Decision Support Architecture for Primary Studies Evaluation

Vilmar Nepomuceno
Informatics Center (CIn)
Federal University of Pernambuco
Recife - PE, Brazil
vsn@cin.ufpe.br

ABSTRACT

Background. A systematic literature review is a process in which all relevant available research about a research question is identified, evaluated, and interpreted through individual studies. The workload required for this process may bias the evaluation of the studies, affecting the result. **Aim**. Creating a decision support architecture to assist participants of a systematic review in the selection process of the individual studies and quality assessment of these studies, possibly improving the execution time and reducing the evaluation bias. **Method**. Improving the primary studies selection and quality assessment processes by using text mining techniques and ontologies to construct a decision support architecture. We will also conduct experiments to evaluate the proposed architecture. **Contribution**. Improve the primary studies selection and quality assessment processes, reducing its workload, and lowering the evaluation bias in systematic literature reviews.

Keywords

Systematic Literature Review, Text Mining, Ontology.

1. INTRODUCTION

The systematic literature review (SLR) is a process in which all relevant available research about a research question, or area, or phenomenon of interest is identified, evaluated, and interpreted through their individual studies, called primary studies during the SLR process. A guide to lead the SLR in software engineering was proposed by Kitchenham and Charters [1] and summarizes the systematic review process into three main phases: planning the review, conducting the review and reporting the review results.

The SLR planning phase generates a protocol that defines how to conduct the review. Once started, the process of conducting the review needs to define a search strategy, which is responsible for finding primary studies available, and, once obtained the studies in potential, it is necessary to perform the selection from these studies through criteria that were defined in the SLR protocol. The criteria for studies inclusion and exclusion must take into account the research question already defined.

The process of primary studies selection is a free interpretation of the criteria of whom is leading the SLR, hence, the large number of papers retrieved during the search process and the poor quality of their abstracts [2] makes the completion of the selection process a hard task and sometimes inaccurate. After the selection process is carried out to evaluate the quality of selected papers to increase the reliability and importance of SLR results, and to perform this task, there are several guidelines, which are usually not properly followed and the use, in general, is not justified by the authors [3]. These two processes of primary studies evaluation, study selection and quality assessment are hard tasks and time consuming [4], much of this time due to the primary study reading procedure.

There are studies that talk about automatic selection of primary studies using text mining techniques, in which the primary studies are classified by the similarity of texts [5] [6]. However, in these studies, the conductor does not assist the process, which can generate a set of papers that is significantly different than would be generated by the manual process. Besides, according to Kitchenham et al. [7], the process of quality assessment for primary studies is essential for SLRs. Our research seeks (i) to identify a way to semi-automatically support the primary studies quality assessment, with the use of text mining techniques and ontologies to describe prior knowledge about the SLR, as well as (ii) identifies, semi-automatically, the criteria for inclusion/ exclusion inside the text of the studies returned by the search, also through text mining techniques, supporting the selection process of the primary studies.

The semi-automatic evaluation of the quality and the inclusion/exclusion criteria semi-automatic search should support the evaluation of primary studies in SLR process. With this, we aim to reduce the execution time of the SLR, due to the primary study reading procedure, decreasing the primary studies evaluation bias that may occur due to the evaluation process subjectivity, increasing the assurance that the outcome of the SLR is not being compromised [8]. In addition, you can increase the studies search space, which today is limited because of the effort spent during the selection process and subsequent quality evaluation of these studies.

2. RELATED WORKS

2.1 Quality Assessment

There are several guidelines available for quality assessment of primary studies, such as [9] which put forward eleven evaluation criteria based on CASP [10]. In Kitchenham et al. [11] was used a checklist for quality assessment, in order to specify an appropriate process for evaluating quality. The study concluded that at least two evaluators are required to improve the reliability and the quality assessment should be represented by the sum of the criteria used. However, Dieste et al. [12] identified trends that should not exist a plethora of items into instruments of quality control on SLRs, and that this assessment should be careful about the limits of this process with respect to aspects of internal validity [11].

