To do this, the authors avoid the n-gram technique that we saw previously and use the generation of a sequence by interleaving a group of sequences. A very important issue is the threshold, because if it is too low there are a lot of results but if it is too high there is no useful results. The core of the approach is represented by mining sequence of patterns from a given project. This feature is realized by applying a best-effort approach to extract the pattern directly from the source code. It follow the same approach show in MAPO by visiting the AST but it not consider conditional statement such as if else structure. At the end of this process, the authors the list of API call in form of method invocation considering their qualified name. Notice that PAM represents an improvement of the original MAPO approach because it is able to dynamically inferred the call sequence. When this extraction phase is finished, we look at the probabilistic model used by the authors for retrieve the most probable and useful API call in form of patterns. The model is based on a generative algorithm that takes as input the API call patterns and generates the interesting patterns for each of them. For interesting, the authors mean pattern that introduce some kind of novelty in the considered original patterns. The probability distribution applied at this point is explicitly defined by the probability to extract a certain pattern considering the sequence until now. However, to extract the more interesting patterns from the original patterns, the generative algorithm is not enough and the authors introduce the inference. Normally, the introduction of the inference lead to NP-hard problem but the authors use in this case a greedy algorithm that approximates the problem using conditional probability. This algorithm maximize at each step the probability to choose more interesting patterns starting from the original sequence by using a parameter.

All this concept are used in the main algorithm used by PAM, called structured EM algorithm. This algorithm requires some form of training data, represented by client methods and the associated probability to determinates the more interesting related API patterns. All steps are independent so the EM algorithm calls the previous algorithms in a parallel way. At the end,

the authors have as results the list of the API most significant patterns with the associated probability. The dataset used is the same of CLAMS paper. PAM is implemented as Maven project available on Github. As input, it takes a file in arff format (the same format used also in CLAMS) that represent the client API methods. We can parametrize the execution with different arff or by setting the maximum number of iteration or structured steps.

### 3 Problem statement

So far, we have seen several approaches to perform recommendations related to API context. The proposed tool from the literature are very different from the point of view of dataset, similarity measure and clustering techniques but they share a conceptual model that starts from the abstract syntax tree of the code to the final recommendations. This procedure is depicted in Figure 1, that describe at high level the main steps to reach the final output. Many of these are optionally but, as we can see from the related works section, are strongly recommended to perform a better analysis and to have good enough results. Although this approach is very common, there are differences in definitions, methodologies and techniques that make an approach very different from another, starting from the basic concepts of recommendation, API and pattern. These key concepts, that compose the core of this work, are often ambiguous in the literature and change meaning based on the considered context. For this reason, before to describe my approach to face the issue, I define the building block of it, looking at the common meaning for each of them as well as the definition taking into account in my development context.

