DOI 10.1109/CESI.2017.6

Empirical Studies in Question-Answering Systems: A Discussion

Jacob Krüger*‡, Ivonne Schröter†‡, Andy Kenner†, Thomas Leich*
* Harz University of Applied Sciences Wernigerode, Germany
{jkrueger, tleich}@hs-harz.de
† METOP GmbH Magdeburg, Germany
{ivonne.schroeter, andy.kenner}@metop.de
‡ Otto-von-Guericke-University Magdeburg, Germany

***Abstract*—Researchers perform empirical studies in industry to gain qualitative insights into a real-world problem. However, common critics are the diversity and selection process of participants. To address these issues, we propose to improve the integration of question-answering systems into empirical study. In this paper, we *i)* describe approaches to conduct studies in such systems, *ii)* exemplify corresponding challenges, and *iii)* discuss their potential. We illustrate the approaches on existing works in which they were partly implemented to show that they can succeed.**

***Keywords*-Question-Answering Systems, Empirical Research, Stackoverflow.com**

I. Introduction

Researchers rely on empirical studies based on industrial interviews to survey qualitative data they cannot measure [6]. However, conducting such studies is a challenging task and may fail for several reasons, such as lacks of trust or missing social skills [5, 13, 16]. Furthermore, reviewers often criticize and reject empirical studies based on industrial settings. Torchiano and Ricca [13] report several points of criticism that resulted in a rejection, two of them being:

- An unsuited selection method or unrepresentative sample of participants may result in a *sampling bias*.
- Empirical studies may not be generalized because of a *limited geographical scope* of the participants.

As both points address the participants of a study, we argue that a broader audience can help to reduce these biases. Therefore, we propose to better integrate community question-answering (CQA) systems [12] into methods for empirical studies. In contrast to on-site studies, with such systems researcher can target a broader pool of participants from all over the world.

CQA systems provide a basis for general discussions but can also be highly specialized, as we show in Table I. For example, Stack Overflow[1] focuses on programming related questions that are answered by a diverse community, including industrial participants. Srba and Bielikova [12] systemically review research methods on CQA systems and identify three main approaches: *Exploratory studies*, *content and user modelling*, and *adaptive support methods*. While CQA systems are analyzed in academia (see Table I) and companies use them, for instance to communicate with their customers [11], they are currently rarely considered for empirical studies. We argue that their wide acceptance can help to achieve broader and more diverse participants. Thus, we can address the aforementioned problems.

[1] http://stackoverflow.com/, 13.01.2017

TABLE I
Examples of CQA systems.

Name	Topics	Example Study
Stack Overflow	Software development	Squire [11]
Lambda the Ultimate	Programming languages	Okon and Hanenberg [9]
Quora	Many	Wang et al. [15]
Yahoo! Answers	Many	Agichtein et al. [1]
Reddit	Many	Okon and Hanenberg [9]

In this paper, we sketch a vision of methods and open issues to use CQA systems for empirical research. To support our arguments, we review existing literature similar to Srba and Bielikova [12]. However, we aim not to summarize or categorize all found articles but discuss empirical research with CQA systems. More precisely, we contribute the following:

- We propose three approaches on how to integrate CQA systems to conduct or support empirical studies. This provides an initial set of solutions and can be used and extended by the research community.
- We overview challenges that have to be addressed when conducting such studies. Our goal is to improve the awareness and initiate solutions for potential problems.
- We discuss our vision to illustrate its potential benefits and pitfalls. While the challenges are directly connected to conducting a study, this discussion focuses on a more general level and problems that we have to consider in advance.

The remaining paper is structured as follows. In Section II, we propose our ideas on conducting empirical studies with CQA systems. We describe challenges of the proposed approaches in Section III and discuss our general concerns in Section IV. Finally, we conclude in Section V.

II. Empirical Studies in CQA Systems

In this section, we discuss ideas on how to use CQA systems for empirical studies. To support our argumentation,

we performed a literature search in the SCOPUS[2] database and identified corresponding articles. We limited our search to articles that name Stack Overflow and state that they perform a study in their title, abstract, or keywords. Furthermore, we considered only the years 2015 and 2016, the period which Srba and Bielikova [12] do not cover in their review. Hence, our search string was:

> *TITLE-ABS-KEY (stackoverflow.com OR stack overflow) AND TITLE-ABS-KEY (study) AND PUBYEAR > 2014*

We received 62 articles from which we found five to illustrate our ideas. These results might be biased as we only consider a single CQA system. However, Stack Overflow is an established platform for software engineers and is well suited to support our discussion. In the following, we describe three different approaches to support empirical studies with CQA systems.

