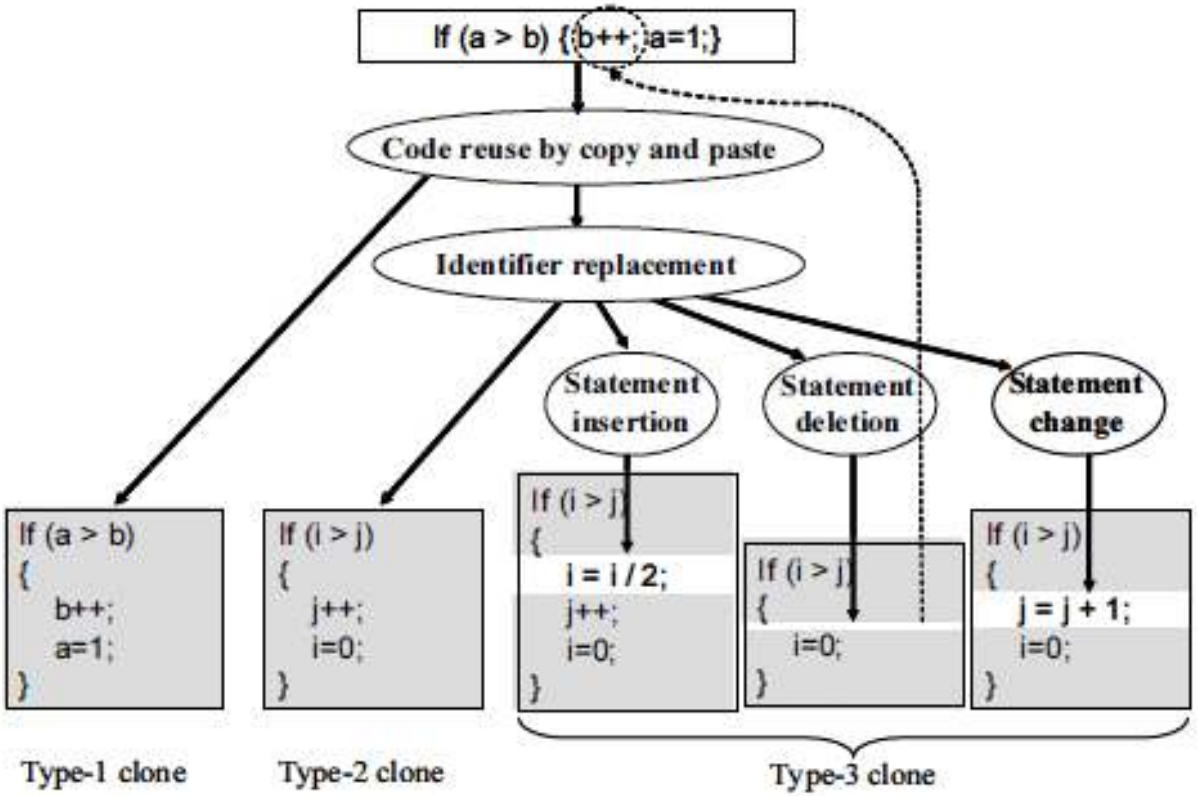
Systematic Literature Review for Detecting and Classifying Code Clones

Alexandre Martins Paiva
ampaiva@gmail.com

Background

- What is code clone?



SLR

- Systematic Literature Review [1] [2]

Planning the Review

1. Identification of the need for a review
2. Commissioning a review
3. Specifying the research questions
4. Developing a review protocol
5. Evaluating the review protocol

Conducting the Review

1. Identification of research
2. Selection of primary studies
3. Study quality assessment
4. Data extraction and monitoring
5. Data synthesis

Reporting the Review

1. Specifying dissemination mechanisms
2. Formatting the main report
3. Evaluating the report

Identification of the need for a review

- State-of-art on clone detection
- Clone detection using concern metrics

Specifying the research questions

- RQ1: What methods have been used to detect code clone?
- RQ2: How code clone has been classified?
- RQ3: Do Concern Metrics have been used to detect code clone?