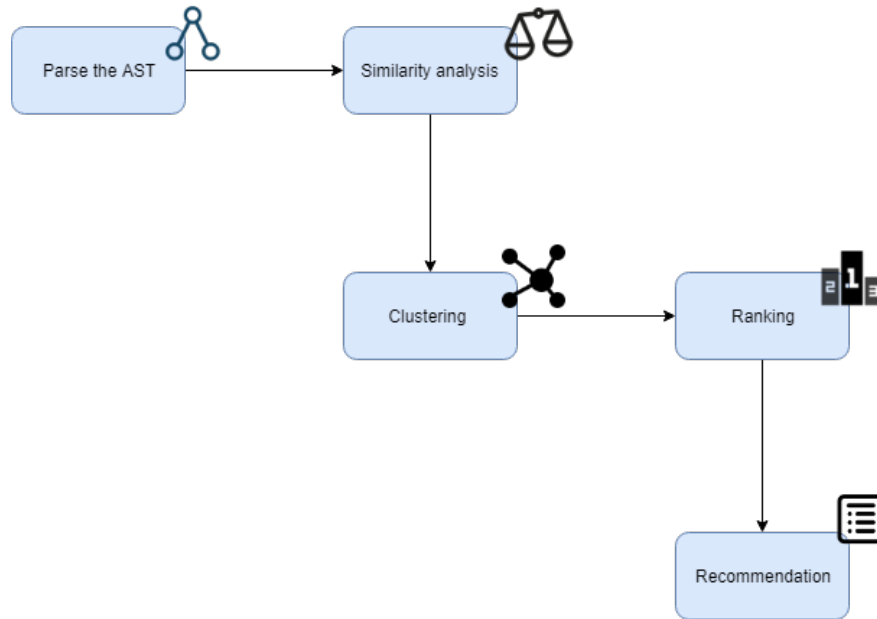


Figure 1: The general approach for recommendation

#### 3.1 Concept of recommendation

First of all, we have to look at definition of recommendation. In general, a recommendation is a suggestion or a proposal as to the best course of action, often provided by an authoritative

body or expert domain. The key concept when we talk about recommendation is the context, because it changes meaning with respect to type of system that a developer is working in. For example, we can find recommendation system in software engineering (RSSE) that gives the necessary information about suitable items for a software engineering task in a specific context. When a developer starting to work on a project, there are several information spaces, also called landscape, that describe all involved information in the overall developing process. In [rsse], the authors define all possible information spaces in software engineering but this definition can be applied in whatever software complex system. Related to the developed project, the main definition is the project source code and project history, that give the context of developing. The project source code is useful to understand the structure of it, especially when we looking for method declarations, method calls and possible most frequent pattern while the project history tells something about the changing happened in the code from a older version from another. This changes are usually captured by a VCS (version control system) although this kind of system isn't easily browsable and techniques like data mining or machine learning are required (as we saw from the related works section). Development environment also belongs to this information spaces classification and it includes all scripts, commands and tools used to test and run the system. Another big information space is defined by API that are linked to the project as well as their documentation, a good starting point to better understand their behaviours. Moreover, the authors consider also two types of traces when a user is developing a project: interaction traces, composed by the list of user's actions like search on website for particular component or interface and normally this kind of information are captured by an IDE (like Eclipse); the execution trace, instead, regarding the software runtime execution and the collected information are in general function calls and results of computation at every time. Finally, a very important information space is represented by the web that becomes more and more relevant in recent years; in particular, Stack Overflow questions and answers site is the most used by the developers as we can find a lot of concrete example about code snippets as well as information about interfaces, technologies and tools.

As we can see, all this information spaces represent an huge data mine and there are many problems to extract the correct information from it. First of all, the available information are heterogeneous and context awareness while a typical user is looking for a quickly solution related to its specific context. Furthermore, the complex software systems are rapidly growing and it can lead the overload information problem. So, looking at this problem, it is necessary to find a proper way to do recommendation, in the manner that the developer find a good solution in very few time without waste it to search in very large context. For the software engineering context, the authors propose RSSE, a recommendation system for software engineering based on capturing the context, giving the proper recommendation. The process involves some kind of data preprocessing, capturing the context, select the correct recommendation using collaborative filtering techniques and show them to the user.

Moving in the context of the IDEs, auto-completion can be a kind of recommendation at runtime because the developer, usually through shortcuts, wants suggestion for specific functions and methods. This technique uses often documentation embedded in the IDE (like JavaDoc in Eclipse) and it adapts itself to the context, in this case the imported libraries in the project. It is based on the word prediction when a user is typing something, such as fragment of code or method declaration. There are many techniques to do this, first of all n-gram technique, probabilistic methodologies and heuristic learning algorithm.

## 3.2 Concept of API

Moving on the concept of API, we have also in this case huge amount of definition in the literature and it's necessary to specify very well the context in we are. In general, an API (application programming interface) is defined as a set of procedures, protocols and objects that gives to the developers the necessary building blocks to implement a specific functionality in easy and understandable way. Based on the context, these building blocks can be classes, interfaces and methods properly declared and used or intermediate software that acts as middleware in different situations as well as in the hardware context. The concept of API is strongly related to libraries that a developer uses and the kind of application that he is developing. So, we can have remote API to interact with different resources like databases deployed anywhere through protocol. An example in Java context is Java Remote Method Protocol. The most famous and used are the web APIs, usually defined in a framework context like MVC or inversion of control. While in the past a web API is associated simply to a web services, nowadays they moving towards SOAP (Simple object access protocol) and REST protocols, as the developers need an easy and quickly usage of the APIs. The basic principle in API development is the information hiding that consists in hide the design principle behind the development of a tool or framework. In this way, when the software needs an upgrade on the functionalities, all the process is done behind the scenes and we will reach a stable interface.

