The Comment Density of Open Source Software Code

Oliver Arafat
Siemens AG, Corporate Technology
oarafat@gmail.com

Dirk Riehle
SAP Research, SAP Labs LLC
dirk@riehle.org, http://dirkriehle.com
http://twitter.com/dirkriehle

Overview

Summary (blog post): The Sweet Spot of Code Commenting in Open Source

Abstract: The development processes of open source software are different from traditional closed source development processes. Still, open source software is frequently of high quality. Thus, we are investigating how open source software creates high quality and whether it can maintain this quality for ever larger project sizes. In this paper, we look at one particular quality indicator, the density of comments in open source software code. In a large-scale study of more than 5,000 projects, we find that active open source projects document their source code, and we find that the comment density is independent of team and project size, but not of project age. In future work, we intend to correlate comment density with project success or failure.

Reference: Oliver Arafat, Dirk Riehle.
Analysis Process and Tool Chain

Raw data source
- Local database (ohloh.net snapshot, crawled sources)
- Web services access (ohloh.net, sourceforge.net, others)

Pre-processing
- Database querying using SQL and scripts
- Java library for computationally heavyweight filters, aggregation

Aggregated data source
- Output of pre-processing stage for specific analytical tasks
- Aggregated data significantly improves analysis speed

Analytical processing
- Mines aggregated (and raw) data for insights, hypothesis testing
- At present basic processing (Excel), machine learning next

Analysis output
- Results of analytical processing: averages, distributions, correlations
- Presented as models, tables, graphs, charts, etc.
Commenting Practice in Open Source

The graph shows the relationship between project size in lines of code (LoC = CL + SLoC) and comment lines. The equation $y = 0.1596x$ with a coefficient of determination $R^2 = 0.872$ indicates a strong linear correlation between the two variables.
Comment Density

- **Comment density**
  - Percentage of lines of code that are comments
  - Measure of how well documented code is
  - Indicator of likelihood of survival

- **Definitions**
  - $CL = \text{comment lines}$
  - $SLoC = \text{source lines of code}$
  - $CD = CL / (CL + SLoC)$

---

**Graph Details**

- Mean: 0.1867
- Median: 0.1674
- Standard Deviation: 0.1088
- Correlation: -0.00787

---

© 2009 Dirk Riehle. All Rights Reserved.
Variation by Programming Language

Java

PERL

More here: How Open Source Comments (by Programming Language)
Comment Density by Commit Size

Total Average Comment Density

sloc size = 1-100
mean = 0.2513
median = 0.2340
stdev = 0.0626

sloc size = 50-100
mean = 0.2234
median = 0.2190
stdev = 0.0169

sloc size = 80-100
mean = 0.2224
median = 0.2171
stdev = 0.0209
Comment Density by Team Size

- **team size = 1-20**
  - mean = 0.1914
  - median = 0.1878
  - stdev = 0.0255

- **team size = 1-50**
  - mean = 0.1922
  - median = 0.1906
  - stdev = 0.0425

- **team size = 1-100**
  - mean = 0.1856
  - median = 0.1857
  - stdev = 0.0641
  - correl = -0.0550

© 2009 Dirk Riehle. All Rights Reserved.
Comment Density by Project Age

Average Comment Density

corr = -0.9054

Project Age [months]

© 2009 Dirk Riehle. All Rights Reserved.
Commenting Practice Summary

• Successful open source projects follow a consistent commenting practice
  – 1 out of 20 commits serves commenting purposes only
  – 1 out of 5 non-empty lines is a comment line

• Average comment density is independent of project size

• Average comment density is independent of team size

• Average comment density slowly falls with project age
References


Thank you! Questions?

For any feedback or questions, please email the authors:

Oliver Arafat
Siemens AG, Corporate Technology
oarafat@gmail.com

Dirk Riehle
SAP Research, SAP Labs LLC
dirk@riehle.org, http://dirkriehle.com
http://twitter.com/dirkriehle