Validating Search Processes in Systematic Literature Reviews

Barbara Kitchenham[1], Zhi Li[2,3] and Andrew Burn[4]

[1] School of Computer Science and Mathematics, Keele University, Staffs, ST5 5BG, U.K.
b.a.kitchenham@cs.keele.ac.uk
[2] College of Computer Science and Information Technology, Guangxi Normal University
No.15 Yu Cai Road, Guilin, Guangxi 541004, P.R. China
[3] Key Laboratory of High Confidence Software Technologies (Peking University)
Ministry of Education, Beijing, 100871, China

[4] School of Engineering and Computing Sciences, Durham University, Durham, DH1 3LE, U.K.

Abstract. Context: Systematic Literature Reviews (SLRs) need to employ a search process that is as complete as possible. It has been suggested that an existing set of known papers can be used to help develop an appropriate strategy. However, it is still not clear how to evaluate the completeness of the resulting search process. **Aim:** We suggest a means of assessing the completeness of a search process by evaluating the search results on an independent set of known papers. **Method:** We assess the results of a search process developed using a known set of papers by seeing whether it was able to identify papers from a different set of known papers. **Results:** Using a second set of known papers, we were able to show that a search process, which was based on a first set of known papers, was unlikely to be complete, even though the search process found all the papers in the first known set. **Conclusions:** When using a set of known papers to develop a search process, keep a "hold-out" sample to evaluate probable completeness.

1 Introduction

Systematic literature reviews (SLRs) aim to find all relevant research papers on a specific topic or research question. In order to refine a search process, [6] recommend comparing the papers found by the search process with a set of known papers, and we adopted the method in a recent SLR [7]. In this paper, we suggest a means of assessing the probable completeness of such a strategy by using an independent set of known papers to provide an estimate of the precision of the search strategy.

We discuss related studies in Section 2 and explain our methods in Section 3 We report our results in Section 4 and discuss them in Section 5. We present our conclusions in Section 6.

Kitchenham B., Li Z. and Burn A..
Validating Search Processes in Systematic Literature Reviews.
DOI: 10.5220/0003557000030009
In *Proceeding of the 1st International Workshop on Evidential Assessment of Software Technologies* (EAST-2011), pages 3-9
ISBN: 978-989-8425-58-4

2 Related Research

Dieste and Padua [1] reported 24 search strings suitable for identifying software engineering empirical studies. They devised their strategies empirically using as a gold standard the 103 human-centric experiments and quasi-experiments found by Sjøberg et al. (2005). The best search string found 93.3% of the 103 papers, although there was a false positive rate of 82.9%. However, their results apply to searching for empirical studies, in general, rather than searching for empirical studies on a specific topic. Furthermore, as already mentioned, their gold standard took a rather restricted view of empirical studies. In this paper we look at topic related search strings and show that the use of a specific set of papers as a gold standard may lead to over-estimates of the precision of the search process.

Skoglund and Runeson [11] took a different approach and investigated a search process based on identifying a "take-off" paper that becomes the starting point of a search and following the references from that paper. They also considered using cardinal papers (i.e. papers that are frequently referenced) and identifying papers that referenced those cardinal papers. Our paper takes a similar approach to this paper but is based on using a set of known papers to develop search strings and investigates how to assess the effectiveness of the resulting search process.

3 Method

We developed a search process to detect empirical papers on unit testing and regression testing based on developing search strings that found as many as possible of the papers used by Juristo et al. [5] in their literature review. The search process involved an automated search of four digital libraries, identification of known papers found by the automated search, followed by checking the references of known papers found by the search (i.e. snowballing). This search process found all the papers used by Juristo et al. [5].

We then compared the set of all the papers found by the automated search process with the set of papers used in a regression testing SLR [3]. Papers found by our search process and selected by Engström et al. [3] were also snowballed to look for known papers missed by the automated search. Although the Engström et al. [3] study included a different set of regression papers to those included by Juristo et al. [5] study, we thought that a search process that found all of the regression testing papers included by Juristo and that used fairly generic search strings ought to find most of the papers included by Engström et al.[3] and would therefore provide us with an independent assessment of the effectiveness of our search.

Thus, the study procedure was as follows:

1. We developed, iteratively, search strings that found the maximum number of papers identified by Juristo et al. [5] using the ACM, IEEE and CiteSeer digital libraries. The search strings were restricted to the time period 1987 to 2005 which covered the time period of the Juristo review. Our search strings were unable to find all the papers found by Juristo et al. [5], so we extended our search process as explained below.

2. The search strings were applied to the SCOPUS digital library (the only changes to the search strings were those necessary to permit the search to be performed using the SCOPUS interface). Since SCOPUS is a general indexing system, that indexes ACM and IEEE we should not have found more of the papers found by Juristo et al. [5] but we wanted to check since it was possible that the SCOPUS searching process was more efficient than the IEEE or ACM searches.

3. The references of known papers found by the automated searches were scanned for otherwise missed papers.

4. The papers found by stages 1 and 2 were searched again looking for papers selected by Engström et al. [3] for their SLR of regression testing.