In this work, we focus on the Java programming environment and in particular on the external API that are not included in the JDK. As we will see, the dataset is composed by at least the same libraries that we analyze for CLAMS tool because the output patters are reused for the final recommendations. At this point, we can define the last concept to better understand the implementation: the concept of pattern.

## 3.3 Concept of Pattern

In the same way of API, it is not easy to define what is a pattern. In a very general way, a pattern is the formalization of a problem/solution pair, used to make decision. In software engineering, we talk about of design patterns, that are widely used in the community with the purpose on reuse the knowledge in different situation like describe system that shares most of functionalities with another. In this case, we can reuse the concepts provide by a particular pattern (for example the classical Model view controller in web application domain) to model a particular situation without waste time in finding new solutions. Moving to our context, a pattern at level of code is the chain of method invocations necessary to use in the correct manner an API. For example, the library Twitter4j in Java requires first the object TwitterFactory in order to call other methods useful to implement a certain functionality of the library. The identification of a particular pattern is called pattern recognition and this techniques usually involves the measurement of the object to identify distinguishing attributes, the application of a similarity measure considering different features and comparison with known patterns to determine a match or mismatch.

# 4 Code clone with Simian tool

## 4.1 About code cloning

To support the recommendation, we choose an approach that involves code cloning analysis. This kind of techniques is very spread in the computer science world and we look at [codecloning1] to have a brief but clear overview on this topic. In general, a clone detector try to find the similarities between two fragment of source code. These analysis depend first of all from the



level of details that the tool wants to reach: to make a very simple example, each code cloner could be set a different similarity function in order to set the level of cloning. They differs also in term of the comparison of two fragment of code, such as AST, textual comparison and so on. As we can see, there are a lot of concepts and techniques in this approach and to avoid get confusing, the authors create a very useful taxonomy to classify the activity of code cloning, showed in table 2.

Table 2: Code cloner tools taxonomy

Code cloner type	Level of similarity
Type-1	The code fragments differs only from the whitespaces, comments ad layout
Type-2	Two code fragments that are syntactically equals except for the same conditions of type-1 plus identifiers, literals and name variables
Type-3	This kind of clone detector looks for variation (add, delete or change) in statement that appears in the fragments, plus the previous conditions
Type-4	We have this kind of cloner when the computation that the fragments perform are equal without considering the syntactic implementation

Although there are a huge number of tools and techniques, there is a common clone detection process that it has to be considered in order to avoid very critical loss in time and spaces. In fact, even using whatever tool the computation became a big issue if the common part among fragments are unknown at the beginning of the process. So, the authors identify an overall process to approach the code cloning activity, even all the steps are not required depending on the situations. As we said, a preprocessing phase is necessary to discard useless elements in the fragments of code like embedded code that appears in some language and to obtain the source units. These units can be very different depending on the purpose of the cloner and sometime they can be partitioned again in comparison units, depending on the structure of the original source unit (the common case is when we have an if-else structure in which the comparison units could be the different branches). After the preprocessing, if the code cloner go further the textual analysis, a transformation phase is required, to bring all the fragments to a common representation. Among the possible normalizations that we can apply to code fragments, we have the removal of whitespaces and comments, the normalization of identifiers (for example, through order sensitive index scheme), pretty-printing that affect the layout and structural transformation (for example, by removing the modifiers in a particular language). When we have a comparable units, a different comparison algorithm is run depending on the tool in order to obtain the list of matches. In this phase, we have to distinguish the fixed-granularity tools, in which the units that belongs to same block have the same granularity from the free-granularity ones in which the aggregation continue until a threshold value is reached. The list of candidates for the comparison are usually source coordinates that must be map on the original source code files. The last step is a post-processing in which clones are ranked or filtered depending on the aim that the tool wants to reach. This phase can be done by human evaluator or through a parametric heuristic algorithm.

