On Searching Relevant Studies in Software Engineering

He Zhang
Lero, The Irish Software Engineering Research Centre
University of Limerick, Ireland
he.zhang@lero.ie

Muhammad Ali Babar
IT University of Copenhagen, Denmark
maba@itu.dk

BACKGROUND: Systematic Literature Review (SLR) has become an important research methodology in software engineering since 2004. One critical step in applying this methodology is to design and execute appropriate and effective search strategy. This is quite time consuming and error-prone step, which needs to be carefully planned and implemented. There is an apparent need of a systematic approach to designing, executing, and evaluating a suitable search strategy for optimally retrieving the target literature from digital libraries.

OBJECTIVE: The main objective of the research reported in this paper is to improve the search step of doing SLRs in SE by devising and evaluating systematic and practical approaches to identifying relevant studies in SE.

OUTCOMES: We have systematically selected and analytically studied a large number of papers to understand the state-of-the-practice of search strategies in EBSE. Having identified the limitations of the current ad-hoc nature of search strategies used by SE researchers for SLR, we have devised a systematic approach to developing and executing optimal search strategies in SLRs. The proposed approach incorporates the concept of 'quasi-gold standard', which consists of collection of known studies and corresponding 'quasi-sensitivity' into the search process for evaluating search performance. We report the case study and its finding to demonstrate that the approach is able to improve the rigor of search process in an SLR, and can serves as the supplements to the guidelines for SLRs in EBSE. We plan to further evaluate the proposed approach using several case studies with varying topics in software engineering.

Search strategy, quasi-gold standard, systematic literature review, evidence-based software engineering

1. INTRODUCTION

Systematic reviews (also referred as systematic literature reviews, SLRs) aim to identify, assess and combine the evidence from primary research studies using an explicit and rigorous method. This method has been widely implemented in some disciplines, such as medicine and sociology. Since their seminal paper of Evidence-Based Software Engineering (EBSE) was published in 2004 Kitchenham et al. (2004), systematic review has become an important methodology of EBSE, and many SLRs have been conducted and reported.

EBSE involves five distinct steps Dyba et al. (2005). The second step, '*search the literature for the best available evidence to answer the question*', builds the basis for evidence aggregation, appraisal and further integration with decision making practise. Kitchenham also states that the aim of an SLR is to find as many primary studies relating to the research questions as possible using an unbiased search strategy Kitchenham and Charters (2007). The rigor of the search process is one factor that distinguishes systematic reviews from traditional (ad hoc) literature reviews.

Similar to other disciplines, many researchers doing SLRs rely on searches of digital libraries for identification of relevant studies in software engineering (SE). However, these database searches have typically been designed using methods lacking in scientific rigor, instead often relying solely on investigator's past experience and knowledge of the subject matter Boynton et al. (1998). In practice, identifying primary studies can be difficult for several reasons, including inadequate search strategy, heterogeneity of language describing the subject matter, and limited range of indexing terms describing study methodology Dickersin et al. (1994). Though Biolchini et al. suggest evaluating search engines to verify if they are capable of executing search strings during the planning phase Biolchini et al. (2005),

no concrete strategy has been provided for search strategy evaluation.

Despite the current state that neither the above EBSE papers nor the SLR guidelines include the practical instructions about how to improve and evaluate the rigor and performance of a search strategy, some issues relating to literature search in SE have emerged and been reflected in SLR reports, such as

- *How to design a rigorous search strategy that maximises the collection of relevant studies?*
- *What are criteria of an affordable and reliable strategy to effectively balance the search sensitivity (quality) and precision (effort)?*
- *Is it possible to evaluate a predefined search strategy and corresponding search strings?*

Moreover, the latest version of guidelines Kitchenham and Charters (2007) also encourage software engineering researchers to develop and publish such strategies including identification of relevant digital libraries. Hence, there is a need for validated search strategies for SLRs that optimise retrieval of relevant studies from digital libraries and electronic databases for researchers and practitioners. This paper attempts to serve as a preliminary response to this need. We have devised a systematic and practical approach for search strategy development in order to improve the rigor of search processes in SLRs. This approach also strives to balance the retrieval of validated set of relevant studies in SE and the effort consumed in this phase.

This paper is structured as follows. Section 2 introduces concepts related to search strategies for SLRs. In Section 3, we describe a systematic and practical approach for implementing a relatively rigorous literature search. This search approach is then demonstrated by a '*replicated*' search (an observer-participant case study) and compared to its original SLR in Section 4. Finally, some discussion and our conclusion are presented in Section 5.

2. SEARCH STRATEGY IN SYSTEMATIC LITERATURE REVIEWS

2.1. Defining Search Strategy

A necessary and crucial step of SLR is the identification of as much relevant literature to research questions as possible. Search strategy, which defines the methods to retrieve the relevant literature, has been developed in many ways, but the typical approach can be for information professionals (in subject matter) to use their combined knowledge of databases (digital libraries), search techniques, thesauri and the field of interest, to explore, often iteratively, combinations of terms which capture the concepts of interest White et al. (2001). An optimum search strategy is expected to provide effective solutions to a series of questions for search process in SLR:

1. ***Which*** *approach to be used in search process (e.g. manual or automated search)?*
2. ***Where*** *(source or venue) to search, and which part of article (field) should be searched?*
3. ***What*** *(subject, evidence type) to be searched, and what are inputs (search strings) to search engines?*
4. ***When*** *is the search carried out, and what time span to be searched?*

Which approach(es)? The guidelines, Biolchini et al. (2005), Kitchenham and Charters (2007), all emphasise the literature search through web search engines provided by digital libraries, i.e. automated search. However, in practice, many reported SLRs also employed manual search, alone or combined with automated search, in specific sources (e.g., Jorgensen and Shepperd (2007)).