5. The known Engström et al. [3] papers found in Stage 4 were then snowballed to look for otherwise missed Engström et al. papers

We used measures of sensitivity and precision to assess the quality of the search process. In particular, sensitivity can be regarded as a measure of the probable completeness of the search process. Sensitivity and precision are defined as follows:

$$\text{Sensitivity} = (\text{KPF})/(\text{KP}) \quad (1)$$

$$\text{Precision} = (\text{KPF})/(\text{TPS}) \quad (2)$$

Where

KP = The number of known papers

KPF=The number of known papers found by the search process

TPS=The total number of papers found by the search process.

4 Results

As shown in Table 1, our search process identified all of the papers used by Juristo et al [5].

Table 1. Success finding papers selected by Juristo et al [5].

Source of papers	Known Papers	Total papers found by search	Sensitivity (%)	Precision (%)
Included by Juristo et al. 2006	24			
Indexed by IEEE/ACM digital libraries	21			
Known papers found by searching IEEE, ACM, CiteSeer digital libraries	17	1480	70.8	1.1
Known papers found by searching the SCOPUS digital library	15	1278	62.5	1.2
Known papers found both searches	22	3758	91.7	0.6
Extra Known Papers Found by Snowballing	2		8.3	
Total found by process	24		100	

However, it should be noted that:

- The search strings only found 17 (70.8%) of the known papers when applied to the ACM, IEEE and CiteSeer digital libraries.
- The SCOPUS search found an additional five papers as well as ten papers found by the first search. These results deomstrate that the search mechanisms used by different digital libraries are not equivalent.
- It required snowballing the references of the remaining papers found by the automated searches to find the final two papers.

Table 2 shows the number of regression testing papers used by Engström et al. (2010) that were also found by our search process (excluding two papers published in 2006 which was outside the time limit of our search). The two automated search stages found a total of 22 (80%) of the papers. Snowballing found another two papers bringing the total to 24 (88%).

Table 2. Success finding Regression Testing Papers.

Source of papers	Known Papers	Total papers found by search	Sensitivity (%)	Precision (%)
Included by Engstöm et al. 2010 (excluding two papers published in 2006)	25			
Indexed by IEEE/ACM digital libraries	22			
Known papers found searching IEEE, ACM, CiteSeer digital libraries	15	1480	60	1.0
Known papers found by searching the SCOPUS digital library	14	1278	56	1.1
Known papers found by both searches	20	3758	80	0.5
Extra Known papers found by Snowballing	2		8	
Total found by process	22		88	

5 Discussion

Although it found all the primary studies used by Juristo et al. [5] which included 10 regression testing studies, the search process did not find all the regression test studies found by [3]. Based on the sensitivity of the search process calculated for the [3] papers, the probable completeness of the search process is 88%.

The results indicate that the search process is unlikely to be complete and, if completeness is essential, additional searching is necessary. In this, case, a more intensive search might be based on searching using the names of specific unit testing methods, and/or contacting well-known researchers to ask if we had missed any of their papers.

Although the papers for which we searched were all indexed by ACM or IEEE, we found different papers when searching the ACM and IEEE digital libraries than we did when using the equivalent search strings on the SCOPUS indexing system. This emphasises that search algorithms differ among different digital libraries and that apparently redundant searches may be necessary to increases completeness.

5.1 Estimating the Number of Missing Papers

If sets of primary studies are obtained by independent search processes, it is also possible to estimate the likely number of missing papers using the capture-recapture approach (Spoor et al., 1996). They identify the maximum likelihood estimator of the total population size (N) to be:

$$N=[(M+1)(n+1)/(m+1)]-1 \tag{3}$$

Where

M=number of publications found by one search
n= number of publications found by the other search
m=number of publications identified by both searchers

The variance of N (Var) is calculated as follows:

$$Var=[(M+1)(n+1)(M-m)(n-m)/[2(m+1)(m+2)] \tag{4}$$

The standard deviation is calculated as the square root of Var and the 95% confidence limits of N will be approximately plus or minus twice the standard deviation of N.

If the set of regression testing papers found by Engström et al. [3] was independent of the set of regression testing papers found by Juristo et al [5], we might be able to estimate the likely number of missing regression testing papers. However, since Engström et al. [3] reference Juristo et al. [5], it is clear that the authors knew about Juristo's study and the two sets of studies are not independent, so a capture-recapture estimate would be biased.

5.2 Limitations of the Study

The fact that our search missed relevant papers might be explained by the fact that the sources we searched differed. Engström et al. [3] used ACM, IEEE, ScienceDirect, Springer LNCS and Web of Science, whereas we applied our search process to the IEEE, ACM, CiteSeer and SCOPUS digital libraries. However, all the papers except three were indexed by the IEEEXplore digital library, and we found those three papers.