A. Mining Existing Data

A first approach to utilize CQA systems is to mine them for information. The idea is to rely on existing questions and answers that industrial participants provide and which are connected to the conducted study. This might be used to determine an initial hypothesis, gain supportive data, assess findings, or even to conduct a full study. Mining data from software repositories and CQA systems is a common technique for several approaches, for example to assess answer quality or identify a user's context [12]. We argue that such exploratory approaches are a first step towards empirical studies. However, empirical research focuses less on analyzing meta-data or behavior but the actual answers to a question. Thus, we need to adopt existing methods and scope them for empirical studies.

One example for this is the work of Venkatesh et al. [14], who analyzed 92.471 discussions on Stack Overflow to gain information on the usage and development of Web APIs. Their goal was to identify the most important topics among API developers and how these evolve. Conceptually, the authors apply our idea and mine all information for their study in a CQA system. Otherwise, they would have to directly question developers of Web APIs, which might be problematic due to locality and a limited sample size.

Based on another study, De Rosso and Jackson [4] developed Gitless, a system to overcome shortcomings of Git. For this, the authors conducted a study on Stack Overflow questions that discussed problems of Git. With their results, De Rosso and Jackson implemented the new Gitless system. Finally, they preformed a user study, but not on Stack Overflow, to evaluate whether they solved the identified problems.

B. Asking New Questions

The previous approach relies on existing data that addresses the considered research question. This will most likely not always be the case. Furthermore, it might be necessary to gather more details than existing answers provide. Hence, researcher have to ask new questions by themselves to conduct their study. We argue that this approach is already more complex and challenging than mining existing data. Questions have to be defined in a way that avoids misunderstandings and encourages answers, which can be difficult.

For instance, Okon and Hanenberg [9] searched code examples for an experiment on dynamic-typed and static-typed languages. They first tried to mine existing data but did not find a suitable answer. Hence, they decided to post an own question in several CQA systems, including Stack Overflow. While this perfectly illustrates our idea, the authors received only few answers. Moreover, code examples were rarely posted and Okon and Hanenberg found only one to be according to their question.

C. Identifying Participants

Finally, we propose to use CQA systems to only identify the participants of an empirical study. This can be necessary or helpful for several reasons, such as:

- The researcher can ensure that the participants have a specific, for example industrial, background.
- It is possible to perform qualitative interviews with detailed questions.
- Direct feedback allows participants to ask questions by themselves, for instance to clarify a statement.

This approach requires considerably more effort than the previous two: Instead of only gathering existing or aiming to conduct new data, researchers have to identify participants and afterwards conduct quasi on-site studies with them. However, while this requires additional effort, researchers can improve the quality of their study due to direct interactions.

Coleman and Lieberman [3] applied this idea on Stack Overflow. Their goal was to identify what motivates participants in this CQA system to share their knowledge and how reputation systems affect them. For this, they gathered contributors of Stack Overflow, using the snowball method. They then conducted semi-structured interviews with four participants.

D. Combinations

The aforementioned approaches can help to scope methods and define research directions. However, they can hardly be separated at any occasion and combinations seem to be useful or even unavoidable. For instance, mining existing data can provide an impression of existing questions and answers. Based on the findings, a researcher may define his own question or identify suitable participants.

One example for such a combination is the work of Squire [11]. She gathered a set of 20, including industrial, projects that moved their developer support to Stack Overflow. However instead of asking questions, the author collected data from the CQA system to measure two metrics. These metrics were used to empirically evaluate the level of participation and the response times of developers (the participants of the study) before and after they changed towards Stack Overflow. Hence,

[2]https://www.scopus.com, 13.01.2017

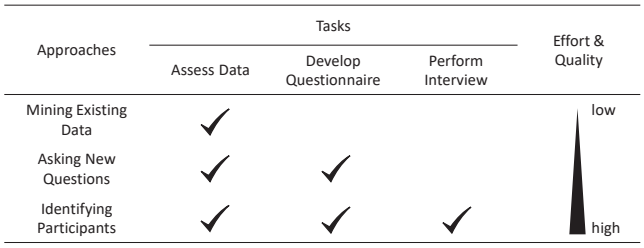

Fig. 1. Proposed approaches and tasks performed in these.

this resembles a combination of *Identifying Participants* and *Mining Existing Data*.

E. Effort and Quality

The examples we found also hint at dependencies between the proposed categories. As we illustrate in Figure 1, the effort of conducting an empirical study increases from mining of existing data towards identifying participants. This is due to the additional tasks that a researcher has to do: Instead of assessing mined data, researchers have to additionally develop questionnaires or perform interviews. However, we also argue that the quality increases correspondingly. Direct contact allows researchers to ask more detailed questions and better assess a participants expertise.

