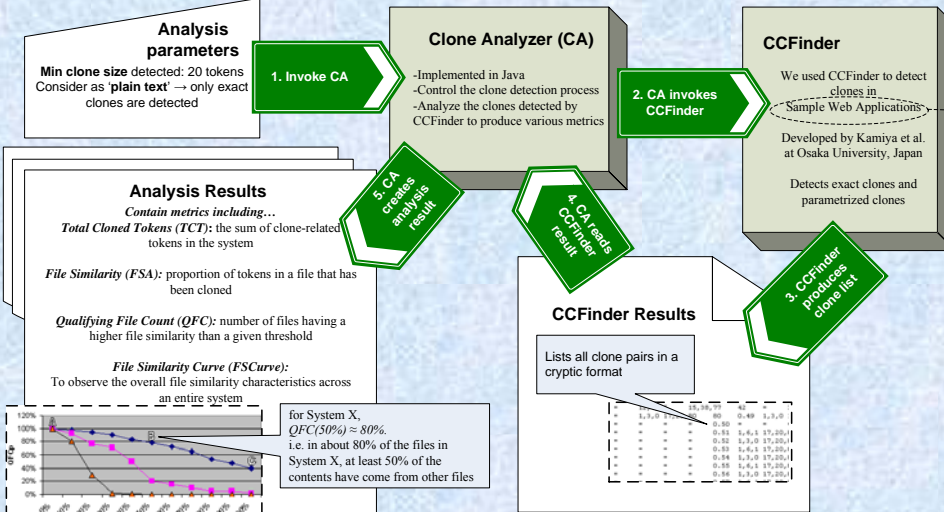


Overview: Cloning (ad hoc reuse by copy-paste-modify) adversely affect maintainability of software. Cloning has been a known problem in traditional software. How big a problem is it in Web engineering? We investigate this issue in our study.

- 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?



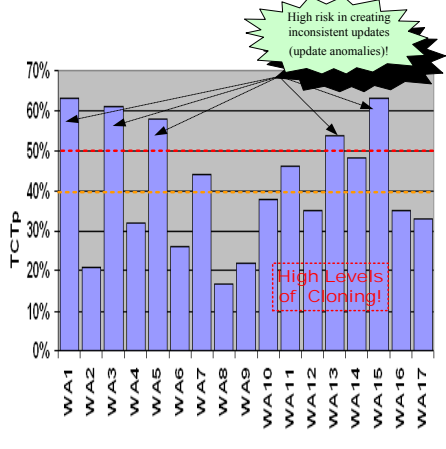
Sample Web Applications

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

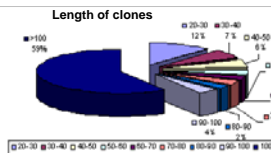
We analyzed all text files that are likely to be maintained by hand
 Total > 11000 files

What have we found out so far?

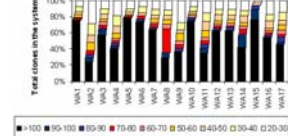
Total Cloning Percentage



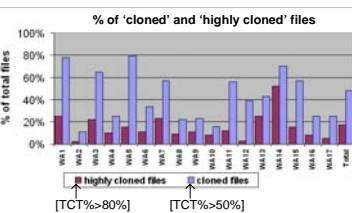
Most clones are longer than 100 tokens



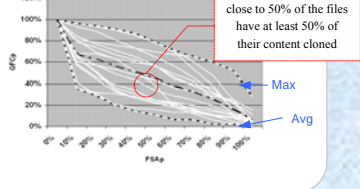
Breakdown of clones by length for each WA



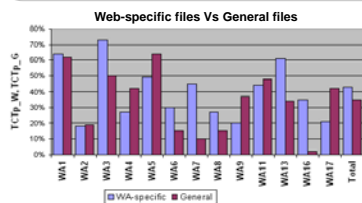
There is good clone concentration within files



FSCurves for all WAs



Cloning level in WAs is higher than traditional applications



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

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?



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

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