Information and Software Technology 98 (2018) 59–73

Contents lists available at ScienceDirect

Information and Software Technology

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

A comparison of automated training-by-example selection algorithms for Evidence Based Software Engineering

Edgar E. Hassler[a], David P. Hale*[,b], Joanne E. Hale[b]

[a] Department of Computer Information Systems, Appalachian State University, Boone, NC 28608, USA
[b] Department of Information Systems, Statistics, and Management Sciences, Culverhouse College of Commerce, The University of Alabama, Tuscaloosa, AL 35401, USA

ARTICLE INFO

Keywords:
Research infrastructure
Evidence Based Software Engineering
Systematic Literature Review
Systematic Mapping Studies
Culling
VSM
LSA
Recall
Precision
Document selection

ABSTRACT

Context: Study search and selection is central to conducting Evidence Based Software Engineering (EBSE) research, including Systematic Literature Reviews and Systematic Mapping Studies. Thus, selecting relevant studies and excluding irrelevant studies, is critical. Prior research argues that study selection is subject to researcher bias, and the time required to review and select relevant articles is a target for optimization.
Objective: This research proposes two training-by-example classifiers that are computationally simple, do not require extensive training or tuning, ensure inclusion/exclusion consistency, and reduce researcher study selection time: one based on Vector Space Models (VSM), and a second based on Latent Semantic Analysis (LSA).
Method: Algorithm evaluation is accomplished through Monte-Carlo Cross-Validation simulations, in which study subsets are randomly chosen from the corpus for training, with the remainder classified by the algorithm. The classification results are then assessed for recall (a measure of completeness), precision (a measure of exactness) and researcher efficiency savings (reduced proportion of corpus studies requiring manual review as a result of algorithm use). A second smaller simulation is conducted for external validation.
Results and conclusions: VSM algorithms perform better in recall; LSA algorithms perform better in precision. Recall improves with larger training sets with a higher proportion of truly relevant studies. Precision improves with training sets with a higher portion of irrelevant studies, without a significant impact from the training set size. The algorithms reduce the influence of researcher bias and are found to significantly improve researcher efficiency.

To improve recall, the findings recommend VSM and a large training set including as many truly relevant studies as possible. If precision and efficiency are most critical, the findings suggest LSA and a training set including a large proportion of truly irrelevant studies.

1. Introduction

Evidence Based Software Engineering (EBSE) is growing in importance and impact, aiding the maturity of the Software Engineering discipline by driving systematic, structured analysis, synthesis, and interpretation of empirical evidence [1]. Within EBSE, Systematic Literature Reviews (SLRs) and Systematic Mapping Studies (SMS) depend on the effective search and selection of primary research studies. While study search algorithms have been the focus for numerous researchers [2–7], study selection algorithms have received comparatively less attention [8]. Effective and efficient study selection (that is, selecting relevant studies and excluding irrelevant studies) is critical to the success of EBSE.

One of the main barriers in conducting SLRs and SMSs is the time-commitment associated with the selection of studies from the extensive set of results returned as part of the search stage of the protocol. In addition to the time required for the selection of studies to be included in the review, a common problem is the consistent application of the inclusion/exclusion criteria of the review [9]. While inconsistences in the application of the inclusion/exclusion criteria are resolved through the discussions among researchers, such inconsistences result in the expenditure of additional time to resolve the inconsistences and re-accomplish the selection process. This additional researcher time and effort may present a significant barrier to the undertaking of such research [10].

The contribution of this research is the validation of automated training-by-example classifier algorithms that are computationally simple, do not require extensive training or tuning, and ensure inclusion/exclusion consistency while reducing researcher study selection time expenditure. The algorithms are evaluated using a series of simulations.

* Corresponding author.
E-mail addresses: hassleree@appstate.edu (E.E. Hassler), dhale@ua.edu (D.P. Hale), jhale@cba.ua.edu (J.E. Hale).

https://doi.org/10.1016/j.infsof.2018.02.001
Received 1 April 2017; Received in revised form 3 January 2018; Accepted 1 February 2018
Available online 06 February 2018
0950-5849/

The remainder of this paper is organized as follows. Section 2 provides an overview of the selection process and associated metrics along with a summary of previous work. Section 3 details the conceptual research model and research questions. The Monte-Carlo Cross-Validation simulation approach is described in Section 4. Section 5 discusses the research findings. Section 6 provides a confirmatory case study to provide additional evidence that extends external validity. Section 7 includes the discussion, implications and limitations of the research. Finally, Section 8 concludes the research and describes future work.

2. Background

To frame and ground this research, this section describes the study selection process used by EBSE researchers, the metrics used to assess the success of the selection process, current tools and methods used to support study selection, as well as the new proposed automated training-by-example classifier algorithms.