There is still not a standard process for primary studies quality assessment in software engineering, several authors have suggested several ways to estimate the validity of the studies. Probably due to this lack of standardization, in our research were not found tools that automate the process, or part of this process.

2.2 Primary Studies Selection

The selection of primary studies is the most time-consuming task for an SLR, which can be affected by the titles and abstracts that do not reflect well the content of the work [4]. Additionally, time constraints may lead the research conductors to reduce the search space. Hence, automate the process, or part of it can help overcome these barriers.

Automatic classification of primary studies, indicating its inclusion/exclusion can be found in some works [5] [6], in other words, in these works, the study selection process is done in an automated way without the intervention of the researcher. With this approach can dramatically reduce the effort spent on that task, however, methods that do this kind of selection may include studies that did not help in the research, because of their low quality, or the algorithm low accuracy, as well as, they may exclude studies with low textual similarity, but with good quality, which could be used by the researcher in some way.

Therefore, this research proposes a semi-automated approach, leaving to the researcher the final decision of the inclusion or not of these studies. It is likely that when comparing our methodology and the automatic classifiers, there will be a loss of time taken to perform the task by using the technique proposed in this work. However, it is possible that the final result of the selection process, using our proposal, is more satisfactory and significantly faster than the manual approach.

3. PROPOSAL

According to the aim of this study, our guiding research question (RQ) is "How *can we improve the process of evaluating primary studies by automating parts of the process?*". This question can be decomposed into:

- RQ1. How can we improve the selection process by automating parts of the process?
- RQ2. How can we improve the quality assessment process by automating parts of the process?

An initial version of the proposed Decision Support Architecture (DS Architecture) is presented in Figure 1.

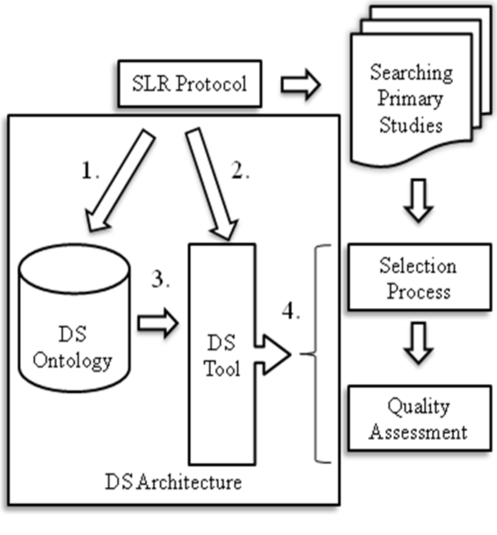

Figure 1. Proposed DS Architecture.

The knowledge about the SLR will be represented by an ontology (1) based on the protocol designed for the execution of the SLR and possible research conductor's refinements. This approach is proposed by Biolchini et al [13], that state it is possible to improve the results obtained during the SLR through standardization of the terminology for the concepts involved using an ontology. It is still undefined in what form the creation of the ontology will be held, but there are free available tools such as *protégé* [14] and Jena [15] and the GATE (*General Architecture for Text Engineering*) framework [16], that was already used in [17] for this purpose.

These criteria should be used as input to a text mining algorithm, which should be available as a tool component (DSTool) and will be able to identify whether the selection criteria have been met and which quality criteria can be identified inside the primary study text. The interest points of the text, where the tool will be based to respond, will be shown to the research conductor, which will decide whether the criteria were actually achieved. The inclusion/exclusion criteria will be provided by the conductor in the protocol creation, and the quality criteria will come from one of the guidelines available in the literature, after to perform a systematic literature review that should indicate the best guideline to be implemented by the text mining algorithm. The main idea to provide a previous guideline is to improve the robustness and accuracy of the algorithm, which does not prevent that other criteria, which are not present in the selected guideline, may be used. Another input to the text mining algorithm is the ontology built upon the protocol (3), many text mining techniques use ontologies as a knowledge base, one of which is the ontology-based question answering system [18], which It is the starting point in the tool development. Still will be evaluated, ways to present the results from the tool (4).

