Mining the Technical Skills of Open Source Developers

João Eduardo Montandon[1], **Marco Túlio Valente**[1]

[1] Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)

{joao.montandon,mtov}@dcc.ufmg.br

Abstract. *Software is "eating the world" as we witness the rise of companies whose business model is totally centered on software. The successful implementation of these systems heavily depends on the quality and expertise of their software development teams. However, software-based companies are facing an increasing software developers shortage issue. On the one hand, technical recruiters are increasingly relying on the information provided by Social Coding Platforms (SCPs)—e.g., GitHub, Stack Overflow, etc—to prospect new talent. On the other hand, the large volume of data available force job recruiters to only assess superficial information of their candidates. In order to tackle this problem, we described in the thesis an extensive investigation of methods and techniques to identify the technical skills of software developers based on their activity in SCPs. We organized the thesis in three major working units, where we first investigated the most demanded technical and soft skills under the eyes of IT companies, and then assessed developers' technical skills from deep and broad perspectives. These studies resulted in contributions to both research and industrial communities.*

Resumo. *Software está "devorando o mundo" à medida em que testemunhamos o surgimento de novas empresas cujo modelo de negócios é totalmente centralizado em um sistema computacional. O sucesso da implantação desses sistemas depende, em grande medida, da qualidade e competência dos desenvolvedores responsáveis pela sua implementação. Contudo, as empresas de desenvolvimento de software estão enfrentando um problema crescente de escassez de profissionais de tecnologia. Se por um lado recrutadores têm dependido cada vez mais de plataformas de codificação social—tais como GitHub e Stack Overflow—para buscar novos talentos, por outro o grande volume de dados disponibilizados por essas plataformas inibe uma avaliação em profundidade de seus candidatos. Neste contexto, foi realizada nesta tese uma ampla investigação de métodos para identificar automaticamente habilidades técnicas de desenvolvidores de software. A pesquisa de doutorado foi organizada em três grandes trabalhos, onde primeiro se investigou as habilidades mais requisitadas por empresas de tecnologia na busca de novos profissionais. Em seguida, realizou-se dois estudos avaliativos com objetivo de identificar as habilidades técnicas de desenvolvedores em duas perspectivas: amplitude e profundidade. Os resultados obtidos a partir desses estudos levaram a contribuições significativas para pesquisadores e praticantes da comunidade de Engenharia de Software.*

1. Introduction

After 25 years of the invention of modern Internet technologies, software has "eaten the world" as we everyday observe the rise of companies centered on software systems. These systems have become increasingly complex artifacts, hence requiring new levels of specialization of their development teams. Therefore, software companies require their professionals to master several specific technologies so they can perform their daily work. This scenario is leading to a worldwide shortage of skilled software engineers.

To fulfill their open positions, companies look for candidates with deep knowledge in their main area but also with a broad understanding of the software development cycle as a whole. People carrying such a profile are known as T-shaped professionals and are highly valued due to their capacity of solving problems in a multidisciplinary environment. In this thesis, we study techniques to identify developers' technical skills from both broad and deep perspectives given their activity in Social Coding Platforms, i.e., GitHub.

2. Conducted Research

This thesis is composed of three major works. First, we study in more detail the side view of the skills required by IT companies when looking for new professionals. For this, we conducted a large-scale analysis of 20,000 job opportunities and revealed which hard and soft skills are demanded in 14 IT professional roles. We observed that programming languages are largely required, followed by libraries and frameworks.

In the second study, we mined developers' expertise from a deep perspective to predict their expertise level in third-party components. For this, we build a ground-truth containing activity-based data of 575 developers experts in three JavaScript libraries: *reactJS*, *socket.io*, and *mongodb*. In summary, we were able to produce clusters where the number of experts ranges from 65% to 75%.

The third study analyzed the expertise of software developers from a broad perspective by identifying their technical roles. We investigated the effectiveness of traditional machine learning strategies in classifying the competence of software developers in six popular technical roles: *Backend*, *Frontend*, *FullStack*, *Mobile*, *DataScience*, and *DevOps*. The proposed model reported high outcomes for precision (88%) and AUC (89%).

3. Thesis Impact

In our view, the thesis summarized in this manuscript has yielded a relevant impact on the software engineering community. For instance, one of these works was organically featured in Reddit's JavaScript forum, gathering the attention of its users; so far this paper has been visualized more than 1,8K times. From a research standpoint, the publications resulting from this thesis have been cited in more than 60 works so far, and their datasets have been downloaded more than 11K times. Lastly, this research was recognized among the best six theses by the committee of the XXXV Thesis and Dissertations Contest from the Brazilian Computer Society [Montandon and Valente 2022].

References

[Montandon and Valente 2022] Montandon, J. E. and Valente, M. T. (2022). Mining the technical skills of open source developers. In *Thesis and Dissertations Contest (CTD) - Brazilian Computer Society (SBC)*, pages 1–10.