Looking now to the different tools, most of them relies on different approaches to identify cloned code. The most immediate approach is the textual one, as the transformation and normalization phases are often very slight. The tools that implement this approach uses fingerprints or substring of the source code. The fixed lines that are used in the comparison are called window and they are hashed. To obtain fragments with different lengths, the tool apply simply a slicing on the window. The lexical approach, instead, works on the tokens obtained from the source code through the compiler-style lexical analysis. This technique is more robust because it avoid the whitespaces and other dirty code that we want to exclude from the comparison. The big issue of this approach

is that it not consider the syntax; so, the founded clones may overlap different syntax units but preprocessing or postprocessing can avoid this situation, like pretty-printing techniques to format the code in a better way. Go further, we now look at the syntactic approaches, that usually rely on the AST of the code. There are two main process that we can apply: the tree matches and structural metrics. The first relies only on the AST extracted from the code and the comparison takes place on the subtrees. Each element of the source code(variables, literals) became a leaf of the tree and subtrees that are hashed into buckets in order to reduce the number of comparisons that take place in each bucket. However, the complexity of this approach is very high and recently there are code cloner that try to mitigate this drawback by serialize the AST as node sequence in order to reach the same speed as in the token based techniques. The second approach the exploit the AST is based on structural metrics. This technique avoid the direct comparison between ASTs by collecting a vector of metrics, usually calculated through fingerprints functions that consider classes, methods and statements for the metrics. The last approach that we look is the semantic technique that relies on static program analysis to provide more information rather than the syntactic one. With respect to the other approaches, the source code is represented by a PDG (Program Dependencies Graph) to keep trace the data dependencies among expression and statements. So in this case, the comparison of the clones turns to the problem of finding isomorphic subgraphs. Finally, in the literature we can find hybrid approaches that involves both syntactic and semantic analysis.

This work includes also a very useful tool comparison, in which the authors show a list of code cloners and their main features. This state of the art is done by taking into account several parameters and metrics, like availability of the tool, IDE integration, comparison algorithm, kind of granularity, pre or post processing, language support, subsystem, possible empirical validation and overall complexity. However, it is not easy to evaluate code cloning tools, because there are several factors and hypothesis to taking into account when we do the comparison. As we have seen, each code cloner have its own techniques, comparison algorithm, approaches, complexity and supported languages, so the risk to do an unfair comparison is concrete. To avoid this situation, the authors set a list of possible scenario that analyze different kind of situations in which the code cloning activity may be useful. The evaluation analyzes the results and if the considered code cloners are able to detects the common part in particular duplicated snippet of code. Looking at this state of the art, now we will see the chosen code cloning tool to implement the proposed approach for API function call recommendations. We choose Simian, as it is developed in Java and it is easy to integrate in the overall platform.

## 4.2 Simian overview

As we see from the related works section, we can perform recommendation at different levels of abstraction (pattern, methods, code snippets) in order to give a complete and useful hints to a developer. In our approach, the overall idea is to perform the API recommendation at the level of code snippets that represent the patterns related to the developer's file. To do this, we can exploit the code cloning analysis that we present in the previous section. As tool for this approach, we choose Simian, a project developed in Java that performs this kind of analysis for many languages as Java, C, C#, Ruby, JavaScript, COBOL, Lisp, SQL, Visual Basic. Following the taxonomy in table 2, we can define Simian as a Type-2 code cloner with flexible options on variables, literals, modifiers and it perform the analysis following the textual approach described before. All possible options are described in table 3, although we discard some options that are related to languages different from Java, like ignoreRegions for C files. To test the main functionalities of the tool, we can simply run the jar file a available on the website [simianref] by specifying options and the input file. As output, we see on the console the textual representation

of source coordinates that describe the number of duplicated lines and the original source files. Notice that we can change the type of output using the formatter option (in table 3). Following the tool classification provided by [codecloning1], Simian has the following features:

- It supports object oriented and web languages;
- It not require additional tools or dependencies;
- It is language and platform independent;
- It has free granularity and it analyze line by line of the source files;
- It uses fingerprints technique for the code representation;
- It applies transformation on variable, types and literals using options.