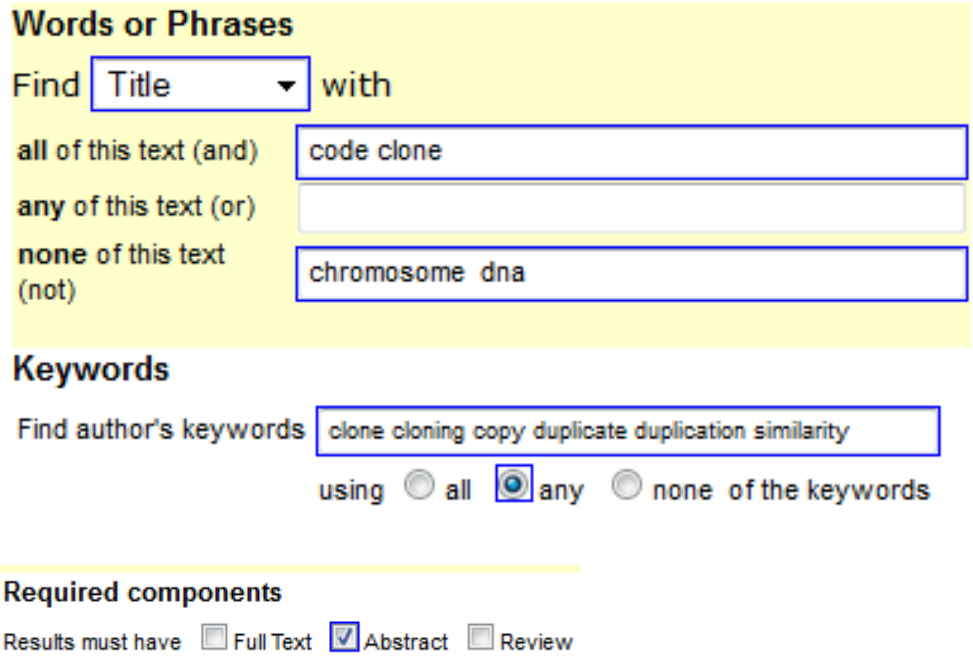
Developing a review protocol

- Source Selection
- Search String
- Primary study identification and evaluation

Source Selection

- ACM Digital
- IEEE Xplore

Search String



Primary study identification and evaluation

For each paper

1. Use the title to exclude any paper that is clearly not related to the research questions
2. Use the abstract and keywords to exclude any paper that has a relevant title but is not related to the research questions
3. Read the paper and include it if any of the research questions are addressed

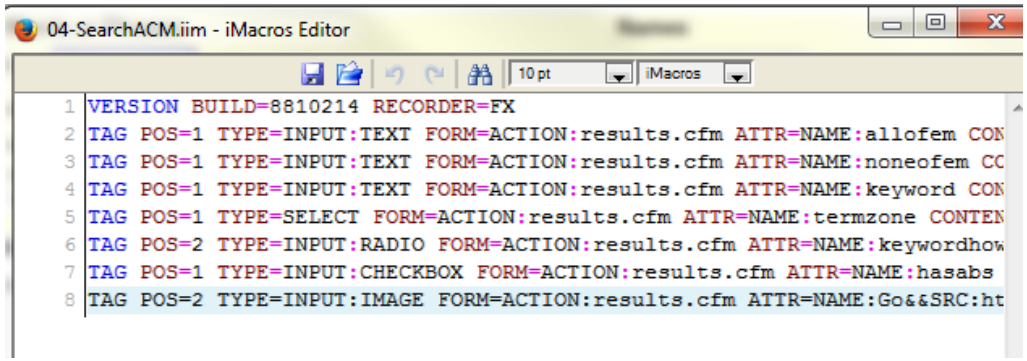
Until consecutive **irrelavants (*)** =5

RQ1	RQ2	Relevance
N	N	0
N	C	1
C	N	2
C	C	3
N	F	4
F	N	5
C	F	6
F	C	7
F	F	8

N=Nothing
C=Cited
F=Focus

Selection of primary studies

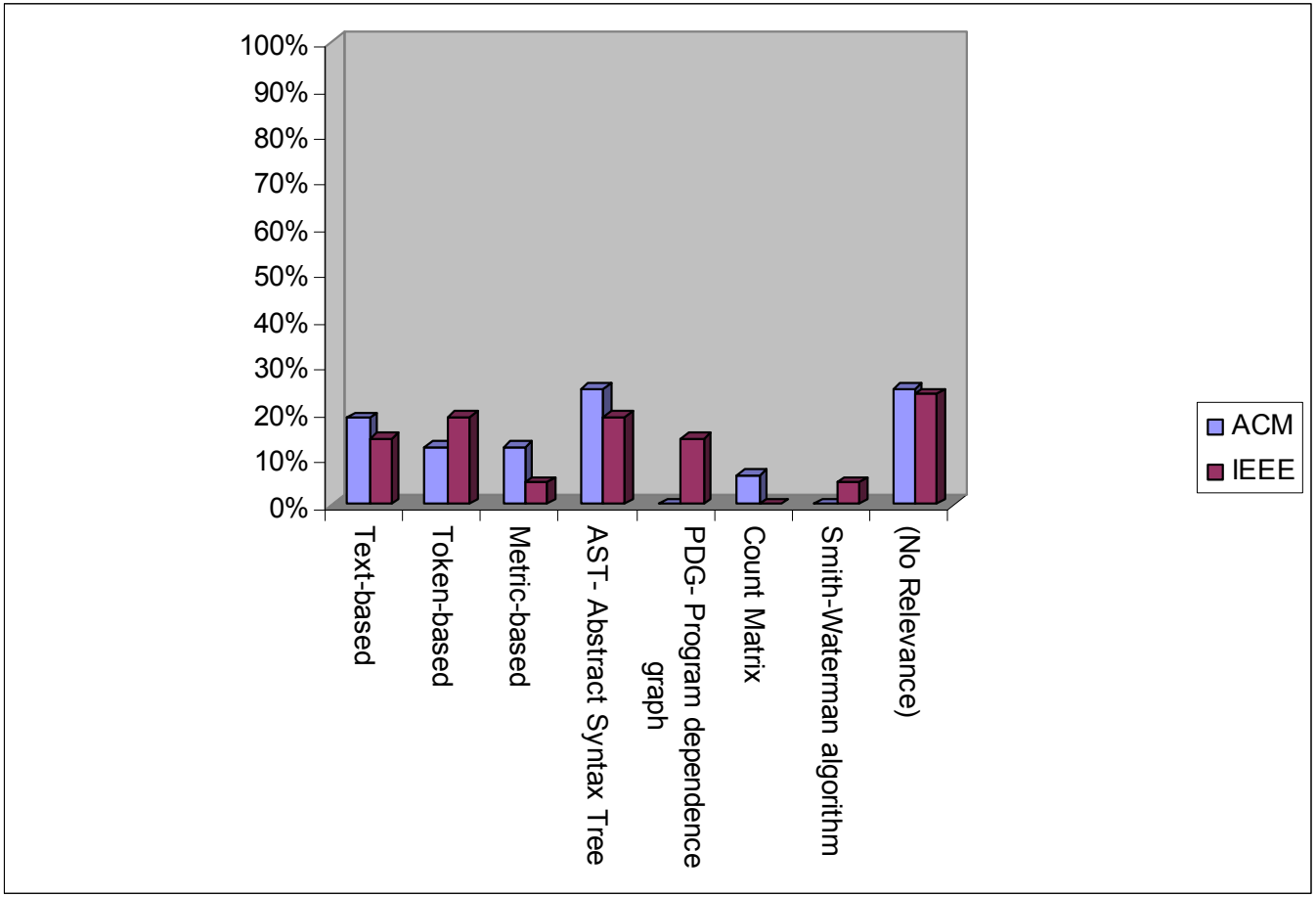
- Criteria refined based on RQs
- Use of Snowballing
- Search database (iMacros support)



