Data Min Knowl Disc (2009) 18:300–336
DOI 10.1007/s10618-008-0118-x

Sourcerer: mining and searching internet-scale software repositories

Erik Linstead · Sushil Bajracharya · Trung Ngo ·
Paul Rigor · Cristina Lopes · Pierre Baldi

Received: 16 September 2007 / Accepted: 18 September 2008 / Published online: 15 October 2008

Abstract Large repositories of source code available over the Internet, or within large organizations, create new challenges and opportunities for data mining and statistical machine learning. Here we first develop Sourcerer, an infrastructure for the automated crawling, parsing, fingerprinting, and database storage of open source software on an Internet-scale. In one experiment, we gather 4,632 Java projects from SourceForge and Apache totaling over 38 million lines of code from 9,250 developers. Simple statistical analyses of the data first reveal robust power-law behavior for package, method call, and lexical containment distributions. We then develop and apply unsupervised, probabilistic, topic and author-topic (AT) models to automatically

Responsible editor: Eamonn Keogh.

Erik Linstead, Sushil Bajracharya, and Trung Ngo have contributed equally to this work.

E. Linstead · S. Bajracharya · T. Ngo · P. Rigor · C. Lopes · P. Baldi (✉)
Donald Bren School of Information and Computer Sciences, University of California, Irvine, USA
e-mail: pfbaldi@ics.uci.edu

E. Linstead
e-mail: elinstea@ics.uci.edu

S. Bajracharya
e-mail: sbajrach@ics.uci.edu

T. Ngo
e-mail: trungcn@ics.uci.edu

P. Rigor
e-mail: prigor@ics.uci.edu

C. Lopes
e-mail: lopes@ics.uci.edu