Among the main drawbacks, Simian not include IDE supports and we do a manual integration that we will see in next section. Moreover, it not includes some preprocessing or postprocessing phases as well as an heuristic algorithm for the threshold. It has no external dependencies and it is free downloadable but empirical evaluation is not available. Also the algorithm complexity is not well define, although it depends first of all by the number of lines of code and on the website there is a time approximation for one comparison.

Table4 describe a very simple scenario in which we pick four pairs of Java project with the description of their main features and how lines of code are in commons.

Table 3: Simian options used in the experiment

Option name	Default value	Description
-threshold	6	This option fix an lower bound on the number of duplicated lines of code (if present)
-formatter	none, possible values: plain, xml, emacs, vs (visual studio), yaml, null	This option is used to obtain results in a specified format
-reportDuplicateText	disable , type + to add	With this option, the duplicated lines of code present in all projects are printed on the console
-language	disable , type + to add	This option specify the language of the input files to compare
-defaultLanguage	disable , type + to add	If not file type is not specified, Simian inferred the type and set it as default
-failOnDuplication	able , type - to remove	If this option is able, it causes an exception when the checker finds duplicate code
-reportDuplicateText	disable , type + to add	With this option, the duplicated lines of code present in all projects are printed on the console
-ignoreRegions	disable , type + to add	It ignores block in regions structures (only for C# programming language)
-ignoreBlocks	disable , type + to add	It excludes specified blocks from the comparison (start/end line must be specified)
-ignoreCurlyBraces	disable , type + to add	The curly braces are ignored so it should be match as duplicate line
-ignoreIdentifier	disable , type + to add	With this option, the variable with different ( identifiers match as equal
-ignoreIdentifierCase	able, type - to disable	This option not consider the case of identifiers present in the code: so Name and name are considered equal)
-ignoreStrings	disable , type + to able	This option consider all strings in the comparison and doesn't take care about the form in which are write
-ignoreStringCase	able, type - to disable	Same as previous option but consider the upper and lower case as the same
-ignoreNumbers	disable, type + to add	This option considers different numbers as equal
-ignoreCharacter	disable, type + to add	With this option, all character type match as equal
-ignoreCharacterCase	able, type - to disable	Same as ignoreStringCase but consider char by char. Useful for more precise analysis
-ignoreLiterals	disable, type + to add	All literals should be seen as equal for Simian
-ignoreVariableNames	disable, type + to add	This option allow to Simian to see different variable names as equal
-ignoreModifiers	able, type - to disable	This option doesn't consider modifiers of methods (public, private, protected as element of diversity in the code

Table 4: Projects considered in the comparison

Projects name	Main features	Similarity level (duplicated LOC)
ADTPlugin, ModiscoPlugin	Plugin projects created with same wizard	39 lines of code in common
CyberGea, NeoEMFExample	Cybergea: Plugin, Servlets and JDBC NeoEMF: Metamodels, Neo4J facilities, EMF framework	No lines in common
CyberGea, Scuna project	Cybergea: Plugin, Servlets and JDBC SCuna: Swing GUI and JDBC	12 lines on common (basic JDBC statement)
Simple Servlet, ServletSession	Web projects with servelts	35 lines of code in common

From the scenario, we can see that similar project share more line of code, like the first two pairs that are both Eclipse plugin projects.

## 5 Proposed approach

### 5.1 Overview

After the problem statement and an brief description of Simian tool, we go to describe a possible approach to solve the problem of API function call recommendations. This approach, described in Figure 2, exploits the CLAMS work, in particular the patterns extracted as output, plus the code cloning features provided by Simian. So, at the beginning of the process and after a preprocessing pahse, we have the patterns files and the devoper's file, represented by a single string. Notice that with this method we keep trace on the context in which the user is developing. Then, all these files is used to extract the recommendations in form of patterns by using Simian integrated in Eclipse platform, following the options specified in next sections. Basically, at the of this phase, Simian retrieves the cloned clone between the developer's file and the CLAMS patterns related to the library that the user is implemented. By using these file, the tool performs recommendations by remove the cloned part and suggest to the user the new lines of code that represent the missing pattern for the user. In next sections, we going in deep to describe the entire system and how the integration of Simian and CLAMS works in practise.