In manual (hand) search, investigators scan the sources (e.g., journals or proceedings) paper by paper and issue by issue. This search method may ensure the capture of relevant studies in the specified sources, but in the meantime, consumes much effort in examining many irrelevant studies. Instead, automated search uses search strings, which represent the identifiers of the subject, to retrieve results from search engines (digital libraries). Compared to manual search, this method is more efficient, but its performance depends on the quality of search strings, capability of search engine, and diversity of the subject.

Where to search? 'Search source' was used as a general term for where relevant studies can be retrieved. We use 'search source' distinct from 'search engine' in defining search strategies. As automated search always retrieves results from search engine, in contrast, the former is dedicated for the sources specified in citations (e.g., journals and proceedings) in this paper, they are specified and scanned in manual search. As illustrated in Figure 1, generally speaking, there is a many-to-many relationship between them: one engine can cover multiple sources, while one source may also be retrievable from more than one engine.

What into search? Subject and article type, which are normally defined in protocol, are two important filters to remove irrelevant studies and low quality studies. For SLRs in SE, the most used subjects are 'computer science' and 'software engineering'. Search strings, which are connected with logic operators, are inputs to search engines in automated search. This paper

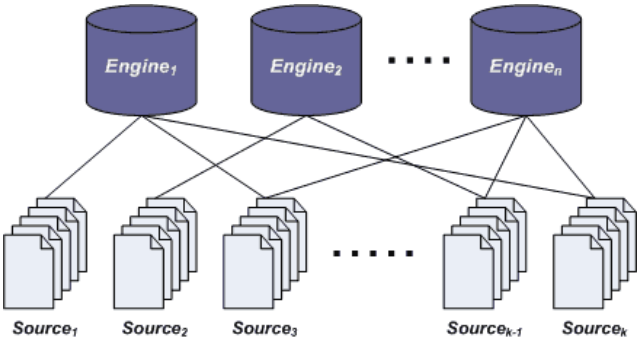

Figure 1: *Search sources and engines*

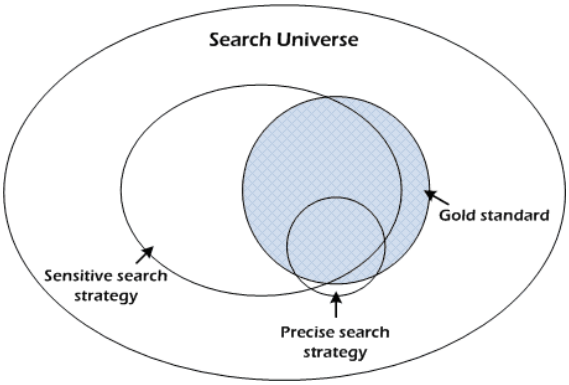

Figure 2: *Search sensitivity, precision, and gold standard*

proposes a systematic search approach that improves search string development and evaluation.

When and what time span to search? Time span of the studies in search is determined by the purposes of an intended SLR and its focused research questions. For example, trend analysis for a given period, or synthesis of collection of full evidence for answering a specified question. As it normally takes at least months from the initial search to the appearance of an SLR for public access, the search date(s) should be addressed in the report as well, i.e. when the search was conducted?

2.2. Evaluating Search Strategy

Subjective vs. objective evaluation. The performance of a search strategy can be evaluated by examining the answers to the above search design questions and the results retrieved from the search process in which the strategy applies. Roughly speaking, the evaluation is implemented in subjective and/or objective forms.

In subjective evaluation, some external experts review the predefined search strategy as a part in an SLR protocol before the stage of *conducting the review*. After the automated search, some pre-indicated studies (based on expert's awareness of domain knowledge) are compared to the search results. However, the reliability of subjective evaluation highly relies on their personal knowledge in the specific domain, which is difficult to be quantified. Apart from the subjective approach, objective evaluation employs a set of quantitative criteria to assess performance of a search strategy.

Sensitivity vs. precision. Two important criteria borrowed from medicine can be used for evaluating the quality and efficiency of a search strategy. *Sensitivity* for a given topic is defined as the proportion of relevant studies retrieved for that topic and *precision* is the proportion of retrieved articles that are relevant studies. Figure 2 shows different search strategies within search universe and the relation with *gold standard*.

In automated search, given search strings, the selected search engine (library) retrieves a certain amount of results (studies). Then the *sensitivity* and *precision* corresponding to the search strings and engine can be calculated as:

$$Sensitivity = \frac{Number\ of\ relevant\ studies\ retrieved}{Total\ number\ of\ relevant\ studies} 100\% \quad (1)$$

$$Precision = \frac{Number\ of\ relevant\ studies\ retrieved}{Number\ of\ articles\ retrieved} 100\% \quad (2)$$

Gold standard. The '**gold standard**' represents, as accurately as possible, the known set of identified primary studies in a collection according to the definition of research questions in an SLR. Gold standard normally plays two distinct roles in the evaluation framework. For SLRs, it is assumed to be *truth* in appraising the sensitivity of a search strategy; it is also a source of training samples for refining search strings White et al. (2001). In practice, it may be appropriate to bifurcate the gold standard for these two purposes.