3.1 Decision Support Algorithm

Based on the system architecture proposed by Bo and Yunqing [18], a question answer architecture is being proposed to the decision support algorithm (Figure 2).

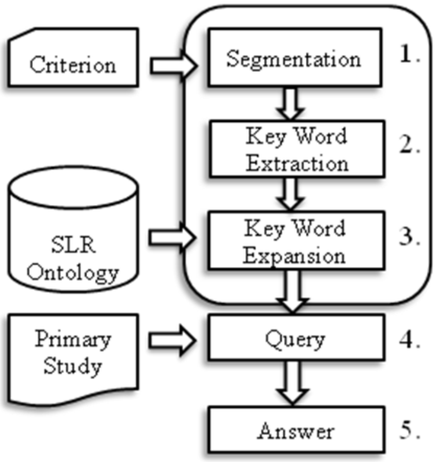

Figure 2. Proposed Algorithm Architecture.

The questions that will serve as input to the algorithm will be provided by the protocol, at this time the criteria for inclusion/exclusion and the quality criteria will be provided one by one to the segmentation process (1). At this point, the question will be broken in terms, and the keywords will be extracted (2). As, for example, the following quality criteria: "*Is there a clear statement (definition) of the aims (goals, purposes, problems, motivations, objectives, questions) of the research?*" [3], where we can draw the following set of terms T = {statement, definition, goals, purposes, problems, motivations, objectives, questions, research}. The keywords found in the criteria will be expanded with the help of the ontology (3). Thus, we can create a larger set of terms that should improve the algorithm accuracy. These three points can be called as the search query creation process.

After creating the search query, we look for the answers in the primary studies (4), at this point we use text mining algorithms to find possible answers to the criteria and return the points in the text where the answers can be found, as well as the text of the response (5).

4. EVALUATION PROPOSAL

To evaluate the effectiveness of the Decision Support Architecture (DSArch) a controlled experiment [19] will be performed. The study aims to answer the following research questions:

- RQ1. Does DSArch decrease the selection time of primary studies?
- RQ2. Does DSArch increase consensus between individuals of the same pair including/excluding studies?
- RQ3. Does DSArch decrease the quality assessment time of the selected primary studies?

4.1 Hypotheses, Variables, and Parameters

The null hypotheses are presented below:

- $H_{0,RQ1}$. There is no difference in the execution time in the selection process with or without the use of DSArch.
- $H_{0,RQ2}$. There is no difference in the consensus among peers with or without the use of DSArch.
- $H_{0,RQ3}$. There is no difference in the quality assessment time of the selected primary studies with or without the use of DSArch.

To examine the hypotheses the following dependent variables will be used:

- Selection Time. The time to complete the process for selecting primary studies.
- Consensus. Measured by inter-rater agreement coefficient. Cohen´s Kappa will be used, due to the decision be taken by pairs.
- Quality Time. The quality assessment time of the selected primary studies.

The experiment factor is the process for evaluating the primary studies, by selecting these studies and quality assessment of the selected primary studies. The factor alternatives (treatments) are executing the inclusion/exclusion process and quality assessment using DSArch and the other is using only the protocol criteria manually, without tool support for the alternatives.

4.2 Material, Tool, and Training

To perform the experiment a subset of primary studies drawn from an SLR will be provided to participants. The protocol created for the SLR will be provided and both the search process and the visualization of the primary studies will be conducted by REviewER$_1$. It belongs to our research group and it is a tool that gives support to search process in some databases (ACM library, Engineering Village, IEEExplorer, Science Direct, Scopus, and Springer Link) in a automated way and gives support to the primary studies selection process. A version of REviewER will be developed, containing the DS Architecture, to be used in the evaluation. However, only the DSArchictecture should be assessed.

A training will be conducted with the participants to present the reviewer tool and how one should analyze the DSArch. For this training some primary studies, chosen from a subset of the primary studies set obtained from the SLR search process, will be selected and participants should evaluate them using the factor alternatives.