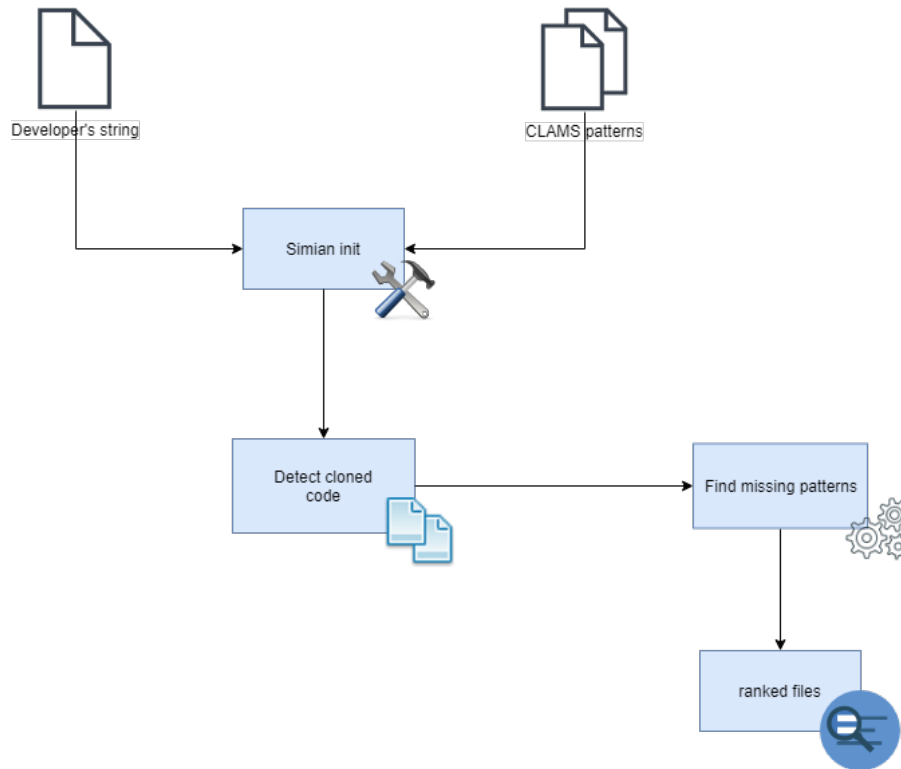


Figure 2: Recommendation for a single file using Simian and CLAMS

### 5.2 Preprocessing

About input files that are necessary to initialise the tool, we have from one side the developer's file with the real snippet of code that he is implementing and may want support for this. From the original file, we extract a portion that represent the context of the recommendation that can be

a list of method invocations or simply a list of variable declaration. This portion is called ground truth and it is the part affected by the recommendation. On other hand, there are patterns mined by CLAMS, in form of ranked Java files sorted by rational specified in CLAMS paper. These file contains patterns, defined as sequence of API method calls that define, instantiate and using class belonging to the APIs contained in the developer's string. The number of these file and also their dimensions in term lines of code depend on the considered libraries. As Simian is a tool based on file comparison, I need to use temporary files of Java to do the comparison; after the process, the files are destroyed. This preprocessing phase is required for Simian because, as we have seen in the related section, the tool not include a built-in preprocessing or postprocessing phases. To extract the snippet of code from the developer's file, we use Rascal, a tool integrated in Java, to visit the AST of the input file and take the body of the method used for retrieving the recommendation. Notice that we consider only compilable files, otherwise Rascal is not able to build the corresponding AST for the analysis. About CLAMS pattern, we have a dataset of 22 libraries plus others related to json and mqtt projects; for each of them, CLAMS retrieves a list of pattern represented by Java files and their number depends on the library that we consider.

### 5.3 Simian within Eclipse platform

Once we define the input, there is another step in order to use Simian to do API recommendations: the integration with Eclipse platform to have a more flexible and usable version of the tool, as Simian doesn't provide any IDE integration. As we said in previous section, the basic version of Simian is a jar file launched from the terminal console with different options (see the Table 1). Although it is very easy to use, in this version is not very suitable for our purposes and it is necessary to integrate directly the Simian jar file, available on this website [ref]. To keep the integration with a Maven project, we create a repository that contain the update version of this jar and put it as reference in the POM file of the project.