A highly *sensitive* search strategy will retrieve most of the studies in *gold standard*, but may also retrieve many unwanted articles (Figure 2). A highly *precise* search strategy will retrieve only a small portion of irrelevant articles, but may miss a large number of papers in *gold standard*. A perfect search strategy would be 100% sensitive as well as 100% precise, capturing exactly the gold standard without any irrelevant ones.

Gold standard has been used for improving literature search in systematic reviews in other disciplines, such as in medical and clinical research and social science Dickersin et al. (1994) and White et al. (2001). Nevertheless, as the retrieval of a *real* gold standard is impossible for most systematic reviews, this paper instead introduces the concept of '**quasi-gold standard**' that is a set of known studies from related literature sources identified to the research topic.

2.3. State of the Practice

Since the introduction of EBSE and SLR, the number of SLRs in SE has been growing rapidly. This subsection briefly summarizes the state-of-the-practice of search strategies in EBSE from the above aspects.

2.3.1. Automated search vs. manual search
To investigate the realistic implementation of search strategies in EBSE, we conducted a search of SLRs published in SE, which extends the SLR search reported in Kitchenham et al. (2009) with the updated records by the end of 2008. This up-to-date SLR search identified 38 SLRs. The search results consists of 68% (26 out of 38) reported studies using automated searches in their SLRs; 39% (15 out of 38) using manual search; and 26% (10 out of 38) combining the both. Several SLRs did not report the search method they used, or were conducted based on the studies identified by other SLRs, such as Hannay et al. (2007).

2.3.2. Search engines and search sources
Table 1-a summarizes 11 engines (digital libraries) used more than once in SLRs for searching relevant studies in SE, which are ranked in order of their frequencies. Among them, IEEE Xplore and ACM Digital Library are the main search portals for most SLRs in SE. Table 1-b lists top sources for manual search used twice or more in SLRs. The sources related to SE in general (e.g., IEEE Software, TSE, ICSE) and empirical software engineering (e.g., ESEM, ISESE) were most used in manual search in the previous SLRs.

Table 1: *Search engines and sources*

Rank	Search engine	# of SLRs	% of SLRs
1	IEEE Xplore	24	92%
2	ACM digital library	21	81%
3	ScienceDirect	15	58%
4	ISI Web of Science	10	38%
5	EI Compendex	9	35%
6	SpringerLink	8	31%
6	Wiley InterScience	8	31%
6	Inspec	8	31%
9	Google Scholar	6	23%
10	SCOPUS	2	8%
10	Kluwer	2	8%

(a) search engines used more than once

Rank	Search source	# of SLRs	% of SLRs
1	IEEE Software	4	27%
1	ESEM	4	27%
1	ISESE	4	27%
4	TSE	3	20%
4	ICSE	3	20%
4	JSS	3	20%
4	IEEE Computer	3	20%
8	Metrics	2	13%
8	TOSEM	2	13%
8	ESE	2	13%
8	WWW	2	13%
8	ICSM	2	13%
8	MISQ	2	13%

(b) search sources used more than once

2.4. Related Work in Software Engineering

Some previous researchers have discussed the issues related to literature search in software engineering. Brereton et al. (2007) identified several issues of electronic search derived from their experience in conducting SLRs. For instance, researchers must select and justify a search strategy that is appropriate for their research questions; primary studies could not be retrieved from single source, etc.

Dieste and Padua (2007) investigated the optimal search strategies using the combination of alternative search strings for automated search in SLR. Nevertheless, the '*gold standard*' used to calculate sensitivity was established from the studies already identified in another SLR by Sjoberg et al. (2005). In most cases of SLR, such a '*gold standard*' is impossible to be accessed by researchers in the *planning stage* of their intended SLRs. In other words, a '*gold standard*' in this case provides no help to search strategy evaluation, and to ensure the retrieval quality of relevant studies in SLRs.

So far, to the authors' knowledge, neither comprehensive definition and rigorous development method of search strategy nor practical evaluation approach has been developed for retrieving relevant studies in SE.

3. QGS BASED SCIENTIFIC SEARCH APPROACH

Based on the concept of Quasi-Gold Standard (**QGS**), this section constructs a systematic, scientific, and also practical literature search approach for SE, which provides capability for search strategy development and evaluation.

3.1. Mechanism and Overview

To avoid the possible limitations of applying single search method (automated or manual) in SLR and to provide a practical and relatively rigorous method for search string evaluation, we propose a systematic literature search approach, as complement to SLR guidelines, in support of retrieval of relevant studies. It recommends that an optimum search strategy should be an effective integration of manual and automated searches, which support each other.

3.1.1. QGS: quasi-gold standard
In terms of our observation (that is confirmed with the results from the case study), most reported SLRs in SE developed their search strategies subjectively. Even for the well-conducted SLRs, search strategies were developed by teams with expertise and tested on collections of '*well-known*' samples to assess the search performance. Unfortunately, such preset '*well-known*' samples cannot replace the *gold standard* for evaluation, as a full set of primary studies is impossible to be accessed prior to the execution of an SLR.

