Evolution of Crowd-Sourced Documentation in Developers Discord Conversations

Marco Raglianti
REVEAL @ Software Institute – Università della Svizzera italiana (USI), Lugano, Switzerland
marco.raglianti@usi.ch

I. Introduction

Software documentation is critical in several development activities [1]. Correct and up to date documentation is an important asset for every software developer [2]–[7]. Software documentation from heterogeneous sources (mailing lists, StackOverflow, issues, and pull requests) has been investigated to produce a taxonomy of issues [8]. The more adopted a language or framework is, the more documentation its users generate. Coding examples, idiomatic patterns, and common mistakes are discussed daily, usually in unstructured form, on instant messaging platforms such as Gitter, Slack, and Discord [9]–[17]. Their potential as knowledge and documentation sources remains largely unexplored. There is a lack of techniques and tools to reliably mine, index, and retrieve coherent content from such platforms.

II. About me

I'm Marco Raglianti, Ph.D. student under the supervision of Prof. Dr. Michele Lanza, in the the Reverse Engineering, Visualization, Evolution Analysis Lab – REVEAL – group at the Software Institute, Università della Svizzera italiana.

The focus of this work is the evolution of the *documentation landscape* of software systems [18] (*i.e.,* the faceted and heterogeneous multitude of information sources that can constitute software documentation). I investigated software developer communities and online interaction for collaborative development, especially in instant messaging applications. In the last two years I explored Discord conversations and the source code shared in public Discord servers providing tools to mine, analyze, and visualize this form of documentation. My main goal is to extend knowledge-mining capabilities to new potential sources of documentation that are still unexplored.

III. Presentation abstract

In my presentation I will start by introducing the context of mining software developers' conversations for Software Engineering research. I will give an interactive demo of DiscOrDance, the tool we developed to mine and analyze Discord communities (Figure 1). I will show insights from our case study: The Pharo Discord server. Source code is present and central to conversations in high-throughput & high-volatility instant messaging platforms [15]. Moreover, aggregating single messages in higher-level constructs (*e.g.,* conversations) has an effect on the granularity at which we can perform code evolution analysis [16].

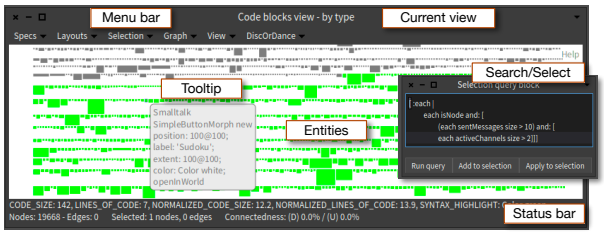

Fig. 1: The User Interface of DiscOrDance

When aggregating messages into conversations we need to solve the disentanglement problem [19]. I will present CoDi (Figure 2), our re-implementation of a state-of-the-art conversation disentanglement model as a micro-service architecture and user interface. It allows experimentation with a two-step disentanglement model [10], [20]–[22], visual analysis with a web interface, and batch processing of datasets with its RESTful API.

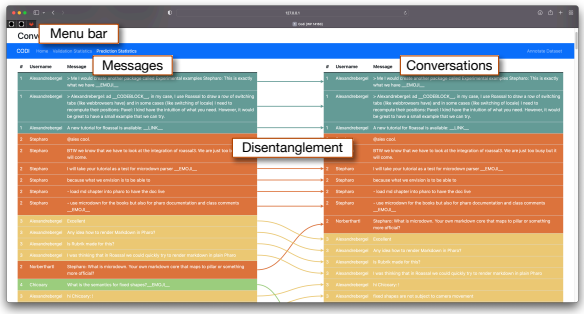

Fig. 2: The User Interface of CoDi

I will share our experience on challenges we faced in mining public Discord servers. Most challenges we are ill-prepared to face are on the frontier of social, legal, and ethical aspects. For example getting access for DiscOrDance to a Discord server from its administrators poses acceptance problems as well as a need for privacy policies of mined data. Discussion will also cover "false-positives" in terms of anonymization. For example, most Discord datasets can actually be "de-anonymized" with a simple search in Discord itself – and the process can be easily automated just slightly breaking Discord's Terms of Service.

Finally, I will present our ongoing work on mining GitHub to automatically extract the documentation landscape [18] of a software system.

REFERENCES

[1] E. Aghajani, C. Nagy, M. Linares-Vásquez, L. Moreno, G. Bavota, M. Lanza, and D. C. Shepherd, "Software documentation: The practitioners' perspective," in *Proceedings of ICSE 2020 (International Conference on Software Engineering)*. ACM, 2020, pp. 590–601.

[2] A. Forward and T. C. Lethbridge, "The relevance of software documentation, tools and technologies: A survey," in *Proceedings of DocEng 2002 (Symposium on Document Engineering)*. ACM, 2002, pp. 26–33.

