WoC Based Developer's Research Profile

Nakul Sharma[1], Balasubramnium Vivekanandam[2], Eugenio Vocaturo[3]
[1] Linclon University College; [2] Linclon University College; [3] Linclon University College
Email ID [1]pdf.nakul@lincoln.edu.my,
[2]vivekanandam@lincoln.edu.my,
[3]eugenio.vocaturo@cnr.it

Abstract: WoC infrastructure has been employed for studying software evolution and supply chain analysis. Mining WoC can help discover various facts and patterns about both the developer as well as the repository. This work proposes the possibility of extracting developer's email's from WoC and passing it across to orchid and Scopus databases. The developer's information is first extracted from WoC using the metadata file. It is then segregated according to domain names of the email id. The present work also proposes a metric to determine the research aptitude of the developer. The present work can be extended to include other scientific databases as well.

Keywords: Developer; World Of Code (WoC); profile;scopus;

Introduction

The software developer profiles are created on different profiles. However there is little work done to explore the research profiles of developer existing in different databases. The developer can also endeavor to change their carrier path to be more academic focused. In academics, the research profiles across different databases have some importance attached to them.

The World of Code (WoC) provides a facility to conduct analysis of repositories as well software developer in their individual contributions. The world of code infrastructure integrates different hosting platform to provide analysis and meta-data information about repositories. The developer information is stored across different locations in WoC servers [1]. The developer related meta-data was extracted from WoC server consisting of following fields:-

1. Name
2. Email Addresses
3. Number of commits done

There exist different research profiles which determine the quality of research work undertaken by authors. There also exist several means of asserting the research aptitude of any individual. This includes the person's ability to communicate research outputs in standard scholarly journals or conferences. Scopus databases indexes large number of research articles and journals. Scopus research database

have certain criteria for including and excluding any journal or conference. Web of Science (WoS) have more strict criteria for including and excluding as their scrutiny.

The WoC infrastructure mainly focuses on analysis of repository information across different platforms. It does not include any interface for checking the research component within the developer's profile. Hence, it is imperative that a separate mechanism is evolved to extract the necessary information in order to ascertain the research profile of developer.

The developer's profile at different research databases is also a testimony of different types of collaboration conducted. The url website of all the resources used in this paper is given in Table-1.

Table 1. Websites of Each Resource

Name of Website	**Website's url**
World of Code [1]	https://worldofcode.org/
Scopus [2]	https://scopus.com

The present work extracted information related to developer's profile from the World of Code infrastructure. The research database of Scopus was searched for the corresponding profiles of each of the extracted developer. There were several issues faced while conducting this mapping of developer profile from WoC to Scopus database.

Related work

The existing related work is categorized into two different sections. The first section is regarding work done on the existing research databases and profiles.

The second section deals with work done in extracting or mapping developer's profile online.

WoC is software infrastructure which deals with study of software evolution and software based supply chain analysis. The infrastructure provides a facility to query large scale repositories along with the developer's information. The meta-data analysis is also made possible by this platform. The mapping between different repositories and related entities make it possible to raise appropriate queries easily [1].
The developer profile related research includes reference [1] [4] [5] [7] [8]. The Scopus and orchid profile information include reference [6], [2].

Section-1 Scopus API scrapping

Author's create a web application to access Scopus profile and create bibliometrics based on the data collected from API. The author's position appearing in Scopus is retrieved through web scrapping and API. The evaluation of the system is done through System Usability Scale (SUS) [2].

The Scopus and open Alex API's are queried for getting citation map of different authors [12]. This paper has a practical hands-on approach in dealing querying both these databases and getting the necessary information related to author's profile [12].

Section-2 Developer's Profile

The authors create a developer's portrait that is multidimensional. This was created using standard text mining, SCA techniques and web-based techniques. The evaluation of the system is done using case studies [4].

A developer's reputation profile is created by authors. This is accomplished by making use of WoC software infrastructure along with identity disambiguation. The evaluation of the system was done using case studies [5].

The authors created web application for researcher's network in Ecuador. This was done by using Data Mining techniques, CRISP-DM, Scrum techniques. The application also provided visualization of the researcher network [6].

The author's extract developer's skills and expertise in the research conducted. The authors make use of API's, mining, web scrapping in accomplishing this task. The artifact created also included the email addresses of developers [7].