Instead, we introduce the concept of '**quasi-gold standard**', which is a set of known studies from key sources, e.g., domain-specific proceedings and journals recognized by the community in the subject, for a given time span. Note that compared to a *gold standard*, there are two more constraints associated with a '*quasi-gold standard*': venues (where) and period (what time). In other words, a 'quasi-gold standard' can be regarded as a 'gold standard' in the conditions where these constraints apply. Accordingly, a more objective

method for devising and testing search strategies is developed and integrated into a systematic search process, which may rely on an analysis of information from the available records (QGS) rather than subjective input from searchers' perceptions (like some SLRs did). On the other hand, for the subjective approach of search string design, QGS can also be used for evaluating the search strategy (see Section 4).

Figure 3 shows the mechanism underpinning the proposed search approach. The results (studies) from manual search are used for establishing a QGS, which can further elicit the search strings for automated search, or later evaluate the search strategy. In the opposite direction, automated search complements manual search, expands the coverage and capture of most relevant studies in a relatively rigorous form.

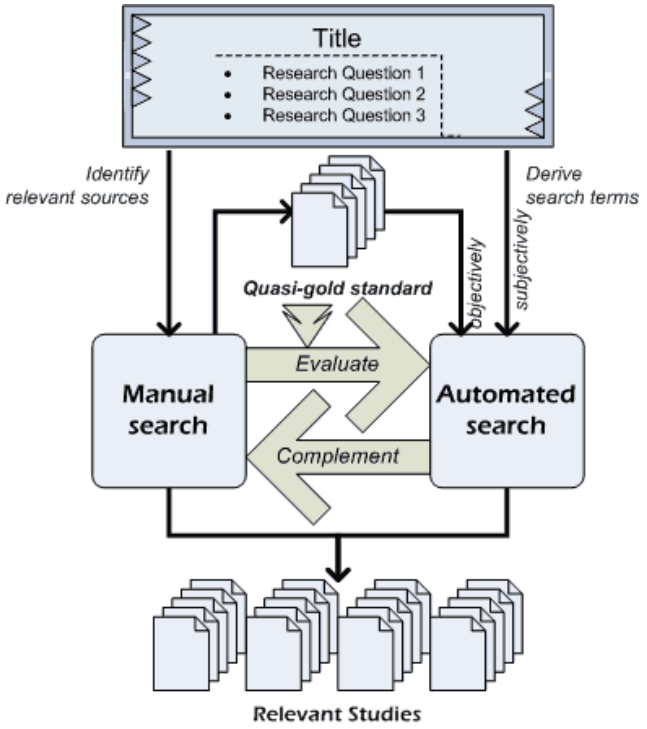

Figure 3: *Mechanism underpinning the approach*

3.1.2. Approach overview

Figure 4 presents an overview of the proposed search approach, which starts with identifying sources for manual search and engines (libraries and databases) for automated search. The QGS is established by performing manual search in the selected sources, and the identified studies are then grouped by their residing libraries and databases.

The design of search string can be in a subjective or objective form. In subjective approach, the search strings are argued by researchers according to their knowledge in the subject (like many previous SLRs), then tested by the 'quasi-gold standard'. The objective method elicits search strings automatically from articles in the QGS

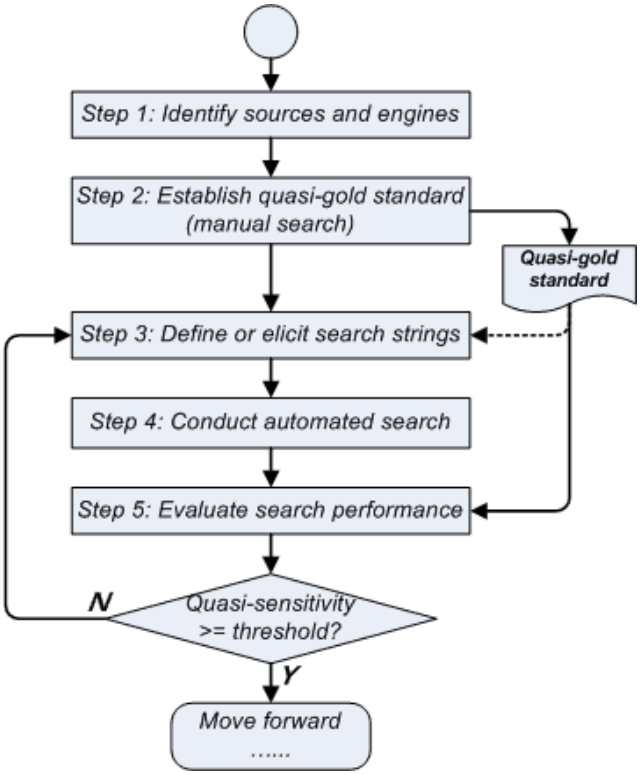

Figure 4: *Proposed scientific search process*

through word frequency or content analysis tools. These search strings are inputs to automated search, and results will be combined with the QGS once they are assessed as '*acceptable*' in evaluation.

3.2. The Search Process

3.2.1. Step 1: Identify related sources and engines

The literature search process starts at the identification of the sources (venues) of relevant publications. In SE, many digital libraries are available for automated search, and even more sources for manual search.