Following this documentation, we have the following main classes:

Table 5: Overview about Simian classes

Simian class	Description
Auditstener	This class is necessary to initialise Simian tool and collect all notification from events that occur
Block	This class represents the duplicated block of code as an object and we can interact using method utilities
FileLoader	It is used to load all files for the comparison, with the method load
Checker	This class is used to perform the real comparison by calling the method check() on preloaded files
StreamLoader	Once we load files and create the Checker, this class load them into the Checker
Options	A data structure that encapsulates all options enabled for the comparison
Option	This class represents a single option and we can specify it by accessing to a static field
Language	This class contains static fields to set all supported languages as type of input files
CheckSummary	It contains all statistical data such as cloned code, number of total files, requested time and duplicated files

#### 5.4 CLAMS adaptation

Regarding the CLAMS output, we describe the original structure as well as the necessary modification to integrate it in our approach. The authors, differently from other works, provide the complete source code and commands to set up the entire environment to have CLAMS working on Linux . CLAMS is written in Python and uses srcXML and Astyle to produce xml files and to formatter in a human-readable way the code respectively, but as claimed by the authors, there is no really constrain about the technologies to use in case of a new implementation. As input, CLAMS takes two kind of files: client files, that represent the real project on Github related to the dataset that authors use for evaluation phase while example files are used as training set. All these files are collected in a folder, that CLAMS loads by getting the path. Moreover, there is a namespace file that identify the name of classes used in the clients and example files by using their complete namespaces such as org.codehaus.jackson. The last input used by the main.py file, that is used to initialise the platform, is the methods calls and caller are represented by an .arff file, used in the machine learning domain. As we want to add some other libraries with respect to the original dataset (MQTT-Json projects), we need to replicate the same structure for CLAMS; to do this, we found on Github several projects related to this libraries and produce the arff files using Rascal.

There is a phase of preprocessing in which CLAMS extracts API call and their AST using JDT utilities and represent them in xml using srcXML. The core of the project is the snippet generator module (represented by summarise.py file) that takes as input a source code file (java in this case) and using srcXML they first replace literals with xml types and delete comment. Then , they separate the API code from code that doesn't contain API call and highlights the variable in local scope of API. Finally, the code without API call is removed and Clams add some com-

ments near the API statement and needed variables. Notice that their approach considers also the classical statement like if-else structure as a part of API statement. For clustering, they use both HDBSCAN and k-medoids algorithm that are quite similar and differs only in the precision of the returned snippet (HDBSCAN is more accurate but k-medoids covers more methods). For both of them, the authors import Python libraries that implement these algorithm quite well and we can switch the algorithm by change the parameter in the main.py. Moreover, they have the file ranking.py to order the generated snippet. The rank is based on the example files that contains a sequence of API call; if the sequence within the file is a super-sequence of the sequence of snippet that we considered, so this snippet is supported, and its rank is increased. In the result folder, CLAMS put the library that we want to analyze, the methods, the source file (both in .java and xml format), some json file that represent all information about a method (class, package, rank, id) and the arff file related to the library.

For the integration step in our platform, it is necessary to slightly modify the original approach to have better results. In particular, if we use the pattern of CLAMS as they are, there are some bias because, through srcML, CLAMS substitutes the literals with its own type and Simian is not able to detect them as cloned code, even using all available options regarding the code. So, to avoid this situation, we must modify the function that substitute literals, putting some default value instead of srcML types. This modification doesn't affect the validity and accuracy of extracted path because is just a matter of modify literals with another.

## 5.5 API recommendations

At the end of these preparatory phases, we describe now the core of this project, the API recommendations. Once the Simian is launched, it performs the detection of code cloning activity on the CLAMS patterns files and the developer's code snippet. Notice that the notion of cloned code depends on the options that we have selected and turn on: the mandatory options to enable is the threshold, that set the minimum line of code in commons, reportDuplicateText, otherwise we couldn't show and manipulate the result and language that is Java because we analyze projects related to it. Without this options, Simian gives us only the fingerprints that represent the source coordinates of the files, that are not significant for our aims. So, to have a better representation, we have put the results in a wrapper class that represent the Pattern object in which we have all attributes to describe in the right manner the recommendation. Other options, such as the strings, identifiers or modifiers that should be introduced in the comparison, can be enable with respect to the level of cloning that we want to reach. To find useful results, it is necessary to set at least ignoreIdentifiers, ignoreIdentifierCase, ignoreLiterals, ignoreVariableName, ignoreNumbers and ignoreModifiers because Simian goes beyond the developer personal implementations and looking only for the structure of the code, in order to use the concept of pattern in a more effective way. Based on these options, Simian applies the proper transformations on the original textual code in order to perform the comparison.

