Connecting Technology with Real-world Problems — From Copy-paste Detection to Detecting Known Bugs

(Keynote Abstract)

Yuanyuan Zhou
UC San Diego, USA

ABSTRACT

In my talk, I will share with you our experience in applying and deploying our source code mining technology in industry. In particular, the most valuable lesson we have learned is that sometimes there is a bigger problem in the real world that can really benefit from our technology but unfortunately we do not know about it until we closely work with industry. In 2004, motivated from some previous research work that pointed out copy-pasting as a major reason for majority of the bugs in device driver code, my students and I applied data mining technology (specifically frequent subsequence mining algorithms) in identifying copy-pasted code and also detecting forget-to-change bugs introduced during copy-pasting. The benefit of using data mining is that it is highly scalable (20 minutes for 4-5 millions lines of code), and can tolerate changes such as statement insertion/deletion/modification as well as variable name changes. When we released our tool called CP-Miner to the open source community, it attracted some inquiries from industry. These inquiries have motivated us to start a company to commercialize our tools.

During the commercialization process, our customers have taught us that our technology can be applied to solve a major headache faced by many embedded system vendors such as storage companies, network devices, telecommunications, handhelds, electronics, etc. These companies typically have to maintain tens and even hundreds of active branches of similar software, one for slightly different devices or with different customization. These branches are 60–70 % similar and are being developed in parallel after split (and are usually never merged back together). So when developers decide to fix a bug or security hole in one branch, it is usually a big challenge for them to check what other branches a similar fix should be also applied. So many costly incidents have happened in the field due to some known bugs (a bug that has already been diagnosed and fixed in some other branches). Since these branches have diverged over the years, many code are similar but are not exactly the same. Also, file names can change, etc. Therefore, it is hard to rely on source control systems such as ClearCase, Subversion to keep track their differences. Being pushed by many customers who suffer this pain, we applied our technology to this problem and build a tool called PatchMiner. Currently PatchMiner has been deployed and widely used in several large companies (more information can be found at http://www.patterninsight.com/). It is a very interesting journey. To me personally, I learned that sometimes what I (as an academic) feel as a solution to a major problem may be only a nice-to-have in the real world; and it really requires close interaction with industry to understand their painpoints.

BIOGRAPHY

Yuanyuan Zhou is currently a Qualcomm Chair Professor at UC-San Diego. Prior to UCSD, she was a tenured associate professor at University of Illinois at Urbana Champaign. She has also worked at NEC Research Institute as a scientist after completing her Ph.D at Princeton in 2000. Her research interests span the areas of operating systems, architecture, system reliability and maintainability. She was the recipient for the Alfred Sloan Fellowship 2007, UIUC Gear Faculty Award 2006, NSF Career-2004 award, the CRA-W Anita-Borg Early Career Award 2005, the DOE Early Career Principle Investigator Award 2005, the IBM Faculty Award 2004 2005, the IBM SUR-2003 award and NetApp Faculty Fellowship 2010. She has 3 papers selected into the IEEE Micro Special Issue on Top Picks from Architecture Conferences and one best paper in SOSP 2005. She and her students have released several software quality assurance tools that are currently been used many developers in many commercial companies as well as open source projects.

MSR'11, May 21–22, 2011, Waikiki, Honolulu, HI, USA
ACM 978-1-4503-0574-7/11/05