2.1. EBSE process

EBSE research studies follow the following process [1,11]:

1. **Planning.** During this phase, the research objectives and questions are defined and the protocol is created. The protocol includes sources of primary research studies, search methods and keywords, study inclusion (relevant) and exclusion (irrelevant) criteria, study quality criteria, a data extraction form, and a data synthesis strategy.
2. **Execution**. During this phase, primary studies are obtained, evaluated, and analyzed.
   a. **Search Execution**. During this Execution step:
      i. First, during Initial Selection, primary studies are identified, collected, and organized in the document ***corpus.***
      ii. During **Selection Execution** (also called here **Study Selection**), studies are evaluated according to the inclusion and exclusion criteria, and **classified** as either **relevant** (included) or **irrelevant** (excluded). If quality criteria dictate, this is followed by additional review of the corpus to remove studies deemed below quality thresholds.
      iii. In Selection Review, the corpus is reviewed to minimize incorrect exclusion of relevant studies.
   b. **Information Extraction.** During this Execution step, relevant information is extracted from those studies classified as included.
   c. **Analysis & Synthesis.** During this phase, the results of the included studies are analyzed and synthesized to accomplish the original research objectives and questions. This phase is highly variant between SLR and SMS research projects.
3. **Documentation.** During this final phase, a report detailing the results and findings is prepared. The report provides a transparent, repeatable account of the study, explicates the results, and provides discussion around meaning, implications and limitations.

This research focuses on improving the **Selection Execution** step (2b above; called **Study Selection** hereafter in this research). This is often accomplished in a series of steps designed to progressively reduce the corpus down to the truly relevant studies and includes [12,13]:

a. Starting with corpus study titles and the predefined inclusion criteria, studies are removed that are clearly irrelevant to the research topic.
b. Abstracts of the remaining studies from step ***a*** are examined, removing studies that are clearly irrelevant to the research topic.
c. Using the set of included studies from step ***b***, the full text of the studies is screened again against the inclusion/exclusion criteria.

In many cases, researchers combine steps a and b above. Two or more researchers conduct these steps independently, with classification disagreements resolved by agreed-upon method, such as consensus or voting. The research then progresses to Information Extraction as discussed above.

2.2. Study selection metrics

The goal of Study Selection is to retain the relevant studies found during the search process while excluding irrelevant studies. Consequently, the effectiveness of the selection algorithm can be measured by the level of True Positive (TP) and True Negative (TN) classifications (studies correctly classified as include and exclude, respectively), False Positive (FP) errors (including a study that should be rejected) and False Negative (FN) errors (excluding a study that should be included).

The traditional proportions of each type of error (FP or FN) with respect to the total number of studies classified provide a general indicator of classifier performance, but may be misleading due to the asymmetry between the portions of relevant and irrelevant documents typically found in the corpus [13]. Therefore, alternative measures the field of information retrieval, recall and precision, are used here.

An indicator of the lack of FN errors is defined by the *Recall* statistic [13]:

$$Recall = \frac{TP}{TP + FN} \tag{1}$$

Thus, recall measures the percent of relevant studies that are retrieved, and is a measure of completeness. A maximum recall measure of 1.0 signifies the inclusion of all relevant studies, with recall measures less than one indicating increased FN errors.

An indicator of FP errors is found in the measure of precision [13]:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Thus, precision measures the percentage of retrieved documents that are relevant, and is an indicator of exactness. A maximum precision measure of 1.0 signifies the exclusion of all irrelevant studies, with measures less than one indicating an increase in the number of FP errors.

Research process efficiency is also an important metric to assess, as required research effort may be an impediment to the undertaking of such research and may reduce the accuracy of results due to fatigue. Using the approaches proposed in this research, the researcher analyzes a subset of the search corpus (hereafter referred to simply as a corpus) to train the classifier, deciding which studies within the training set are relevant (included) and irrelevant (excluded). Once the classifier is applied, the researcher completes the Study Selection step with only the algorithm-Included studies, manually removing those included erroneously by the classifier.

In this context, researcher efficiency savings are obtained by reducing the number of irrelevant studies that must be read by the researcher, by reducing the proportion of studies in the algorithm training set and reducing classifier FP errors. This is consistent with prior research which argues that the time required to review articles is a target for optimization [see 14,15]. Therefore, this research defines the efficiency savings as the percentage of study inclusion / exclusion decisions that the researcher must make in training the selection algorithm and within the classified include set.

$$Efficiency\ Savings\ = 1 - \frac{TSS + FP}{Corpus} \tag{3}$$

where:

- TSS = Size of the algorithm training set
- Corpus = the number of studies in the corpus, which is also equal to