**Summary of the Paper: Detecting Similar Repositories on GitHub**

Yun Zhang1, David Lo2, Pavneet Singh Kochhar2, Xin Xia1 , Quanlai Li3, and Jianling Sun1 1College of Computer Science and Technology, Zhejiang University, Hangzhou, China 2School of Information Systems, Singapore Management University, Singapore 3University of California, Berkeley, USA

DOI: [https://xin-xia.github.io/publication/saner17.pdf](https://xin-xia.github.io/publication/saner17.pdf#:~:text=In%20this%20paper%2C%20we%20propose%20a%20novel%20approach,files%29%20which%20are%20not%20considered%20in%20previous%20works.)

Team members: Goian Sergiu-Rares

MetaSim is a search engine created to retrieve similar GitHub repositories beyond traditional keyword based searches. Existing search techniques often rely on simple text-matching methods, which fail to capture deeper semantic and contextual relationships between repositories. To address this, MetaSim leverages repository metadata (name, description, topics), content analysis (programming languages used) and 3 heuristics (readme files similarity, user starring behavior, temporal trends in starring). A main improval is the use of Repo2Vec, a vector based embedding technique inspired by Word2Vec to represent repositories as numerical vectors. These embeddings enable more accurate similarity calculations using cosine similarity, allowing MetaSim to identify repositories with related functionalities, even across different programming languages.

The aim of this app is to enhance Github repositories search integrating rich metadata beyond just source code analysis. Unlike existing approaches, which primarily focus on code similarity, MetaSim integrates both textual and behavioral data to improve retrieval accuracy. Its core features include Meta Similarity, which ranks repositories based on cosine similarity of their embeddings, and Meta Representation, which converts repository metadata into vectorized formats using Repo2Vec. By considering factors such as readme content similarity, shared interests of users who star repositories, and time-based patterns in starring activity, MetaSim provides a more intelligent and context-aware search experience.

To evaluate MetaSim’s efectiveness, the authors conducted a series of experiments comparing its retrieval accuracy against GitHub’s built-in search functionality. The validation process involved benchmarking MetaSim’s results against ground truth datasets of known related repositories. The study demonstrated that MetaSim significantly outperforms traditional keyword based searches showing higher precision and recall in retrieving functionally similar repositories. Additionally, qualitative assessments indicated that MetaSim's heuristic-driven approach provides more relevant repository recommendations, supporting its utility for developers, researchers, and open-source contributors.