An Investigation of Cloning in Web Applications

Damith C. Rajapakse and Stan Jarzabek

Department of Computer Science, School of Computing, National University of Singapore

{damithch, stan}@comp.nus.edu.sg
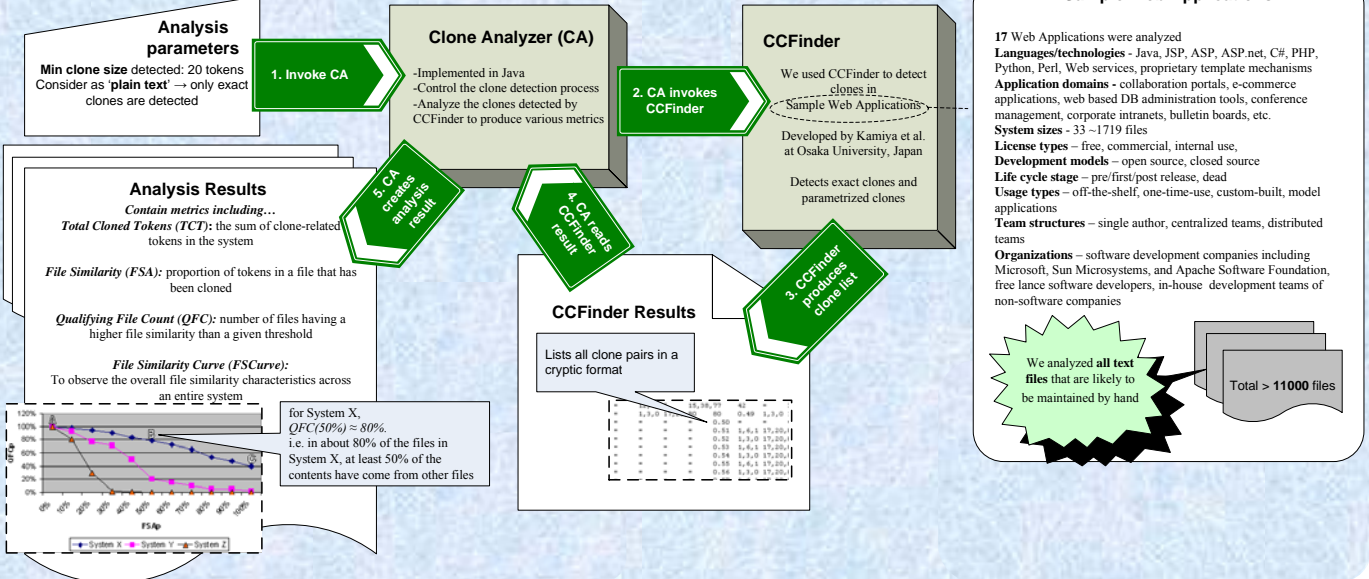
NUS National University of Singapore

Overview: Cloning (ad hoc reuse by copy-paste-modify) adversely affect maintainability of software. Cloning has been a known problem in traditional software. How big a problem is it in Web engineering? We investigate this issue in our study.
- Our results confirm potential benefits of reuse-based methods in Web engineering.
- A framework of metrics and presentation views that we defined and applied in our study may be useful in other similar studies.

What are clones? We define clones as similar text fragments (similar classes, functions, markup segments, text descriptions, etc.)

How do we detect clones?

Analysis parameters

Min clone size detected: 20 tokens
Consider as '**plain text**' → only exact clones are detected

1. Invoke CA

Clone Analyzer (CA)

-Implemented in Java
-Control the clone detection process
-Analyze the clones detected by CCFinder to produce various metrics

2. CA invokes CCFinder

CCFinder

We used CCFinder to detect clones in Sample Web Applications

Developed by Kamiya et al. at Osaka University, Japan

Detects exact clones and parametrized clones

5. CA creates analysis result

4. CA reads CCFinder result

3. CCFinder produces clone list

Analysis Results

Contain metrics including...
Total Cloned Tokens (TCT): the sum of clone-related tokens in the system

File Similarity (FSA): proportion of tokens in a file that has been cloned

Qualifying File Count (QFC): number of files having a higher file similarity than a given threshold

File Similarity Curve (FSCurve): To observe the overall file similarity characteristics across an entire system

for System X,
QFC(50%) ≈ 80%.
i.e. in about 80% of the files in System X, at least 50% of the contents have come from other files

CCFinder Results

Lists all clone pairs in a cryptic format

Sample Web Applications

17 Web Applications were analyzed
Languages/technologies - Java, JSP, ASP, ASP.net, C#, PHP, Python, Perl, Web services, proprietary template mechanisms
Application domains - collaboration portals, e-commerce applications, web based DB administration tools, conference management, corporate intranets, bulletin boards, etc.
System sizes - 33 ~1719 files
License types – free, commercial, internal use,
Development models – open source, closed source
Life cycle stage – pre/first/post release, dead
Usage types – off-the-shelf, one-time-use, custom-built, model applications
Team structures – single author, centralized teams, distributed teams
Organizations – software development companies including Microsoft, Sun Microsystems, and Apache Software Foundation, free lance software developers, in-house development teams of non-software companies

We analyzed **all text files** that are likely to be maintained by hand

Total > **11000** files

What have we found out so far?

Total Cloning Percentage

High risk in creating inconsistent updates (update anomalies)!

High Levels of Cloning!

Most clones are longer than 100 tokens

Length of clones

Breakdown of clones by length for each WA

There is good clone concentration within files

% of 'cloned' and 'highly cloned' files

[TCT%>80%] [TCT%>50%]

FSCurves for all WAs

close to 50% of the files have at least 50% of their content cloned

Max
Avg

What about false positives?

We detect only exact clones
→ false positives are minimum
→ but we miss many parametric/gapped clones

So actual cloning level can be even higher?

WSFinder Vs CCFinder

Yes, A web-specific clone detector detected more clones.

Cloning level in WAs is higher than traditional applications

Web-specific files Vs General files

We analyzed cloning in Web-specific files (e.g. HTML, JSP...) Vs General files (e.g., C++, C#, ...)
Assumption: cloning in traditional applications is not worse than cloning in general files inside WAs

Future Work

- Complement this quantitative analysis with more qualitative analysis
- Address design-level similarities, so-called structural clones
- Find synergies between run time and construction time technologies in solving the cloning problem of WA domain