4.3 Task and Data

To measure the dependent variables two tasks will be realized, which requires no prior knowledge on SLR from the subjects, hence, we aim to facilitate the process of selection of these participants. The tasks to be performed in the experiment are the selection of primary studies from a subset of the primary studies set obtained from the SLR search process and the quality assessment of these selected studies. No other SLR procedures will be performed.

At the end of the execution, the participants must provide a list of accepted primary studies, the time taken for completion of the selection process and the time taken for completion of the quality assessment process. A questionnaire will be performed after the experiment to evaluate the experiment itself and what the participants thought about the proposed architecture.

4.4 Execution

The experiment will take place in a lab with the presence of all participants at the same time. Participants will be divided into pairs following the proposed by Kitchenham et al [7], such choice will be at random.

Table 2 shows the proposed design of the experiment:

Table 2. Experimental Design

Pairs/EU	***EU1***	***EU2***
P1	A	B
P2	B	A

Where P1 and P2 are two pairs of participants, EU1 and EU2 are the experimental units, each is one subset of primary studies obtained from the SLR proposed in the experiment, the subsets will be chosen at random, as well as, A and B that are the applied treatments. This design is being proposed to facilitate internal replication of the experiment.

To collect the data the participants should submit the list with the selected primary studies, the time taken to complete the selection process and the time taken to complete the quality evaluation of the selected studies.

4.5 Data Analysis

Latin squares have an analysis procedure very similar to the factorial experiments (multiple factors). In the case of factorial experiments, we can consider the lock variable as a factor. For the Latin square, we have two lock variables and a factor of interest. Also, there is another complicating factor: the experiment design has multiple replications. This leads to a replicated Latin square design with equal columns (processes) and different lines (participants).

A possible statistical test for the analysis is the ANOVA, as proposed in [19], however, for being a parametric test, some preconditions should be evaluated, and if one of them is violated a equivalent non-parametric test can be used.

[1] http://sites.google.com/site/eseportal/tools/reviewer

4.6 Threats to Validity

Some possible threats to validity of the experiment are already being assessed.

4.6.1 Internal Validity

The completion of the training can generate an apprenticeship in relation to DSArch, which may influence the evaluation process of the primary studies when participants are not using the DSArch. However, if there is influence, it will be in favor of the null hypothesis.

4.6.2 External Validity

The results may not be generalizable to all the researches that perform SLRs, because we do not sample from the population of SLRs researchers, but we intend to make a satisfactory outcome that will bring evidence of the effectiveness of DSArch.

5. CONTRIBUTIONS, FUTURE WORK AND ADVICES

The use of the DSArch might reduce the effort to complete SLR, which is one of the major problems encountered in conducting this type of research. Another problem to be addressed by using DSArch is the primary studies subjective selection, which is biased, by automating part of the process. Another gain by automating part of the process is decreasing the number of conflicts generated during the evaluation of studies within the pairs. An expectation for the proposed algorithm is that it can be used to assist in the SLR data analysis process, since, the question answer principle can also be used at this stage, however, this analysis is out of the context of this thesis.

So far, it was conducted an ad-hoc literature review on the proposed theme, as well a first design of the controlled experiment that will be conducted to evaluate the DSArch. The planned next steps are:

- Execute a systematic literature review of techniques for assessing the quality of primary studies, aiming at choosing a technique to be semi-automated;
- Select and implement an ontology to represent the prior knowledge on the SLR;
- Development of the DSArch;
- Refine the plan and execute the controlled experiment to assess the DSArch;
- Analyze the obtained results;
- Write the Thesis.

The main points where advices are needed:

- The proposed architecture is consistent with the problem found?
- The experimental design can evaluate the proposed architecture?
- The statistical test, ANOVA can evaluate the data generated by the experiment?

6. REFERENCES

[1] Kitchenham B. A. and Charters S.. Guidelines for performing systematic literature reviews in software engineering, Technical Report: 2007.

