Ten years of hunting for similar code for fun and profit (keynote)

Stephane GLONDU
INRIA

Lingxiao JIANG
Singapore Management University, lxjiang@smu.edu.sg

Zhendong SU
ETH Zurich

Citation

Ten Years of Hunting for Similar Code for Fun and Profit (Keynote)

Stéphane Glondu
Inria, Nancy Grand Est
France
stephane.glondu@inria.fr

Lingxiao Jiang
Singapore Management University
Singapore
lxjiang@smu.edu.sg

Zhendong Su
ETH Zurich
Switzerland
zhendong.su@inf.ethz.ch

ABSTRACT

In 2007, the Deckard paper was published at ICSE [1]. Since its publication, it has led to much follow-up research and applications. The paper made two core contributions: a novel vector embedding of structured code for fast similarity detection, and an application of the embedding for clone detection, resulting in the Deckard tool. The vector embedding is simple and easy to adapt. Similar code detection is also fundamental for a range of classical and emerging problems in software engineering, security, and computer science education (e.g., code reuse, refactoring, porting, translation, synthesis, program repair, malware detection, and feedback generation). Both have buttressed the paper's influence.

In 2018, the Deckard paper received the ACM SIGSOFT Impact Paper award. In this keynote, we take the opportunity to review the work's inception, evolution and impact on its subsequent work and applications, and to share our thoughts on exciting ongoing and future developments.

CCS CONCEPTS

• **Information systems** → **Clustering**; **Nearest-neighbor search**; **Similarity measures**; • **Software and its engineering** → **Software maintenance tools**; **Reusability**; *Search-based software engineering*;

KEYWORDS

code vectorization, code similarity, code search, code learning

REFERENCES

[1] Lingxiao Jiang, Ghassan Misherghi, Zhendong Su, and Stephane Glondu. 2007. DECKARD: Scalable and Accurate Tree-Based Detection of Code Clones. In *Proceedings of the 29th International Conference on Software Engineering (ICSE '07)*. IEEE Computer Society, Washington, DC, USA, 96–105. https://doi.org/10.1109/ICSE.2007.30

BIOGRAPHIES
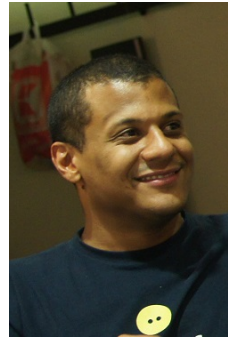
Stéphane Glondu is a research engineer at Inria Nancy. He received his PhD in Computer Science from Université Paris Diderot in 2012, and a Master's degree in Computer Science from ENS Cachan in 2007. His research spans programming languages and logic, formal methods, cryptographic protocols and more specifically voting systems. He has been involved in the development of the Coq proof assistant. He now develops the Belenios online voting system.

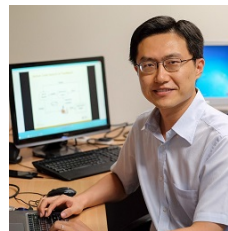**Lingxiao Jiang** is an Associate Professor in the School of Information Systems at Singapore Management University. His research focuses on software analysis and mining, exploring the combination of program analysis and machine learning techniques for software engineering problems. He has been working on various program representations for code similarity measurement, refactoring, code search, and automated testing and debugging. He received his PhD in Computer Science from University of California, Davis in 2009, and a Master's degree in Applied Mathematics and a Bachelor's degree in Information Science from the School of Mathematical Sciences at Peking University in 2003. He also had working experience as a test strategist at Nvidia Corporation before joining SIS at SMU.

Zhendong Su is a Professor in Computer Science at ETH Zurich. Previously, he was a Professor in Computer Science and a Chancellor's Fellow at UC Davis. He received his PhD in Computer Science from UC Berkeley. His research spans programming languages and compilers, software engineering, computer security, deep learning and education technologies. His work was recognized by an ACM SIGSOFT Impact Paper Award, a Google Scholar Classic Paper (2017) Award, multiple best/distinguished paper awards at top venues, an ACM CACM Research Highlight, an NSF CAREER Award, a UC Davis Outstanding Faculty Award, and multiple industrial faculty awards. He serves on the steering committees of ISSTA and ESEC/FSE, served as an Associate Editor for ACM TOSEM, co-chaired SAS 2009, program chaired ISSTA 2012, and program co-chaired SIGSOFT FSE 2016.