Another issue is that the Engström et al. [3] paper only considered regression test selection methods, so we cannot be sure that the completeness estimate based on regression test studies would apply equally to conventional unit testing methods. However, in an extremely rigorous mapping study of mutation testing, Jia and Harmon [4] (In press) identified 10 papers relating to seven unique empirical studies. Three of the papers were technical reports and one paper was published in 2006, leaving six papers which were within the scope of our automated searches. We found five of the six, achieving a sensitivity of 83% (admittedly on a very small number of papers) which is comparable to the 88% sensitivity we achieved for the Engström papers.

6 Conclusions

The technique of using a separate set of papers to validate a search process allows us to assess the probable completeness of the search process. In this case, we used one set of papers to help derive the search process (i.e. the specific search strings for automated searches of digital libraries) and a separate set of papers to assess the completeness of the search process. This is similar to using separate model building and validation datasets in data mining studies.

In most cases, researchers performing SLRs will not have access to two independent reviews addressing the same topic. However, in order to both improve their search process and to provide a quantitative assessment of completeness of the process, we suggest researchers obtain an initial set of relevant papers (based on expertise, a manual search of relevant sources, or a published literature review) and split the set of papers, at random, into a set of papers to be used to refine/improve the search process and a separate hold-out sample to provide an independent estimate of the completeness of the search. The independent estimate of completeness can be used to assess whether additional search effort is required. In some cases, particularly mapping studies, further searching might be unnecessary. However, if the completeness value is reported when a systematic literature review is documented, it would provide a useful quality indicator for readers. This is similar to reporting the Kappa agreement value of the inclusion/exclusion process.

In terms of future work, we have already identified eight additional unit testing papers and two additional regression testing papers from an initial search process performed by novice researchers searching only the ACM and IEEE digital libraries without the aid of the list of known papers [8]. Using Spoor et al.'s method and comparing Juristo et al.'s selection of 21 papers with the joint set of 13 relevant papers found by the novice researchers (with three overlapping papers) suggests that the total population of empirical unit and regression testing papers might be N=76 leaving a total of 45 more empirical testing papers to find. However, the standard deviation of N is rather large (i.e. 37) due to the small overlap between the sets of studies (only three papers in all) giving a lower 95% confidence limit on N of about 2 and an upper 95% confidence limit of about 150. Nonetheless, we intend to review all the papers found by our automated search process to see whether we can identify further unit testing and/or regression testing studies with the aim of extending the published reviews.

Acknowledgements

The work described in this paper was funded by the UK Engineering and Physical Sciences Research Council project EPIC/E046983/1.

References

1. Dieste, O. and Padua, A. G., 2007. Developing Search Strategies for Detecting Relevant Experiments for Systematic Reviews. Proceedings 1st International Symposium on Empirical Software Engineering and Measurement, ESEM ’07, IEEE Computer Society.
2 Elbaum,S., Mailshevsky, A. G. and Rothermel, G., 2000. Prioritizing Test Cases for Regression Testing. Proc. Int’l Symp. Software Testing and Analysis, ACM Press, pp. 102–112.
3. Engström, E., Runeson, P. and Skoglund, M., 2010. A systematic review on regression test selection techniques. Information and Software Technology, 52, pp 14-30.
4. Jia, Y. and Harmon, M. In press. An Analysis and Survey of the Development of Mutation Testing, IEEE TSE. Doi 10.1109/TSE.2010.62.
5. Juristo, N. Moreno, A. M., Vigas, S. and Solari, M., 2006. In Search of What We Experimentally Know about Unit Testing, IEEE Software, vol. 23, pp. 72-80.
6. Kitchenham, B. A. and Charters, S., 2007. Guidelines for performing Systematic Literature. Reviews in Software Engineering, Technical Report EBSE-2007-01.
7. Kitchenham, B., Pretorius, R., Budgen. D., Brereton, P., Turner, M., Niazi, M., and Linkman, S. G., 2010. Systematic Literature Reviews in Software Engineering – A Tertiary Study. Information and Software Technology, 52(8), pp 934-944.
8. Kitchenham, B. A. Breerton, O. P. Li, Z, Budgen, D., Burn, A., 2011. Repeatability of systematic literature reviews, EASE 20011.
9. Leon, D., Masri, W. and Podgurski, A., 2005. An Empirical Evaluation of Test Case Filtering Techniques Based on Exercising Complex Information Flows, Proc. 27th Int’l Conf. Software Eng. (ICSE 05), IEEE CS Press, pp. 412–421.
10. Rothermel, G. Untch, R. H. Chengyun C, Harrold, M. J., 1999. Test case prioritization: an empirical study, Proc. Int’l Conf. Software Maintenance, IEEE CS Press, pp. 179–188.
11. Skoglund, M. and Runeson, P., 2009. Reference-based search strategies in systematic reviews. Proceedings EASE’09, BCS eWic.
12. Spoor, P., Alrey, M., Bennett, C., Greensill, J. and Williams, R. Using capture-recapture technique to evaluate the completeness of systematic literature searches. British Medical Journal, 1996, 313(7053), p 342.