[2] Brereton P., Kitchenham B. A., Budgen D., Turner M. and Khalil M.. Lessons from applying the systematic literature review process within the software engineering domain. Journal of Systems and Software, vol. 80 (4), pp. 571-583, 2007. DOI: 10.1016/j.jss.2006.07.009.

[3] Zhou Y., Zhang H., Huang X., Yang S., Babar M. A., and Tang H.. 2015. Quality assessment of systematic reviews in software engineering: a tertiary study. In Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering (EASE '15). ACM, New York, NY, USA, Article 14 , 14 pages.

[4] Carver J. C., Hassler E., Hernandes E., and Kraft N. A.. Identifying barriers to the systematic literature review process. In Empirical Software Engineering and Measurement, 2013 ACM/IEEE International Symposium on, pages 203-212. IEEE, 2013.

[5] Cohen A. M., Hersh W. R., Peterson K., and Yen P. Y.. Reducing workload in systematic review preparation using automated citation classification. Journal of the American Medical Informatics Association: JAMIA, 13(2):206–219, 2006. ISSN 1067-5027.

[6] Tomassetti F., Rizzo G., Vetro A., Ardito L., Torchiano M., and Morisio M. (2011). Linked Data approach for selection process automation in Systematic Reviews. In Proceedings of the 15th Annual Conference on EASE, pp. 31 – 35.

[7] Kitchenham B., Brereton P., Turner M., Niazi M., Linkman S., Pretorius R., Budgen D., Refining the systematic literature review process – two observer participant case studies, Empirical Software Engineering 15 (6) (2010) 619–653.

[8] Kitchenham B., Sjøberg D. I., Brereton O. P., Budgen D., Dybå T., Höst M., and Runeson P.. Trends in the quality of human-intensive software engineering experiments: a quasi-experiment. ieee, 2013.

[9] Dyba T., Dingsøyr T.. Strength of evidence in systematic reviews in software engineering. Proceedings of the Second ACM-IEEE international symposium on Empirical software engineering and measurement. 2008, p. 178-187.

[10] P. H. R. U. in Oxford. Critical appraisal skills programme. http://www.casp-uk.net/, 2013.

[11] Kitchenham B., Sjøberg D. I., Brereton O. P., Budgen D., Dybå T., Höst M., and Runeson P.. Can we evaluate the quality of software engineering experiments? In Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement, ACM, 2010.

[12] Dieste O., Grimán A., Juristo N., and Saxena H.. “Quantitative Determination of the Relationship between Internal Validity and Bias in Software Engineering: Consequences for Systematic Literature Reviews,” Proc. Int’l Symp. Empirical Software Eng. and Metrics, pp. 285-288, 2011.

[13] Biolchini J., Mian P., Natali A., Conte T., Travassos G.. Scientific research ontology to support systematic review in software engineering. Advanced Engineering Informatics, vol. 21 (2), pp. 133-151, 2007. DOI: 10.1016/j.aei.2006.11.006.

[14] Knaublock, H.. Protégé-OWL. 2003. Avaiable at http://protege.stanford.edu. Accessed on 07/03/2015.

[15] Labs, H.. Jena: A free and open source Java framework for building Semantic Web and Linked Data applications. 2010. Avaiable at http://jena.apache.org. Accessed on 07/03/2015.

[16] Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A framework and graphical development environment for robust NLP tools and applications. In: Proc. of the 40th Anniversary Meeting of the ACL. (2002).

[17] Witte R., Li Q., Zhang Y. and Rilling J., Ontological Text Mining of Software Documents, 12th International Conference on Applications of Natural Language to Information Systems (NLDB 2007, Paris, France, June 27-29, 2007.

[18] Bo W. and Yunqing L., "Research on the Design of the Ontology-Based Automatic Question Answering System," Computer Science and Software Engineering, 2008 International Conference on , vol.5, no., pp.871,874, 12-14 Dec. 2008.

[19] Juristo N., Moreno A. M., Basics of Software Engineering Experimentation, Springer Publishing Company, Incorporated, 2010.