The Journal of Systems & Software 144 (2018) 439–449

Contents lists available at ScienceDirect

The Journal of Systems & Software

journal homepage: www.elsevier.com/locate/jss

A Systematic Mapping Study driven by the margin of error

Tomaž Kosar*[,a], Sudev Bohra[b], Marjan Mernik[a]

[a] University of Maribor, Faculty of Electrical Engineering and Computer Science, Koroška cesta 46, Maribor 2000, Slovenia
[b] Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213-3890, USA

ARTICLE INFO

Keywords:
Software engineering
Systematic review
Systematic Mapping Study
Reliability
Margin of error

ABSTRACT

Until recently, many Systematic Literature Reviews (SLRs) and Systematic Mapping Studies (SMSs) have been proposed. However, when SMS is performed on a broad topic with a large amount of primary studies, the cost of assessment of all primary studies requires unjustified resources. In this paper, a new approach is introduced for performing SMSs, called SMS driven by the margin of error. The main objective of the described work was to decrease the assessment cost of primary studies by stopping the process of classification of primary studies when enough evidence has been collected. We introduced a statistical approach with random sampling and a margin of error into the design of SMSs when a topic under discussion is broad with a large number of primary studies. In this paper, SMS driven by the margin of error was applied on three different use cases: SMS on Domain-Specific Languages, SMS on Template-based Code Generation, and SMS on Software Reliability Modeling, where it was shown that the proposed approach reduced the cost of assessing primary studies and quantified the reliability of SMS.

1. Introduction

This paper proposes a new method for performing Systematic Mapping Studies (SMSs) suitable when the number of identified primary studies is very large and their classification would require enormous resources. SMS is a secondary study, a special form of Systematic Review (SR), that reviews primary studies with the aim of synthesizing evidence on a research topic (Kitchenham and Charters, 2007). Yet another form of SR is a Systematic Literature Review (SLR), with the aim to identify, analyze and interpret all available evidence on specific research questions. The differences between SMS and SLR are subtle, but important. A good discussion on the differences between SMS and SLR can be found in Kitchenham et al. (2011). In brief, SLR includes a more comprehensive and thorough investigation of primary studies while pursuing more specific research questions with high requirements of research synthesis (Cruzes and Dybå, 2011). Furthermore, a quality assessment of the primary studies is necessary during SLR. According to the guidelines in Kitchenham and Charters (2007), all relevant studies should be found whilst performing SLR. On the other hand, the main goal of SMSs is to provide an overview of a broader research topic. Hence, the search requirements for SMSs are less stringent (Kitchenham et al., 2010; 2011), especially for research topics that are very broad, and an enormous amount of primary studies exist. In that case, it is hard to expect that we deal with all relevant primary studies (Wohlin et al., 2013). The authors of Cruzes and Dybå (2011) have also come to similar conclusions: "*SRs that involve the transformation of raw data, or that include large numbers of primary studies, require greater resources, and where the review question and/or range of evidence is very broad, it may be necessary to sample.*" However, current practices in performing SMSs (e.g., Engström and Runeson, 2011; Barney et al., 2012; Ampatzoglou et al., 2013; Kosar et al., 2016c) do not include random sampling, but the inclusion of primary studies obtained from two or more Digital Libraries (DLs) with a possible additional search strategy such as snowballing (Kitchenham et al., 2010). Without random sampling, we can't make inferences and produce proper generalization. Hence, we are convinced that an inclusion of statistical methods is needed whenever all relevant primary studies cannot be classified. Indeed, one of the advantages of statistical methods is their abilities to use smaller numbers of subjects to make inferences about whole populations that would otherwise be prohibitively expensive to study. A part of the experimental design is the error, commonly called the margin of error (confidence interval) (Cochran, 1977; Moore et al., 2009), the researcher is willing to accept for a study. It tells us the level of precision and the range within which the true value of an estimate (e.g., population mean, proportion of subjects) lies. This approach is not uncommon also in other sciences where there is a need to gather evidence about different programs' impacts (e.g., adoption of new technology). The important question is

* Corresponding author.
E-mail address: tomaz.kosar@um.si (T. Kosar).

https://doi.org/10.1016/j.jss.2018.06.078
Received 26 January 2017; Received in revised form 8 December 2017; Accepted 30 June 2018
0164-1212/ © 2018 Elsevier Inc. All rights reserved.

how to determine a sample size with an acceptable margin of error (Bartlett et al., 2001). The main contributions of this paper are:

- A new approach for performing SMSs that decreases the assessing cost of primary studies' screening and classifications (in a particular use case presented in this paper, instead of classifying 476 primary studies, it is sufficient to classify only 301 randomly selected primary studies when the acceptable margin of error is 5%) and increases trustworthiness of SMSs without compromising the quality of results. The presented approach is suitable whenever SMS is performed on a broad topic with enormous existing primary studies.
- The approach is based on random sampling and the margin of error. In this paper, the proposed approach was applied on three different use cases: SMS on Domain-Specific Languages, SMS on Template-based Code Generation, and SMS on Software Reliability Modeling. From these three use cases, similar aggregated results were obtained and similar conclusions derived as in the previously performed SMSs (Kosar et al., 2016c; Syriani et al., 2017; Febrero et al., 2014). It is shown that random sampling can indeed be introduced into the design of SMSs.
- The approach has been empirically evaluated in terms of the saved time along the execution of the study and the loss in terms of accuracy.
- In this paper, our previous SMS on Domain-Specific Languages (DSLs) (Kosar et al., 2016c) has been extended by the inclusion of two additional DLs. It is shown that our method, driven by the margin of error and random sampling, produces similar and reliable results when compared to the original study (Kosar et al., 2016c) and extended SMS on DSLs presented in this paper.

