Mining Ultra-large Scale Software Repositories with Boa

Robert Dyer
2013-14 Tom Miller Fellow, PhD student
Computer Science, Iowa State University
Thursday October 17, 2013
3:40pm, B29 Atanasoff Hall

Mining software repositories provides developers and researchers a chance to learn from previous development activities and apply that knowledge to the future. Ultra-large-scale open source repositories (e.g., SourceForge with 350k+ projects) provide an extremely large corpus to perform such mining tasks on. This large corpus allows researchers the opportunity to test new mining techniques and empirically validate new approaches on real-world data. However, the barrier to entry is often extremely high. Researchers interested in mining must know a large number of techniques, languages, tools, etc., each of which is often complex. Additionally, performing mining at the scale proposed above adds additional complexity and often is difficult.

The Boa language and infrastructure was developed to solve these problems. Boa provides a web-based interface for mining ultra-large-scale software repositories. Users use a domain-specific language tailored for software repository mining and submit queries to the website. These queries are then automatically parallelized and executed on a cluster. This talk demonstrates how to efficiently perform mining tasks on over 600k open-source software projects. I introduce the language and supporting infrastructure and then show how to perform several mining tasks.

Robert Dyer is a Ph.D. candidate in the department of Computer Science at Iowa State University, where he received his M.S in 2008 and B.S. in Computer Science and Computer Engineering in 2006. His research interests are in software engineering and programming languages. Robert is the lead researcher on the Boa project, where he helped implement the language and supporting infrastructure. Previously, he worked on the Ptolemy and Nu languages. Robert has served on the program committee for the OOPSLA’13 artifact evaluation and was an external reviewer for conferences such as OOPSLA, TOPLAS, GPCE, TAOSD, and AOSD. Robert is currently a member of ACM, SIGPLAN, and SIGSOFT.

http://www.cs.iastate.edu/~colloq/new/

ALL SCIENCE IS COMPUTER SCIENCE