Towards Mining Informal Online Data to Guide Component-Reuse Decisions

Sanchit Karve
McAfee Software
20460 NW Von Neumann Dr,
Beaverton, OR 97006
Sanchit_Karve@mcafee.com

Christopher Scaffidi
School of Electrical Engineering and Computer Science
Oregon State University
Corvallis, OR, USA, 97331
cscaffid@eeecs.oregonstate.edu

ABSTRACT
Online repositories provide components available for reuse, but this does not mean all such components are equally reusable. Components might be unreliable, overly specialized, or otherwise inappropriate for reuse. Repositories collect reviews, ratings, and other data intended to help software engineers choose components. But do these data actually provide any information related to reusability? If so, then how can such information be extracted from the data?

To address this question, we analyzed online ratings, reviews and other data for nearly 1200 online components, computed statistics for each component based on these data, and used factor analysis to identify three groups of statistics (factors) that were each internally correlated. We then interviewed software engineers about the reusability of 36 other components and used linear regression to test how well the 3 factors actually corresponded to component reusability.

We found that 2 of the 3 factors were indeed related to reusability. Specifically, the reusability of components could be predicted on the basis of component authors' prior work and the documentation provided about components. This result could be used in future work to develop enhanced search engines that highlight components which are potentially reusable and perhaps worthy of more time-consuming evaluation such as by applying formal methods. Additionally, our results reveal opportunities to improve online repositories through specific simplifications as well as enhancements.

This paper will appear at CBSE’2013. For the full version of the paper, refer to the IEEE website.