This paper is organized as follows. Related works are discussed in Section 2. A description of a new method for performing SMSs driven by the margin of error is presented in Section 3. A case study using the newly proposed method is given in Section 4. Case study results are presented in Section 5. A discussion and threats to validity are presented in Section 6. Key findings and concluding remarks are summarized in Section 7.

2. Related work

The guidelines for performing SRs in software engineering (Kitchenham and Charters, 2007) outlined three phases: Planning the review (identification of the need for a review, commissioning a review, specifying the research questions, developing a review protocol, evaluating the review protocol), conducting the review (identification of relevant primary studies, selection of primary studies, study quality assessment, data extraction and monitoring, data synthesis), and reporting the review (specifying dissemination mechanisms, formatting the main report, evaluating the report). The outlined three phases (planning the review, conducting the review, and reporting the review), with the aforementioned sub-phases, were later simplified for SMSs in Petersen et al. (2008) into five stages:

- Defining research questions,
- Conducting a search for primary studies,
- Screening primary studies based on inclusion/exclusion criteria,
- Classifying the primary studies, and
- Data extraction and aggregation.

This simplified structure has been adopted by many researchers (e.g., Laguna and Crespo, 2013; Riaz et al., 2015; Ameller et al., 2015; Yang et al., 2016; Rodríguez et al., 2017), and even by the authors of the original guidelines (Kitchenham and Charters, 2007). By examining the literature on existing SMSs, we have found out that some of them were performed on research topics with very little existing evidence. In an extreme case as little as 13 primary studies have been identified and examined in Barreiros et al. (2011), while other studies we have come across identified and examined a modest number of primary studies (e.g., 31 in Abdellatief et al. (2013), 32 in Silva et al. (2011), 45 in Neto et al. (2011), 47 in Laguna and Crespo (2013), 55 in Li et al. (2013), 64 in (Engström and Runeson, 2011), 65 in Mehmood and Jawawi (2013)). In these cases, a selected research topic is probably too narrow for SMS, and SLR would be more appropriate. We are convinced that SMSs would be of much greater use if they would be applied to broader research topics. Indeed, there are SMSs which we come across that classified a substantial number of primary studies (e.g., 481 in Syriani et al. (2017), 503 in Febrero et al. (2014), 679 in Haghighatkhah et al. (2017), 1,440 in Nascimento et al. (2012)). Identifying primary studies is one of the most crucial steps in any SRs. In Webster and Watson (2002) backward and forward snowballing have been proposed as the main method to find primary studies, which has been shown later in Jalali and Wohlin (2012), Badampudi et al. (2015) and Afzal et al. (2016) as equally, if not more, reliable and efficient as using well defined search strings in DLs. Yet another approach for identifying primary studies in SLRs has been presented in Zhang et al. (2011), where the search step for relevant primary studies has been improved by recommending the concept of Quasi-Gold Standard (GQS), a collection of known primary studies, and quasi-sensitivity as a measure of evaluating the reliability of SRs, but this approach (Zhang et al., 2011) depends too much on a good QGS. If a collection of known primary studies is of low quality, then the reported reliability is questionable and the requested quasi-sensitivity level, which should be between 70% and 80% (Zhang et al., 2011), can be achieved easily. Overall, identifying a representative pool of primary studies is challenging in performing SRs. After identification of primary studies, a time consuming screening phase follows, where primary studies are checked against inclusion/exclusion criteria. In the works Malheiros et al. (2007) and Felizardo et al. (2011), the time consuming screening process has been reduced slightly by incorporating Visual Text Mining techniques, such as visualizations based on the content of primary studies and based on their citation relationship. However, only few participants have been used in their pilot case studies (Malheiros et al., 2007; Felizardo et al., 2011) and larger replications of this work are needed. In work Miwa et al. (2014), the active learning approach has been used to reduce the workload of screening using Support Vector Machine (SVM) to carry out categorization of primary studies after manual screening of a random sample of relevant and irrelevant primary studies. Our aim is not only to reduce the screening process, but also the classification process, which is yet another challenging task in performing SRs. However, this activity is not different in the proposed SMS driven by the margin of error than in other approaches and, therefore, not discussed further in this paper. The main difference is in the number of primary studies which need to be screened and classified to achieve reliable results.

On the other hand, the main goal of SMSs is to provide an overview of a research topic. Hence, the search requirements for SMSs are less stringent than those for SLRs (Kitchenham et al., 2011). Due to the possible immense amount of primary studies included into SMS, the cost of assessing all the studies would be unreasonably high. For a very broad topic with a large body of primary studies included into SMS, an even more important question to answer is: When have we collected enough evidence? In this paper, yet another approach for performing SMSs is proposed by introducing random sampling of primary studies until the aggregated data are not within the acceptable margin of error or confidence interval.

3. The proposed approach: a Systematic Mapping Study driven by the margin of error

The following procedure is proposed for an SMS driven by the margin of error (Fig. 1). The quality of any SR, SLR and SMS, depends heavily on the identified primary studies. Hence, a concerted effort

Download English Version:

https://daneshyari.com/en/article/6885253

Download Persian Version:

https://daneshyari.com/article/6885253

Daneshyari.com