Select sources for manual search. Research questions for an SLR are motivated by the research in a particular subject matter (domain) in SE. For an experienced and knowledgeable researcher working in this area, the *related* domain-specific sources can be identified without much difficulty. These sources consist of a collection of proceedings of the conferences specialized in that domain and major journals where the community often publishes their research.

As manual search is time-consuming, a large number of selected sources may lag behind the overall progress of SLR. In order to improve the efficiency of manual search, as well as to secure the quality of QGS, the nominated sources for manual search also need to be evaluated by independent experts in this domain, and any emerging disagreements must be resolved before the next step.

Select library engines for automated search. The selection depends on the distribution of related sources across libraries, the coverage and overlapping among them, and their accessibility to the searchers. Whereas, by observing the most reported SLRs, IEEE Xplore and ACM Digital Library become the must-have literature portals that are recommended for consideration of any automated search of future SLRs in SE.

Given the many-to-many relationship between search sources and engines (Figure 1), an optimum combination of both should cover a maximum number of sources with a minimum set of search engines (libraries), in other words, eliminate as much overlapping as possible.

3.2.2. Step 2: Establish QGS

The manual search is conducted by screening all articles, one by one, published in the selected sources (e.g., proceedings and journals) and during a given period. The *title-abstract-keywords* fields of a paper are first checked. The inclusion and exclusion criteria should be explicitly defined in advance. As recommended in the guidelines Kitchenham and Charters (2007), the reliability of inclusion decision should be assessed using the Kappa statistic between researchers, or reviewed by an external panel. If selection decision could not be made, the other fields (like conclusion or even full text) need to be further examined.

One important assumption underlying the manual search processes in the previous SLRs is that all relevant studies within the indicated sources could be identified by carefully screening all the articles. Hence, once the screening is completed and agreement on the selection is reached, all these identified studies are used to form the QGS.

As quasi-gold standard is *source-* (engine) and *period-specific*, the sources selected in Step 1 can also be grouped by search engines. For a large scale SLR, in addition to an overall QGS, this step may produce more than one subset of QGS, each of which corresponds to one dedicated search engine. They enable testing search string's performance for individual engine.

3.2.3. Step 3: Define or elicit search strings

Since search strings for automated search can be defined based on subjective expertise or elicited from the '*quasi-gold standard*', the search process bifurcates at this step.

Subjective search string definition. Most previously reported SLRs in SE performed automated search in a subjective form. The reviewers defined their search strings based on their domain knowledge and past experiences. Though the strings they choose can be evaluated later by QGS, in the subjective approach, it would be inspected by experts in the subject to reduce the number of possible iterations and further save effort. Figure 4 displays there might be backward link from the 'decision' to Step 3. In this case, the set of search terms has to be refined or enriched in order to capture more samples included in QGS through next round of automated search.

Objective search string elicitation. One of the uses of QGS is to elicit the recommended search strings using text mining. In the objective approach, a frequency analysis of citation information of the studies in QGS is undertaken followed by a statistical analysis of the most frequently occurring words or phrases. This analysis determines which terms would best distinguish relevant studies from irrelevant ones.

Some textual analysis packages, such as SimStat and WordStat Provalia (2009), are able to facilitate the identification of the frequently occurring terms in particular items of studies. For instance, the *title-abstract-keywords* of the papers in QGS are imported into the analysis software for frequency analysis. This may produce all the words or phrases being ranked according to the number of records in which each word appears by case. This technique is able to identify the candidate search terms with exception of some stop words which are deliberately excluded White et al. (2001) (e.g., 'the' and 'of').

Note that although the statistical software for textual analysis can help the search string elicitation, especially for a large scale QGS, subjective judgement might also be needed to finally construct the string to automated search based on the frequency list generated through the computer aided analysis.

3.2.4. Step 4: Conduct automated search

This step uses the strings for automated search, which are (subjectively) defined or (objectively) elicited. As the search syntax varies between search engines, the search strings need to be coded correspondingly in advance by following the specific syntax and criteria of each search engine (library). Given the capability limitations of some search engines (for example ACM Dyba et al. (2007)), the automated search sometimes has to be implemented by splitting the combination of search terms into multiple simple ones. Note that due to the overlapping (such as between IEEE and ACM), the duplicate studies retrieved from different search engines also need to be identified and removed in this step.

3.2.5. Step 5: Evaluate search performance

If the search strings for automated search are defined in the subjective approach, the search results need to be evaluated for securing the quality of automated search.

Calculate 'quasi-sensitivity'. In EBSE, missing important studies from an SLR may lead to the generation

of inaccurate evidence. Accordingly, compared to *precision*, *sensitivity* becomes the top criteria considered when evaluating the search performance in most SLRs. Unfortunately, as the *gold standard* for the subject is unknown, the corresponding *sensitivity* cannot be calculated (Equation 1) at this stage. Whereas, our search approach uses the *quasi-gold standard* (from the manually selected sources) to measure *sensitivity* instead of the search universe (Figure 2).

Researchers calculate the number of relevant studies retrieved from the selected sources (Step 1) through automated search (Step 4). Obviously, this number must not be greater than the number of studies identified in Step 2. Divided by the pool size of QGS, the corresponding '*quasi-sensitivity*' can be calculated.