The developer's characteristics as extracted in regard to the technical debt by the authors. The code maturity and different roles of developers are also taken into consideration. The proposed methodology was evaluated using empirical testing [8].

The authors extracted 5 developer's profile after performing ML techniques on the dataset extracted. The ML techniques used were dimensionality reduction and K-means to complete this task [9].

The author's proposed a semantic model and recommendation system to extract expert's specific information. The author's used time-based characterization of network to accomplish this task. The social and global aspects were also investigated by authors in global software development setting [10].

A dev2vec ML based model is developed by authors. It makes use of doc2vector for creating different representation about developer. The proposed method is evaluated using standard ML techniques [11].

Table-2 provides a summary of work undertaken on WoC project related to developer's profile. It also provides a summary of work done on mapping retrieving records from scopus and orchid profiles.

Table 2. Related Work

Reference	**Focused Area of Research**	**Specific Methodology**	**Artifact Created**	**Evaluation Metrics Employed**
[1]	Software Infrastructure, Supply chain analysis	-	World of Code	-
[2]	citation collection, API, Scopus, Bibliometrics, scientometrics	Author position retrieval using API, web scrapping	Web application for author position retrieval	System Usability Scale (SUS)
[4]	Developer portraying, OSS platforms	text mining, web-based analysis, SCA techniques	multi-dimensional portrait model	Case Study
[5]	Developer reputation, software ecosystem, identity clarity	WoC developer identity, identity disambiguation	Author IDs totaling 34 Million	case study
[6]	Researcher network, Ecuador, data mining, visualization	CRISP-DM, Scrum	web application for data querying about researchers	-
[7]	Doc2Vec,developer expertise, skill set, API, Mining	ML, web scrapping, repository mining	Skill Space of developers and their email addresses.	Hypothesis testing
[8]	Developer characteristics, Mining Software Repositories	Mining for developer specific attributes	Developer profile in relation to technical debt, code maturity, roles.	Empirical testing
[9]	Developer profile, code quality , ML techniques	Dimensionality reduction, K-means	5 developer profiles	-

[10]	expert detection, sematic and social network, global software development	temporal characterizatio n of network	semantic model, recommendation system	-
[11]	developer vector,doc2vec	doc2vec	Dev2vec	Standard ML evaluation techniques
[12]	Citation map, Bibliographic analysis, Bibliometric analysis	Scopus API, OpenAlex API	code for interacting with Scopus and OpenAlex	-

Method, Experiments and Results

The proposed methodology is categorized into two sub-sections. The first is data extractor section and second is data analysis section. Figure-2 shows both the sections:

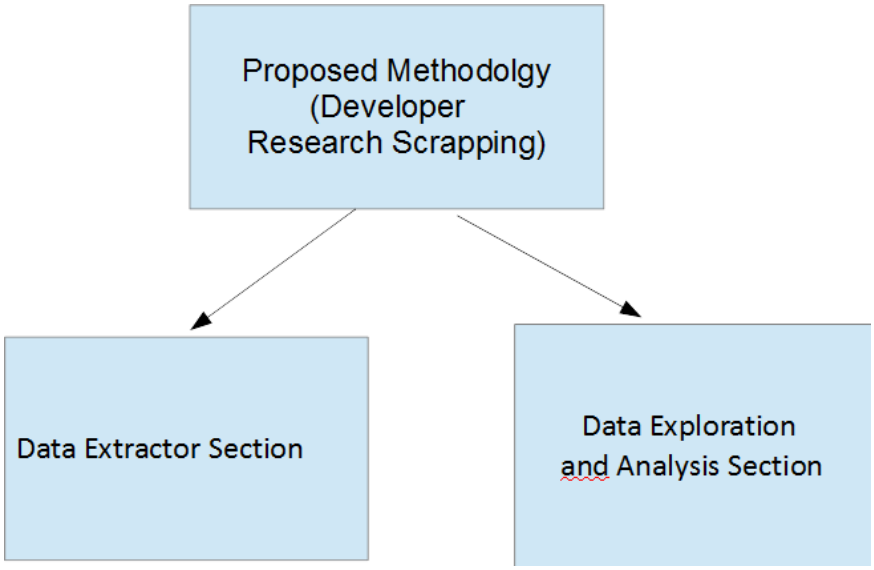

Figure-1 Developer Research Scrapping

Data Extractor Section

Data extractor module is presented in this section. The data extractor module finds following information from the WoC metadata file present in mongoDB file.