Furthermore, Simian compares the pair developer' snippet - pattern because some CLAMS pattern includes some duplicated lines of code and this can bring some bias. Once we load the files, the check is performed and the results that include lines of code, name of pattern file and time to perform the comparison and put all in the wrapper class mentioned before. At the end of this step, we have the patterns (a complete one or only partial) that the developer is start to implement and we can discard it from the comparison, as the developer is not interested to see what he have done so far. Moreover, new pattern can introduce something new or suggest an alternative implementation for the developer. The last step is remove the duplicated line of code from the suggested pattern and show to the developer only the novel part, in form of some that integrate his code or completely new pattern, related



of course to the APIs that he is implementing. About the ranking, we order the pattern by considering the number of cloned lines, so the first pattern is the contains more duplicated lines rather than second and so on. The rank phase is simply performed on the APIRecommendation object that we produce as output.

---

```
Current file
Twitter mTwitter;
    final String CONSUMER_KEY;
    final String CONSUMER_SECRET;

    final Action action;
    final User first_source;
    final int sources_length;
```

---

---

```
New pattern found
mTwitter = new
    TwitterFactory().getInstance();
mTwitter.setOAuthConsumer(CONSUMER_KEY,
    CONSUMER_SECRET);
```

---

---

```
CLAMS pattern
{
    Twitter mTwitter;
    final String CONSUMER_KEY;
    final String CONSUMER_SECRET;
    mTwitter = new
        TwitterFactory().getInstance();
    mTwitter.setOAuthConsumer(CONSUMER_KEY,
        CONSUMER_SECRET);
}
```

---

## 6 Validation

### 6.1 Evaluation framework

### 6.2 Comparing results

### 6.3 Performances

## 7 Conclusion

## References

- [1] Lu Zhang JianPei Hao Zhong, TaoXie and Hong Me. Mapo: Mining and recommending api usage patterns. pages pp 318–343.

## 8 References

Here there are the related works that I consulted until now:

- [1] Hao Zhong, TaoXie, Lu Zhang, JianPei and Hong Me, MAPO: Mining and Recommending API Usage Patterns, In ECOOP pp 318-343, 2009
- [2] Jue Wang, Yingnong Dang, Hongyu Zhang, Kai Chen, Tao Xie, Dongmei Zhang, Mining succinct and high-coverage API usage patterns from source code, MSR pp 319-328, 2013
- [3] Summarizing Software API Usage Examples using Clustering Techniques, Nikolaos Katirtzis, Themistoklis Diamantopoulos and Charles Sutton
- [4] Anh Tuan Nguyen, Michael Hilton, Mihai Codoban, Hoan Anh Nguyen, Lily Mast, Eli Rademacher, Tien N. Nguyen, Danny Dig, API code recommendation using statistical learning from fine-grained changes. SIGSOFT FSE pp 511-522, 2016
- [5] Raymond P. L. Buse, Westley Weimer, Synthesizing API usage examples. ICSE pp 782-792, 2012
- [6] Hudson S. Borges, Marco Tulio Valente, Mining usage patterns for the Android API. PeerJ Computer Science 1: e12 2015
- [7] Haoran Niu, Iman Keivanloo, Ying Zou, API usage pattern recommendation for software development. Journal of Systems and Software 129: 127-139, 2016
- [8] Ferdian Thung, Shaowei Wang, David Lo and Julia Lawall, Automatic Recommendation of API Methods from Feature Requests, ASE 2013, Palo Alto, USA
- [9] Gerhard Fischer and Yunwen Yu, Reuse-Conducive Development Environments, Automated Software Engineering, 12, 199–235, 2005