In this section, we described three different approaches to support empirical studies in CQA systems and discussed their dependencies. Which approach is suitable for a study is difficult to answer and has to be reasoned for the specific situation and task ahead.

III. Challenges

While we reflected on the described approaches and examples, we identified several issues that have to be addressed. In this section, we will discuss the initial challenges that may hamper performing and accepting industrial studies in CQA systems.

A. Selecting CQA Systems

As we illsutrate in Table I, there exist several CQA systems that have more general or rather specific purposes [12]. It is essential to identify these systems and assess for which topics they are suited and how the community composes. Stack Overflow, which we considered, is specialized on programming tasks and attracts developers from industry, academia, and open-source projects. However, there might be systems that are better scoped for industrial settings. An according overview can support researchers in selecting the best platform for their study.

B. Participants

Considering the participants of CQA systems, we have to address several points. These are also current research fields, as Srba and Bielikova [12] show, with some examples being:

- It is important to understand how to motivate contributors to answer a question or participate in a study. Otherwise, a significant sample size may not be achieved or the results could be useless. This might be achieved based on an existing reputation system or rewards.
- Before conducting the study, researchers have to assess the background and expertise of the participants. We need to separate, for instance, between industrial developers and students. While both groups can be helpful to answer a research question, we cannot mix their results of an empirical study. Furthermore, a researcher has to determine the knowledge and expertise of his participants to rate their competence on the asked questions.
- Suitable sample strategies have to be developed to reduce biases and methodologically select participants. Otherwise, a sampling bias could mean that the study is not representative and will be rejected. Still, a sampling is necessary to remove unsuited participants, for instance, due to missing background or expertise.

We are aware that these are only examples of potential issues regarding the participants. Still, they provide a glimpse on the corresponding challenges.

C. Quality

For empirical studies it is essential to guarantee a certain quality. In CQA systems, we have to assess qualities for two purposes.

Firstly, the quality of questions must be assessed. An unclear question may lead to misunderstandings and wrong answers. This can not only affect existing data but also questions posted by researchers. In particular, a bad question may prevent any answers of competent developers. Based on the experiences reported by Okon and Hanenberg [9], it seems important to clearly describe the question and state its purpose. Experiments in this direction can help to identify suitable question styles.

Secondly, it is important to assess the quality of answers [1, 15]. While Okon and Hanenberg [9] received more than 90 answers, only three of those addressed the actual question and only one answered it. There might be several reasons for this few correct responses, for example, because the contributors misunderstood the question or did not have the expertise. Still, this illustrate the importance of assessing an answer's quality.

D. Documentation

Finally, we emphasize the importance of documentation and guidelines. It seems necessary to develop suitable guidelines that adopt existing ones, for instance by Kitchenham et al. [7], for CQA systems. This way, a uniform and reliable methodology can be implemented. While reading the example papers, we found that such a method was missing. As a result, some details of the approaches were hidden in the text. Mainly, we argue that detailed investigations of potential *threats to validity* (e.g., which participants are sampled or how to validate their answers) are essential and differ highly from on-site studies [10]. Thus, it seems problematic to understand or replicate empirical studies until according guidelines exist.

Overall, there are several challenges the research community has to address. In this section, we discussed an initial set of those to provide a starting point.

IV. Discussion

In this section, we discuss our vision on a more general level. Therefore, we rely on our experiences and problems we had during other and ongoing empirical studies and that also affect the usage of CQA systems. We reason on the potential pitfalls that come with the described approaches.

Conducting empirical research is a time consuming task with several limitations [2, 5, 7, 13, 16]. Especially in industrial settings it is problematic to cooperate and still determine good results. Because the critic addresses mainly which participants are selected, we argue that a better integration of CQA systems can improve this situation. However, the described approaches face the same, as well as new, challenges as any empirical study. Thus, we are aware that we need to apply our ideas several times before we can provide a detailed and more complete overview on how to improve this integration.

Furthermore, it will be interesting to which extend CQA systems accept empirical studies. Conducting to many studies may lead to the community leaving. In this situation, the provider of the CQA system will most surely object against further research activities. Hence, in advance we need to determine how to implement studies without repelling the community.

An important aspect, which we discussed in Section III, is the motivation of participants. Contributors of CQA systems are often driven by intrinsic motivation and point systems [8] rather than extrinsic stimuli. Thus, empirical studies in this context need to address the same motivation and are ideally integrated into the system. For this, coordinating with the system's provider might be a suitable approach.