The data extractor module has following steps:-

1. Establishing connection with the WoC server
2. Extracting developer's name, email id, number of commits, from mongoDB server.

The data was extracted from the individual da5 server using mongodb queries. The WoC have several repositories stored to the format such that their analysis becomes easier. In this section, data is extracted from the Woc folder from da5 server. The results were analyzed and passed to Scopus database for checking the developer's research aptitude. The specific metrics is proposed also proposed to indicate the research aptitude of the developer.

Data Exploration and Analysis Section

In this section the extracted data is subjected to feature engineering to analyze the quality of development undertaken by the developer. The data consisted of names of the developer, their email ids and number of commits. The names were not clearly present in the extracted were removed from analysis. These included having special characters, numbers in the email ids and names.

Metric Definition

Proposed metric checks the developer's research profile at Scopus database. There are different measures that can be applied to check developer's research aptitude. A new metric is proposed for checking if the developer has published any Scopus database papers.

Let P be the number of paper's published by the developer D_i within Scopus. If count of papers in scopus database exceeds 0, then developer D_i is considered ActiveResearcher. Let devP be the total count of paper's published in scopus database in developer's profile. The categorization of developer is named as PDi.

Metric Formulation

if p>1
PD_i="Researcher"
else if p==0
PD_i="NotResearcher"

NumPapers is a column indicating the number of papers published by the developer.
The search was conducted using names of the developer and was further authenticated by their email addresses. In order to confirm if the orchid and scopus profile where actually of developer mentioned, an email was sent to all the author's searched.

Table 2. Results of applying cD_i *metric to mongodb dataset*

Total Number of mongodb entries scanned	**Count of** PD_i = *Researcher*	**Count of** PD_i*="NotResearcher"*	**Number of Ambiguous Developers**
66	14	40	12
%	21.21	60.60	18.18

Challenges faced while implementing metric are as follows:-

There were following issues faced while executing the methodology:-

1. Precise mapping from WoC to Scopus

There were issues of same name occurring with different affiliations. In such instances, the mapping was completed by checking the domain names of email address were matched. If the domain name was no-reply@github.com, then the name resolution and area of research was used to find the match.

2. Precise Name Resolution
There were issues related to first and last name of the developer's. This was resolved by first searching for the domain name and then manually mapping them to the correct entity.

Results and Discussions

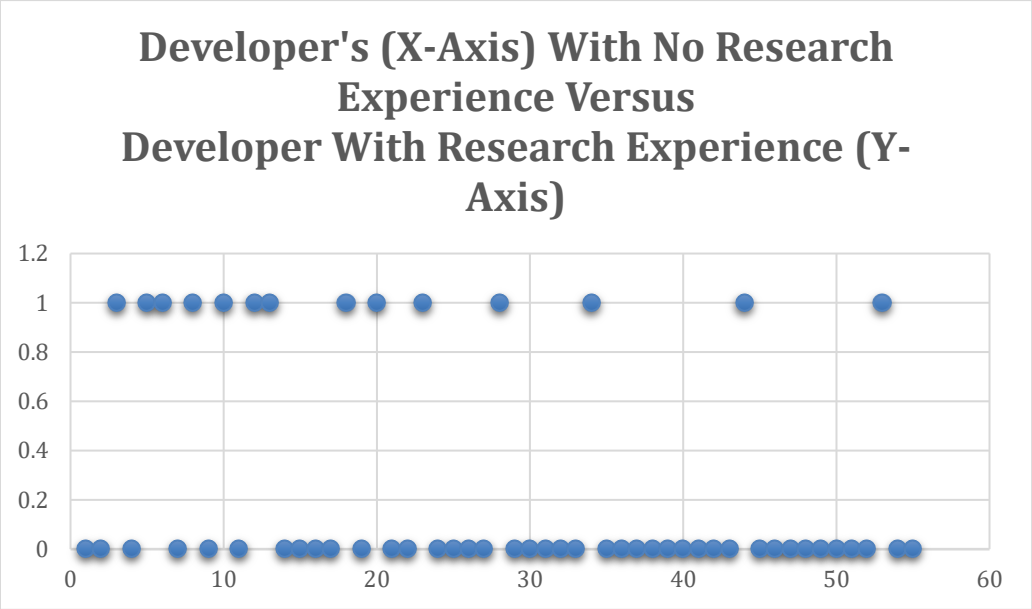

Figure-2 Visualization of Result

An assessment is done in this work about the developer's ability to be active in publishing research papers. The information was extracted from WoC infrastructure in order to ascertain the developer's name, email addresses. A metric was also proposed to categorize developer as researcher or not. This categorization can help in deciding different research roles which may be offered to developer's. The number of commits undertaken can also indicate the developer's reputation.