Evaluate performance. The **quasi-sensitivity** could be 100% or less. It needs to be compared against a rational threshold to finally determine if the performance of automated search is acceptable. Although *sensitivity* and *precision* are the important criteria for evaluating search strategies and a tradeoff is always being pursued between them in search strategies, a high *sensitivity* is usually more desired than a high *precision* in terms of the goals of SLRs.

Table 2 displays the search strategy scales used for evaluating search terms in Dieste and Padua (2007), which was inferred from the sensitivity and precision ranges of SLRs in medicine. Based on the scales, we suggest a threshold between 75% and 85% as a reference for sensitivity evaluation of search performance.

Table 2: *Search strategy scales*

Strategy	Sensitivity	Precision	Comments
High recall	85-90%	7-15%	max sensitivity despite low precision
High precision	40-58%	25-60%	max precision rate despite low recall
Optimum	80-99%	20-25%	maximize both sensitivity & precision
Acceptable	72-80%	15-25%	fair sensitivity & precision

For example, if we choose 80% as the threshold for search string evaluation, then

$$\textit{quasi-sensitivity} \begin{cases} \geq 80\%, & \text{then, move forward...,} \\ < 80\%, & \text{then, go back to Step 3.} \end{cases} \quad (3)$$

If the search performance is considered acceptable (quasi-sensitivity $\geq$ 80%), the results from the automated search can be merged with the 'quasi-gold standard', and the search process terminates. Otherwise, the process has to go back to Step 3 for search string refinement, which may form an iterative improvement of search strings until the performance becomes acceptable.

4. CASE STUDY

This section investigates the proposed search approach using a participant-observer case study (defined by Yin (2003)), in which the literature search of a published SLR is performed and compared.

4.1. The Original SLR

In order to avoid any subjective bias during the search and screening process, the original SLR should be carefully selected as the reference. Some criteria were applied:

1. Relevant studies can be identified with minimum possible ambiguity. That minimizes the subjective bias due to knowledge difference between the researchers in the original and the replicated searches.
2. The articles in the original SLR must be explicitly constrained in definite time frame. Some SLRs with search end date open '*to present*' are excluded here.
3. The publication that reports SLR must include the list of identified studies, which may enable a detailed comparison with the results from the replicated search.

In terms of the above criteria, The SLR by Kitchenham et al. (2009) that summarizes and reports the impact of SLRs in software engineering is selected as reference in the case study. This SLR performed a manual search in 13 sources with explicit time span from Jan 2004 through mid of 2007. As an SLR is a type of secondary study, their work can be regarded as a tertiary study. It retrieved 34 relevant studies, among which 20 SLRs were identified as secondary studies.

4.2. Search Implementation

4.2.1. Identification of search sources and engines

At manual search stage, we chose the sources (journals and proceedings) related to empirical software engineering (ESE) and EBSE. By carefully considering the sources available in SE community, 9 of them were selected by the authors for this study (Table 3). Note that the selected sources for manual search in this paper are different from the original SLR somehow for two reasons: (1) though the replicated search strategy is designed for the same research questions, the authors may have slightly different recognition of the '*related*' sources from the original researchers; (2) the purpose of the manual search in this case study is to establish the *quasi-gold standards*, rather than to strive to capture as many relevant results as possible. Therefore, some originally used sources were ignored at manual search stage, and two additional sources, EASE and ESEM, were added into the list in terms of their tight linkage to EBSE.

The nominated sources were grouped into 5 libraries (Table 3), 4 of which were selected for the automated search, i.e. IEEEXplore, ACM Digital Library, ScienceDirect and SpringerLink. Note that other libraries can be employed for automated search, but the QGS is only valid for evaluating the search through them.

4.2.2. QGS and automated search

In this case study, the searched articles should be 'systematic reviews in software engineering'. Accordingly, we refined the inclusion and exclusion criteria reported in the original SLR Kitchenham et al. (2009). Two researchers screened all papers published in the sources from 2004 to 2008 in manual search independently until reached joint agreements on all included studies. In total, 21 studies were retrieved and 20 of them were used for building the QGS. Table 3 shows the source names and their numbers of relevant studies (by 2007 and 2008).

Table 3: *Selected sources for manual search*

Source	Library/publisher/engine	2007 mid	2008 end
TSE	IEEE	4	4
IEEE-SW	IEEE	1	1
ESEM('07,'08)	IEEE/ACM	0	2
ISESE('04-'06)	IEEE/ACM	2	2
Metrics('04,'05)	IEEE	0	0
IST	Elsevier	2	7
JSS	Elsevier	2	2
EMSE	Springer	0	2
EASE('06-'08)	IEE/BCS	0	1
Total		11	21

The case study implemented automated search by following the *subjective* definition approach, in which the search strings are nominated based on the authors' knowledge relating to the subject of EBSE, and their observation of the studies included in the QGS. As we were looking for SLRs in SE, We intuitively initiated the automated search with the string `(software AND systematic AND review)` into the fields of *title-abstract-keywords* through the above engines. The search strings then were coded to fit the syntax requirements and capability of each engine.

