Automatic Mining and Summarization of Crowd Sourced Software Knowledge

Speaker: **Gias Uddin**

Assistant Professor, Department of Electrical and Computer Engineering, University of Calgary, Canada

Gias Uddin is a tenure-track Assistant Professor at the University of Calgary from July 2020. Prior to that he was a Senior Data Scientist at the Bank of Canada and a software engineer at IBM Watson Analytics. He received a PhD from McGill University and a Master's from Queen's University (Canada).

He has been working in the industry (software startup, software consulting firm, IBM, and Banking) since 2008, in parallel to doing his PhD. In various increasingly senior roles in the industry, Gias Uddin has designed and developed software using innovative machine learning techniques. The software are deployed successfully in production and are being used by thousands/millions of users. His research focuses on AI (Artificial Intelligence) for SE (Software Engineering) that spans both applied and empirical software engineering based on the application of machine learning, artificial intelligence, and natural language processing techniques to the domain of software engineering. He has published more than 20 papers in peer-reviewed topmost conferences and journals in software engineering. His recent paper at the 32nd IEEE/ACM Automated Software Engineering Conference (ASE) was nominated for a best paper award.
Website: `https://giasuddin.ca/`

Abstract

Software is ubiquitous. We rely on software to drive the economy and to improve quality of life. As such, there is now an increasing need to educate software developers to produce software quickly but efficiently. Unfortunately, my surveys of 330 IBM developers found that the official learning resources of software can be often incomplete, obsolete, and incorrect. In subsequent surveys of 178 software developers, I found that developers rely on online software forums to compensate for the shortcomings in official software documentation, but the huge volume and scattered nature of crowd-sourced software forums present significant challenges to get quick, concise and correct insights. Indeed, among the numerous online crowd-sourced platforms for developers, Stack Overflow alone has over 120 million posts with over 11 million registered users. However, in a study of thousands of C++ code examples shared in Stack Overflow, we observed that the code examples can contain critical security vulnerabilities. I have developed Opiner, an online engine to search and summarize reviews (i.e., positive and negative opinions) about APIs (Application Programming Interfaces) from online developer forum. APIs are interfaces to reusable software libraries. To support API documentation from crowd-sourced data with information about code quality, Opiner provides summaries of the shared API usage examples. The summaries consist of code examples along with reviews to inform of their quality attributes. We observe that our proposed domain-specific summarization algorithms are significantly more useful than off-the-shelf techniques. In seven user studies, we find that Opiner positively impacts the productivity of developers to learn and use APIs for diverse development tasks. Opiner website is available online and is visited by developers across the world. I conclude by outlining my research vision, around my two long-term goals:

1. Summarization of Big Crowd Sourced Software Knowledge. I study the diverse nature of contents shared in crowd-sourced developer platforms and design tools and techniques to automatically and efficiently aggregate (i.e., collect and summarize) those contents using machine learning, text analysis and software analytics.

2. Analyzing Trust of Crowd-Source Software Knowledge. I study the quality of software usage information shared in online developer forums and design techniques to recommend the best usage practices.

When: Wednesday, 29th July, 2020 (8:00 PM)
Where: Online. Zoom Meeting ID: 637 1057 9182, Password: 247994
Organized by BUET ACM Chapter, Dept. of CSE, BUET.