Finally, we want to call attention on ethical aspects, which we did not consider until now. Still, there are some open issues that we have to address and communicate to the participants. For instance, it is important to ensure that we can guarantee anonymity, which requires further research. This is especially important in industrial context where information protection often hampers empirical studies. Another question is, whether community members should become part of a study without their acceptance. While this is almost certain when we simply mine data, it seems to be rather questionable. Thus, we need further investigations to which extend this should be possible and to decide whether specific methods should be applied or discarded. In our opinion, this is also up to discussion for analyses of any community on which research is reported.

Within this section, we argued that the proposed approaches can resolve some problems of empirical studies. However, we also discussed some aspects that can be challenging to address and may hamper the implementation of our ideas.

V. Conclusions

In this paper, we proposed to better integrate CQA systems into empirical studies by contacting a broader audience. For this, we sketch three different approaches and several challenges that we have to address. While we argue that our ideas can help to improve studies, we also discussed shortcomings that may lead to failure.

During our future work, we aim to apply our approaches to conduct empirical studies and report on the results. Furthermore, adopted guidelines, suitable methods, and a catalog of CQA systems will be helpful.

Acknowledgment

This research is supported by DFG grant LE 3382/2-1 and Volkswagen Financial Services AG.

References

[1] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding High-Quality Content in Social Media. In *WSDM*, pages 183–194. ACM, 2008.

[2] M. A. Babar and H. Zhang. Systematic Literature Reviews in Software Engineering: Preliminary Results from Interviews with Researchers. In *ESEM*, pages 346–355. IEEE, 2009.

[3] E. Coleman and Z. Lieberman. Contributor Motivation in Online Knowledge Sharing Communities with Reputation Management Systems. In *SAICSIT*, pages 1–12. ACM, 2015.

[4] S. P. De Rosso and D. Jackson. Purposes, Concepts, Misfits, and a Redesign of Git. *SIGPLAN Notices*, 51(10):292–310, 2016.

[5] P. Grünbacher and R. Rabiser. Success Factors for Empirical Studies in Industry-Academia Collaboration: A Reflection. In *CESI*, pages 27–32. IEEE, 2013.

[6] S. E. Hove and B. Anda. Experiences from Conducting Semi-Structured Interviews in Empirical Software Engineering Research. In *METRICS*, pages 23–32. IEEE, 2005.

[7] B. A. Kitchenham, S. L. Pfleeger, L. M. Pickard, P. W. Jones, D. C. Hoaglin, K. El Emam, and J. Rosenberg. Preliminary Guidelines for Empirical Research in Software Engineering. *IEEE Transactions on Software Engineering*, 28(8):721–734, 2002.

[8] K. K. Nam, M. S. Ackerman, and L. A. Adamic. Questions in, Knowledge in?: A Study of Naver's Question Answering Community. In *CHI*, pages 779–788. ACM, 2009.

[9] S. Okon and S. Hanenberg. Can We Enforce a Benefit for Dynamically Typed Languages in Comparison to Statically Typed Ones? A Controlled Experiment. In *ICPC*, pages 1–10. IEEE, 2016.

[10] J. Siegmund, N. Siegmund, and S. Apel. Views on Internal and External Validity in Empirical Software Engineering. In *ICSE*, pages 9–19. IEEE, 2015.

[11] M. Squire. "Should We Move to Stack Overflow?" Measuring the Utility of Social Media for Developer Support. In *ICSE*, pages 219–228. IEEE, 2015.

[12] I. Srba and M. Bielikova. A Comprehensive Survey and Classification of Approaches for Community Question Answering. *ACM Transactions on the Web*, 10(3):18:1–18:63, 2016.

[13] M. Torchiano and F. Ricca. Six Reasons for Rejecting an Industrial Survey Paper. In *CESI*, pages 21–26. IEEE, 2013.

[14] P. K. Venkatesh, S. Wang, F. Zhang, Y. Zou, and A. E. Hassan. What Do Client Developers Concern When Using Web APIs? An Empirical Study on Developer Forums and Stack Overflow. In *ICWS*, pages 131–138. IEEE, 2016.

[15] G. Wang, K. Gill, M. Mohanlal, H. Zheng, and B. Y. Zhao. Wisdom in the Social Crowd: An Analysis of Quora. In *WWW*, pages 1341–1352. ACM, 2013.

[16] C. Wohlin. Empirical Software Engineering Research with Industry: Top 10 Challenges. In *CESI*, pages 43–46. IEEE, 2013.