4.2.3. Evaluation and refinement

Table 4 summarizes the number of studies retrieved by each database with the initial and refined search strings. For example, there are 12 studies retrieved by IEEE Xplore, 5 in the QGS. In total, 13 studies in QGS were retrieved in the initial automated search. In terms of the sample size of QGS, the 'quasi-sensitivity' was calculated to be 65%(13/20), which is unacceptable compared to the threshold (80%). As defined in Step 5, the search process had to go back to improve the string.

By carefully checking the studies included in QGS but ignored in the initial automated search, we found most of them published in the early years in the period (2004-2008) when the method 'systematic review' was just introduced to SE. Their authors claimed the review studies using other terms (e.g., 'survey'). So we refined the string as `(software AND (systematic OR controlled OR structured OR exhaustive OR comparative) AND (review OR survey OR ''literature search''))`, then performed the automated search again.

Table 4: *Results from automated search*

Search engine	#Results	#In quasi-gold	#Identified
Initial search			
IEEE Xplore	146	5	12
ACM digital library	34	1	6
DirectScience	31	6	6
SpringerLink	42	1	6
Overall	*253*	*13*	*30*
Refined search			
IEEE Xplore	270	8	15
ACM digital library	160	1	6
DirectScience	82	7	7
SpringerLink	145	1	6
Overall	*657*	*17*	*34*

The revised automated search is able to capture 17 studies included in the quasi-gold standard, which increases the 'quasi-sensitivity' up to 85% (i.e. acceptable). By combining the studies from manual search, the proposed search approach finally retrieves 38 SLRs for the tertiary study.

4.3. Performance Comparison

Although the similar inclusion and exclusion criteria are employed in both the original and this replicated searches, we exclude several '*relevant*' studies that were selected in the original SLRs during the manual search and selection due to the deviation caused by how strictly the inclusion/exclusion criteria were followed.

Because of the disagreement between the original and the replicated searches, we cannot directly compare the numbers of identified studies from them given the page limit. Instead, we focus on the comparison of performance between the implementations of different search strategies. Table 5 shows the study numbers retrieved by following different strategies for the same research questions. The row headed with 'manual only' indicates how many studies can be identified if manually searching the sources given in Kitchenham et al. (2009) from 2004 till 2008. Two more SLRs could be found when screening their specified sources (more than our sources in manual search). The 'automated only' row shows the search performance by search engines but without refinement; the bottom row presents the results through the QGS based systematic search approach.

Table 5: *Comparison among 3 strategies*

Method	SLRs identified	Quasi-sensitivity
Manual only	22	n/a
Automated only (initial)	30	65%
Systematic	38	85%

5. DISCUSSION

The limitations of applying automated or manual search alone are illustrated in the case study. Manual search is difficult to scan a large number of sources within a limited effort; on the other hand, the performance of automated search highly relies on the quality of search string, which may need continuous refinement in most cases. Although some previous SLRs employed both methods, most of them simply merged the search results only. In contrast, the QGS based systematic search approach not only combines their results together, but establishes linkage between them for supporting each other with their own advantages. This approach also suggests quantitative measurement for when you can stop the iterative refinement of automated search, and captures considerable identified studies with reasonable effort.

Some *secondary* studies related to a research topic (subject matter), which have been screened and filtered already by external researchers, could be introduced into *quasi-gold standard* to further reduce the effort in manual search. For instance, some previous SLRs directly used studies identified by Sjoberg et al. (2005) as their full set of primary studies. As another example, the results from the mapping study by Jorgensen and Shepperd (2007) can be used to build QGS for more specific SLRs in software cost estimation. In such cases, the results may need to be tailored in terms of *subject* and *time* that conform to the new SLR.

As an alternative to search engine based search strategy, reference list based search strategy can be another option for retrieving relevant studies. This strategy was innovated with the concepts of co-citation and bibliographic coupling Skoglund and Runeson (2009). However, as most of the major digital libraries in SE are not designed for supporting this kind of search, it is very time-consuming in manually retrieving studies from reference list. Thus this search approach is not yet practical enough at present in software engineering, but is suggested as a supplementary source for a full SLR by Kitchenham and Charters (2007).

As '*sensitivity*' is the top priority in defining search strategies in most SLRs, another criteria '*precison*' is less discussed here due to the page limit. It is however important to measure the productivity of search process.

As automated search mostly consults the fields of *title-abstract-keywords*, the search performance is also related to the quality and structure of these fields. An indicative title/abstract will increase search sensitivity. Budgen et al. (2008) investigated the possible influence of the quality of abstract to SLRs by experiments, and suggested *structured abstract* for improving understanding and study identification, which may further improve the search accuracy.

6. CONCLUSION

Systematic literature reviews have become an important empirical research methodology in software engineering, and more and more SLRs are being conducted and reported. In SLR, an effective and rigorous literature search takes a critical role in evidence aggregation. In order to enhance the rigor and comprehension of methodology, with reference to the experience of SLRs in other disciplines (e.g., medicine and sociology), this paper proposes a systematic search approach based on the concept of *quasi-gold standard* for retrieving and identifying relevant studies in software engineering. The major contributions can be concluded as

- Provide a clear scope of search strategy and its evaluation in searching relevant studies in SE.
- Introduce the concepts of '*quasi-gold standard*' and '*quasi-sensitivity*' for developing and evaluating the search strategy for a given SLR.
- Propose a systematic, scientific, and rigorous approach for practical search strategy development, implementation and evaluation.