[3] J.-C. Chen and S.-J. Huang, "An empirical analysis of the impact of software development problem factors on software maintainability," *Journal of Systems and Software*, vol. 82, no. 6, pp. 981–992, 2009.

[4] B. Dagenais and M. P. Robillard, "Creating and evolving developer documentation: Understanding the decisions of open source contributors," in *Proceedings of FSE 2010 (International Symposium on Foundations of Software Engineering)*. ACM, 2010, pp. 127–136.

[5] M. P. Robillard and R. DeLine, "A field study of API learning obstacles," *Empirical Software Engineering*, vol. 16, no. 6, pp. 703–732, 2011.

[6] G. Garousi, V. Garousi-Yusifoğlu, G. Ruhe, J. Zhi, M. Moussavi, and B. Smith, "Usage and usefulness of technical software documentation: An industrial case study," *Information and Software Technology*, vol. 57, pp. 664–682, 2015.

[7] I. Sommerville, *Software Engineering*, 10th ed. Pearson, 2015.

[8] E. Aghajani, C. Nagy, O. L. Vega-Márquez, M. Linares-Vásquez, L. Moreno, G. Bavota, and M. Lanza, "Software documentation issues unveiled," in *Proceedings of ICSE 2019 (International Conference on Software Engineering)*. IEEE/ACM, 2019, pp. 1199–1210.

[9] B. Lin, A. Zagalsky, M.-A. Storey, and A. Serebrenik, "Why developers are slacking off: Understanding how software teams use Slack," in *Proceedings of CSCW/SCC 2016 (Conference on Computer Supported Cooperative Work and Social Computing Companion)*. ACM, 2016, pp. 333–336.

[10] P. Chatterjee, K. Damevski, L. Pollock, V. Augustine, and N. A. Kraft, "Exploratory study of Slack Q&A chats as a mining source for software engineering tools," in *Proceedings of MSR 2019 (International Conference on Mining Software Repositories)*. IEEE/ACM, 2019, pp. 490–501.

[11] O. Ehsan, S. Hassan, M. E. Mezouar, and Y. Zou, "An empirical study of developer discussions in the Gitter platform," *Transactions on Software Engineering and Methodology*, vol. 30, no. 1, pp. 1–39, 2020.

[12] V. Stray and N. B. Moe, "Understanding coordination in global software engineering: A mixed-methods study on the use of meetings and Slack," *Journal of Systems and Software*, vol. 170, p. 110717, 2020.

[13] E. Parra, A. Ellis, and S. Haiduc, "GitterCom: A dataset of Open Source developer communications in Gitter," in *Proceedings of MSR 2020 (International Conference on Mining Software Repositories)*. ACM, 2020, pp. 563–567.

[14] L. Shi, X. Chen, Y. Yang, H. Jiang, Z. Jiang, N. Niu, and Q. Wang, "A first look at developers' live chat on Gitter," in *Proceedings of ESEC/FSE 2021 (European Software Engineering Conference and Symposium on the Foundations of Software Engineering)*. ACM, 2021, pp. 391–403.

[15] M. Raglianti, R. Minelli, C. Nagy, and M. Lanza, "Visualizing Discord servers," in *Proceedings of VISSOFT 2021 (Working Conference on Software Visualization)*. IEEE, 2021, pp. 150–154.

[16] M. Raglianti, C. Nagy, R. Minelli, and M. Lanza, "Using Discord conversations as program comprehension aid," in *Proceedings of ICPC 2022 (International Conference on Program Comprehension)*. ACM, 2022.

[17] K. M. Subash, L. P. Kumar, S. L. Vadlamani, P. Chatterjee, and O. Baysal, "DISCO: A dataset of Discord chat conversations for software engineering research," in *Proceedings of MSR 2022 (International Conference on Mining Software Repositories)*. ACM, 2022.

[18] M. Raglianti, "Topology of the Documentation Landscape," in *Proceedings of ICSE 2022 Companion (International Conference on Software Engineering Companion)*. ACM, 2022.

[19] D. Shen, Q. Yang, J.-T. Sun, and Z. Chen, "Thread detection in dynamic text message streams," in *Proceedings of SIGIR 2006 (International Conference on Research and Development in Information Retrieval)*. ACM, 2006, pp. 35–42.

[20] M. Elsner and E. Charniak, "Disentangling chat," *Computational Linguistics*, vol. 36, no. 3, pp. 389–409, 2010.

[21] P. Chatterjee, K. Damevski, N. A. Kraft, and L. Pollock, "Software-related Slack chats with disentangled conversations," in *Proceedings of MSR 2020 (International Conference on Mining Software Repositories)*. ACM, 2020, pp. 588–592.

[22] P. Chatterjee, "Extracting archival-quality information from software-related chats," in *Proceedings of ICSE 2020 Companion (International Conference on Software Engineering Companion)*. ACM, 2020, pp. 234–237.