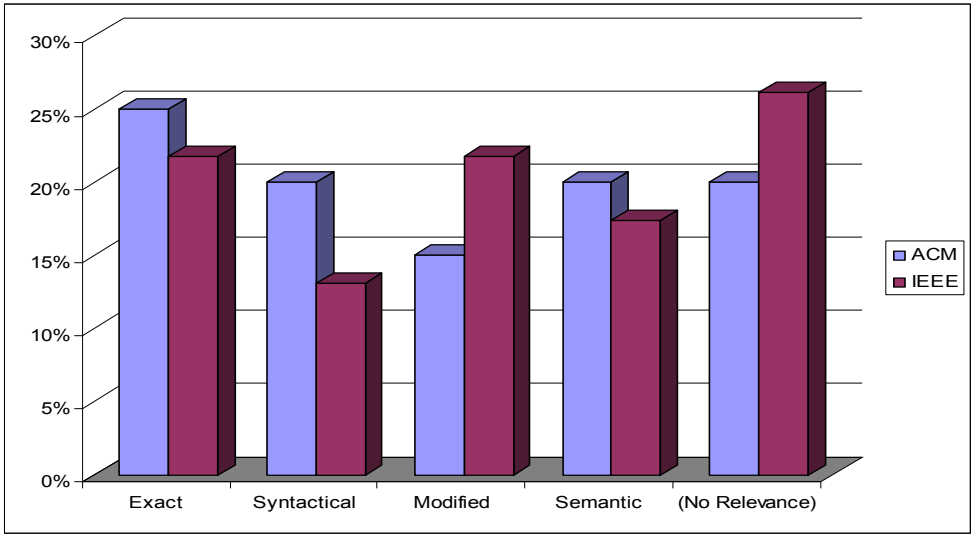
Data extraction and monitoring

Source	First Search
ACM	29
IEEE	60

RQ1: Methods


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Text-based
Token-based
Metric-based
AST- Abstract Syntax Tree
PDG- Program dependence graph
Count Matrix
Smith-Waterman algorithm
(No Relevance)
ACM
IEEE

RQ2: Classification



Exact: Identical code fragments except for variations in whitespace, layout and comments.

Syntactical: Syntactically identical fragments except for variations in identifiers, literals, types, whitespace, layout and comments.

Modified: Copied fragments with further modifications such as changed, added or removed statements, in addition to variations in identifiers, literals, types, whitespace, layout and comments.

Semantic: Two or more code fragments that perform the same computation but are implemented by different syntactic variants.

RQ3: Concern Metrics

- No papers found

Conclusion

- RQ1: Text, Token, Metric, AST, PDG
- RQ2: Exact, Syntactical, Modified, Semantic
- RQ3: No papers found
- SLR: Powerful tool
  - Snowballing: important step [3]

References

1. Kitchenham (2007), “Guidelines for performing Systematic Literature Reviews in Software Engineering”, EBSE.
2. Pate, J. Tairas, R. Kraft, N. (2013), “Clone evolution a systematic review”, J. Softw.: Evol. and Proc.
3. Roy, K. (2009), “Comparison and Evaluation of Code Clone Detection Techniques and Tools: A Qualitative Approach”, Science of Com Programming.

Questions