The relation between user age and programming knowledge

An exploration of stack overflow

Tianpei Xia
North Carolina State University
txia4@ncsu.edu

ABSTRACT

Computer programmers spend years of deliberate practice becoming an expert. But what happens after that? Do programming experts continue to improve, remain the same level, or become worse at some points? A diversity of opinions exist on this matter, but many seem to think that aging brings a decline in adoption and absorption of new programming knowledge.

On this theme, Morrison et al. [3] develop several research questions about the relation between programming knowledge and user age, and collect data from Stack Overflow (SO) to address these questions. They observe that programmer reputation scores increase relative to age well into the 50's, that programmers in their 30's tend to focus on fewer areas relative to those younger or older in age.

KEYWORDS

Stack Overflow, Aging, Programming Knowledge, Data Mining, Software Repositories

1 INTRODUCTION

As the requirement for software keep increasing, the demand for programmers start to expand. The authors turn their attention to a diversity of opinion on the role of age in programming performance. It's not easy to directly assess programming performance, but exploring programming knowledge however is possible. Stack Overflow (SO) is a programming discussion website based on questions and answers about programming that are provided by the online community using the site. As of February 2013, Stack Overflow has over 1.6 million registered users and 4.5 million questions. SO provides access to its underlying data through online queries, and through various forms of data export files.

Over 300,000 SO users have specified their age, which suggests that it might prove a suitable candidate for exploring questions regarding age and programming.

In this poster, I introduce the work of Morrison et al. [3], who use SO data to explore the relationship between aging and programming knowledge. They conclude that age has a positive effect on quality and breadth of programming knowledge, at least up to some point.

The research questions of this poster are:

(1) RQ1: Does age have a positive effect on programming knowledge?
(2) RQ2: Do older programmers possess a wider variety of technologies and skills?

Foundations of Software Science, Fall 2018, Computer Science

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$9999.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

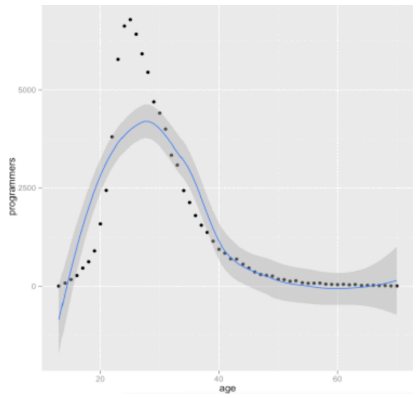

Figure 1: Stack Overflow user count by age

2 DATA COLLECTION AND ANALYSIS

The authors initially explored the SO data via downloading the 2013 MSR Mining challenge PostgreSQL data dump [1]. They then built equivalent SQL queries to extract data from the live site. In order to emphasize the evaluation of knowledge over time, they choose data by using the following criteria:

- Users aged 15 to 70, forming a 'working age' interval, and excluding outlier age values like 99, and Null.
- Users who answered questions in 2012. The 'Posts' of SO may be questions or answers, they are assigned a creation date, allowing each post to be associated with a year. The authors wanted to avoid rolling more than a single year's worth of data in to a comparison based on age.
- Reputation between 2 and 100,000. SO calculates a 'reputation' value for each user, reflecting site familiarity, subject expertise and peer respect. For the purposes of the analysis, the authors treat SO reputation as a proxy for programming knowledge. They eliminated 1 since it is the default reputation assigned at user creation, and dropped reputations over 100,000 as outliers.

The resulting sample contains 84,284 users, with a mean age of 29.02 and a mean reputation of 1073.9. Fig 1 presents a breakdown of user count by age group for the sample.

3 RESEARCH METHODOLOGY AND RESULTS

3.1 Does age have a positive effect on programming knowledge?

As discussed above, the authors treat reputation as a proxy for programming knowledge. They queried the SO data query site