Although the QGS based literature search approach is proposed for improving the search processes in SLRs and EBSE, it can be used in other literature reviews in SE, and benefit the researchers and practitioners who intend to retrieve a relatively comprehensive collection of relevant studies (for the *subject* and *time* given) within reasonable effort.

Currently this approach is being effectively applied in some systematic reviews in SE. We will continue the evaluation and improvement of this approach by conducting more case studies (with the objective and subjective search string elicitation methods) on varying topics in software engineering. In addition, the future methodological work in ESE and EBSE community may include to identify other issues and limitations of the SLRs reported in software engineering, and further to suggest practical improvements to the guidelines of systematic literature reviews.

ACKNOWLEDGEMENTS

This work was supported, in part, by Science Foundation Ireland grant 03/CE2/I303 1 to Lero - the Irish Software Engineering Research Centre (`www.lero.ie`).

7. REFERENCES

Boynton, J. and Glanville, J. and McDaid, D. and Lefebvre, C. (1998) *Identifying Systematic Reviews in MEDLINE: Developing An Objective Approach to Search Strategy Design*, Journal of Information Science, 24(3), 137-154.

Brereton, Pearl and Kitchenham, Barbara A. and Budgen, David and Turner, Mark and Khalil, Mohamed, (2007) *Lessons from Applying the Systematic Literature Review Process within the Software Engineering Domain*, Journal of Systems and Software, 80(1), 571-583.

Biolchini, Jorge and Mian, Paula Gomes and Natali, Ana Candida Cruz and Travassos, Guilherme Horta, (2005) *Systematic Review in Software Engineering*, Universidade Federal do Rio de Janeiro.

Dyba, T. and Dingsoyr, T. and Hanssen, Geir K. (2007) Applying Systematic Reviews to Diverse Study Types: An Experience Report. *In Proceedings of 1st International Symposium on Empirical Software Engineering and Measurement (ESEM'07)*, Madrid, Spain, September, pp. 225-234. IEEE Computer Society.

Dyba, T. and Kitchenham, Barbara and Jorgensen, M. (2005) *Evidence-Based Software Engineering for Practitioners*, IEEE Software, 22(1), 158-165.

Dieste, Oscar and Padua, Anna Griman. (2007) Developing Search Strategies for Detecting Relevant Experiments for Systematic Reviews. *In Proceedings of 1st International Symposium on Empirical Software Engineering and Measurement (ESEM'07)*, Madrid, Spain, September, pp. 215-224. IEEE Computer Society.

Dickersin, K. and Scherer, R. and Lefebvre, C. (1994) *Systematic Reviews: Identifying Relevant Studies for Systematic Reviews*, British Medical Journal, 309(6964), 1286-1291.

Hannay, Jo E. and Sjoberg, Dag I.K. and Dyba, Tore (2007) *A Systematic Review of Theory Use in Software Engineering Experiments*, IEEE Transactions on Software Engineering, 33(2), 87-107.

Jorgensen, Magne and Shepperd, Martin (2007) *A Systematic Review of Software Development Cost Estimation Studies*, IEEE Transactions on Software Engineering, 33(1), 33-53.

Kitchenham, Barbara and Brereton, O. Pearl and Budgen, David and Turner, Mark and Bailey, John and Linkman, Stephen, (2009) *Systematic Literature Reviews in Software Engineering: A Systematic Literature Review*, Information and Software Technology, 51(1), 7-15.

Budgen, David and Kitchenham, Barbara A. and Charters, Stuart M. and Turner, Mark and Brereton, Pearl and Linkman, Stephen G. (2008) *Presenting software engineering results using structured abstracts: A randomised experiment*, Empirical Software Engineering, 13(4), 435-468.

Kitchenham, Barbara and Charters, Stuart (2007) *Guidelines for Performing Systematic Literature Reviews in Software Engineering (version 2.3)*, Keele University and University of Durham.

Kitchenham, Barbara and Dyba, T. and Jorgensen, M. (2004) Evidence-Based Software Engineering. *Proceedings of 26th International Conference on Software Engineering (ICSE'04)*, Edinburgh, Scotland, May, pp. 273-284. IEEE Computer Society.

SimStat v.2.5 and WordStat v.5.1, (2009) Provalia Research, `http://www.provalisresearch.com/`.

Sjoberg, Dag I.K. and Hannay, Jo E. and Hansen, Ove and Kampenes, Vigdis By and Karahasanovic, Amela and Liborg, Nils-Kristian and Rekdal, Anette C. (2005) *A Survey of Controlled Experiments in Software Engineering*, IEEE Transactions on Software Engineering, 31(9), 733-753.

Skoglund, Mats and Runeson, Per (2009) Reference-based search strategies in systematic reviews. *Proceedings of 13th International Conference on Evaluation and Assessment in Software Engineering (EASE'09)*, Durham, England, April. BCS.

White, V.J. and Glanville, J.M. and Lefebvre, C. and Sheldon, T.A. (2001) *A Statistical Approach to Designing Search Filters to Find Systematic Reviews: Objectivity Enhances Accuracy*, Journal of Information Science, 27(6), 357-370.

Robert K. Yin (2003) *Case Study Research: Design and Methods* (3rd edn). Sage Publication.