Conclusions

Scopus is one of the largest database indexing research papers and conferences. This paper proposed a metric to evaluate the developer presence in Scopus database. The developer information is extracted from WoC software infrastructure. A Scopus profile is created for any author who publishes paper in Scopus indexed conference or journal. The extracted names and email addresses were used to fire query on Scopus database for checking respective author's research profile. A scale was proposed to gauge the research aptitude of developer. A sample size of 66 developers were evaluated.

References

1. Y. Ma, C. Bogart, S. Amreen, R. Zaretzki and A. Mockus, "World of Code: An Infrastructure for Mining the Universe of Open Source VCS Data," *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*, Montreal, QC, Canada, 2019, pp. 143-154, https://doi.org/10.1109/MSR.2019.00031.
2. Rochim, Adian Fatchur, Tendi Nugeraha Wijaya, and Dania Eridani. "A citation data collector tool of author's profiles in scopus based on web and application programming interface (API)." In *IOP Conference Series: Materials Science and Engineering*, vol. 1077, no. 1, p. 012017. IOP Publishing, 2021.
3. "ORCHID", https://orcid.org/ accessed on 20.7.2025.
4. Yang, W., Pan, M., Zhou, Y., & Huang, Z. (2020). Developer portraying: A quick approach to understanding developers on OSS platforms. *Information and Software Technology*, *125*, 106336, https://doi.org/10.1016/j.infsof.2020.106336.
5. S. Amreen, A. Karnauch and A. Mockus, "Developer Reputation Estimator (DRE)," *2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, San Diego, CA, USA, 2019, pp. 1082-1085. https://doi.org/10.1109/ASE.2019.00107
6. Arias, Josué, and Lorena Recalde. "Search and visualization of researcher networks: co-authorship in Ecuador." In *Conference on Information and Communication Technologies of Ecuador*, pp. 448-463. Cham: Springer Nature Switzerland, 2023. https://doi.org/10.1007/978-3-031-45438-7_30
7. Dey, Tapajit, Andrey Karnauch, and Audris Mockus. "Representation of developer expertise in open source software." In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pp. 995-1007. IEEE, 2021.

8. Codabux, Zadia, and Christopher Dutchyn. "Profiling developers through the lens of technical debt." In *Proceedings of the 14th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pp. 1-6. 2020. https://doi.org/10.1145/3382494.3422172
9. González, Cristina Aguilera, Laia Albors Zumel, Jesús Antonanzas Acero, Valentina Lenarduzzi, Silverio Martínez-Fernández, and Sonia Rabanaque Rodríguez. "A preliminary investigation of developer profiles based on their activities and code quality: Who does what?." In *2021 IEEE 21st International Conference on Software Quality, Reliability and Security (QRS)*, pp. 938-945. IEEE, 2021. https://doi.org/10.1109/QRS54544.2021.00103
10. Lopes, Tales, Victor Ströele, Regina Braga, Jose Maria N. David, and Michael Bauer. "A broad approach to expert detection using syntactic and semantic social networks analysis in the context of Global Software Development." *Journal of Computational Science* 66 (2023): 101928. https://doi.org/10.1016/j.jocs.2022.101928
11. Dakhel, Arghavan Moradi, Michel C. Desmarais, and Foutse Khomh. "Dev2vec: Representing domain expertise of developers in an embedding space." *Information and Software Technology* 159 (2023): 107218. https://doi.org/10.1016/j.infsof.2023.107218
12. Harder, R. (2024). Using Scopus and OpenAlex APIs to retrieve bibliographic data for evidence synthesis. A procedure based on Bash and SQL. *MethodsX*, *12*, 102601. https://doi.org/10.1016/j.mex.2024.102601