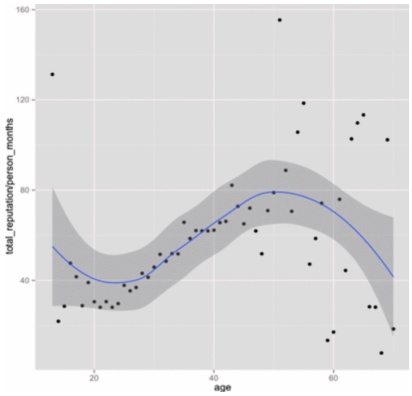

Figure 2: Stack Overflow average reputation by age

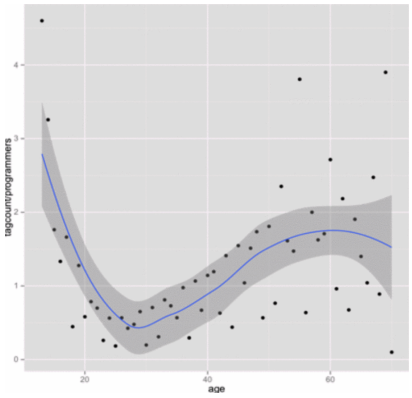

Figure 3: Unique tags by age

according to the criteria specified above, and collected the number of programmers, total reputation, and total number of SO membership months for each age in the range 15 to 70. They normalized total reputation by person months in order to correct for length of membership on SO. They calculate 'membership months' as the number of months between the user creation date and the day the query was run. As shown in Fig 1, the number of programmers is roughly normally distributed around age 29, though skewed right. Reputation (Fig 2) is roughly linearly increasing from age 25 in to the 50's. They ran a linear regression (**R**, **lm**(**total_reputation/person_months** ∼ **age**), which indicated a positive slope of. 52 total reputation points/person month per year, This suggests that there is a positive relationship between age and reputation on SO.

3.2 Do older programmers possess a wider variety of technologies and skills?

The authors theorize that programmers acquire technology and skill knowledge as they progress in their careers.

SO provides a 'tag' feature, allowing each question asked to be annotated with one or more terms indicating the subject matter of the question, e.g. 'javascript', 'python', 'algorithm', 'design-patterns'. We can trace users to the questions they ask and answer, and to the tags associated with each question. The authors count the references to each tag, grouped by age and normalize by number of programmers reporting that age. They would expect a plot of this quantity to increase in some proportion to age.

The authors built an SQL query counting unique tags used by each age, and normalized this total by the number of programmers of that age. A plot of the resulting data, Fig 3, indicates that their expectations were incorrect; there is initially a decline in the mean number of tags per programmer, bottoming around age 30, followed by an increase in the 40's and 50's and dispersion in the 60's.

4 LIMITATIONS

Does the SO population represent the programmer population? US statistics on programmer employment [4] suggest that the age distribution of professional programmers skews older than the user distribution of SO. It is possible that SO represents a kind of 'early adopter' rather than the programming profession. Determining the relationship between the SO user base and the programmer population is necessary before inferences can be made for prediction, planning or other purposes.

Does SO reputation measure programming knowledge? High reputation scores may reflect efforts in resume building and self-promotion as well as programming knowledge. It is possible that the causation between age and programming knowledge exhibited in the data is because higher-knowledge individuals choose to stay active and engaged later in life, rather than because individuals gain knowledge over time, a point made by Hultsch et al. [2].

Does measuring programming knowledge say anything about programming ability? High SO question and reputation scores may indicate a talent for explanation and for clever writing more than an ability to translate knowledge in to code.

5 CONCLUSIONS AND FUTURE WORK

The authors have shown a correlation between age and SO reputation, which may indicate that programming knowledge can be maintained at a high level in to a person's 50's and 60's. The result shows that age have a positive effect, older SO users can acquire additional knowledge.

This has implications for career planning. As an individual programmer, awareness of high performance on the part of others can inspire efforts to continue or to improve.

As for future work, it will be important to investigate how SO measures, such as finding methods to translate user reputation and question scores into programming knowledge and ability.

REFERENCES

[1] Alberto Bacchelli. 2013. Mining challenge 2013: Stack overflow. In *The 10th Working Conference on Mining Software Repositories*.

[2] David F Hultsch, Christopher Hertzog, Brent J Small, and Roger A Dixon. 1999. Use it or lose it: engaged lifestyle as a buffer of cognitive decline in aging? *Psychology and aging* 14, 2 (1999), 245.

[3] P. Morrison and E. Murphy-Hill. 2013. Is programming knowledge related to age? An exploration of stack overflow. In *2013 10th Working Conference on Mining Software Repositories (MSR)*. 69–72. https://doi.org/10.1109/MSR.2013.6624008

[4] US Department of Labor and Bureau of Labor Statistics. 2011. *Occupational Outlook Handbook ch. Computer Software Engineers and Computer Programmers*. JIST Publishing.