Information and Software Technology 53 (2011) 625–637

Contents lists available at ScienceDirect

Information and Software Technology

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

Identifying relevant studies in software engineering

He Zhang [a,*], Muhammad Ali Babar [b], Paolo Tell [b]

[a] National ICT Australia, University of New South Wales, Australia
[b] IT University of Copenhagen, Denmark

ARTICLE INFO

Article history:
Available online 28 December 2010

Keywords:
Search strategy
Quasi-gold standard
Systematic literature review
Evidence-based software engineering

ABSTRACT

Context: Systematic literature review (SLR) has become an important research methodology in software engineering since the introduction of evidence-based software engineering (EBSE) in 2004. One critical step in applying this methodology is to design and execute appropriate and effective search strategy. This is a time-consuming and error-prone step, which needs to be carefully planned and implemented. There is an apparent need for 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 undertaking SLRs in software engineering (SE) by devising and evaluating systematic and practical approaches to identifying relevant studies in SE.
Method: We have systematically selected and analytically studied a large number of papers (SLRs) 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 SLRs, we have devised a systematic and evidence-based approach to developing and executing optimal search strategies in SLRs. The proposed approach incorporates the concept of 'quasi-gold standard' (QGS), which consists of collection of known studies, and corresponding 'quasi-sensitivity' into the search process for evaluating search performance.
Results: We conducted two participant–observer case studies to demonstrate and evaluate the adoption of the proposed QGS-based systematic search approach in support of SLRs in SE research.
Conclusion: We report their findings based on the case studies that the approach is able to improve the rigor of search process in an SLR, as well as it can serve as a supplement to the guidelines for SLRs in EBSE. We plan to further evaluate the proposed approach using a series of case studies on varying research topics in SE.

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 the publication of the seminal paper of Evidence-Based Software Engineering (EBSE) [25] in ICSE 2004, systematic review has become an important methodology of EBSE, and many SLRs have been conducted and reported.

EBSE involves five distinct steps [16]. 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 practice. Kitchenham and Charters [24] also state that the aim of an SLR is to find as many primary studies relevant to the research questions as possible using an unbiased search strategy. The rigor of the search process is one critical 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 identifying 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 [8]. 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 [11]. Though Biolchini et al. suggest evaluating search engines to verify if they are capable of executing search strings during the planning phase [7], no concrete instruction has been provided for search strategy evaluation.

Despite the current state that neither the above EBSE methodology papers nor the SLR guidelines include the practical instructions

* Corresponding author. Tel.: +61 2 9376 2227; fax: +61 2 9376 2023.
E-mail addresses: he.zhang@nicta.com.au (H. Zhang), maba@itu.dk (M.A. Babar), pate@itu.dk (P. Tell).

0950-5849/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.infsof.2010.12.010

about how to improve and evaluate the rigor and performance of a search strategy, some issues related to literature search in SE have emerged and been reflected in SLRs on different topics in SE, such as

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

Moreover, the most recent version of guidelines [24] also encourage software engineering researchers to develop and publish such strategies including identification of relevant digital libraries. Hence, there is an apparent need for validating search strategies for SLRs that optimize retrieval of relevant papers from digital libraries and electronic databases for researchers and practitioners. This paper purposes to contribute to the efforts aimed at addressing the above mentioned needs. 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 papers 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 and briefs the state-of-the-practice of literature search in SLRs in SE. In Section 3, we describe the proposed systematic and practical approach for implementing a relatively rigorous literature search. This search approach is then demonstrated and evaluated by two '*replicated*' literature searches (participant-observer case studies) and compared to their original SLRs in Sections 4 and 5 respectively. We discuss the findings from the case studies designed to assess the proposed systematic search approach and the threats to validity in Sections 6 and 7 respectively. They are followed by an overview of the related research in Section 8. Finally, Section 9 draws the conclusions of this paper.

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 [29]. 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*** *(venues or databases) to search, and which part of article (field) should be searched?*
3. ***What*** *(subject, evidence type) to be searched, and what are queries (search strings) fed into search engines?*
4. ***When*** *is the search carried out, and* **what time** *span to be searched?*

Which approach(es)? The guidelines [7,24] emphasise the literature search through web search engines provided by digital libraries, i.e. automated search. However, in practice, many reported SLRs in SE have also employed manual search, alone or combined with automated search, in specific venues (e.g., [19]).

In manual (hand) search, investigators scan the venues (e.g., journals or conferences) paper by paper and issue by issue. This search method may ensure the capture of relevant studies in the specified venues, but in the meantime, consumes a significant amount of 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 string, capability of search engine, and diversity of the subject.

Where to search? 'Search venue' was used as a general term for where relevant studies can be published and retrieved. We use 'search venue' distinct from 'search engine' in defining search strategies. As automated search always retrieves results from search engines [24], in contrast, the former is dedicated to the venues specified in citations (e.g., journals and conferences) in this paper, they are specified and scanned in manual search. As illustrated in Fig. 1, there exist many-to-many relationships between them: one engine can cover a number of venues, while one venue may also be retrievable from more than one engine.

What to 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 commonly used subjects are '*computer science*' and '*software engineering*'. Search strings, which connect keywords with logic operators, are inputs to search engines in automated search. This paper 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 purpose 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) several months from the initial search to the appearance of an SLR for public access, the search date(s) should be denoted in the report to make an SLR transparent and repeatable, 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. The evaluation can be 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

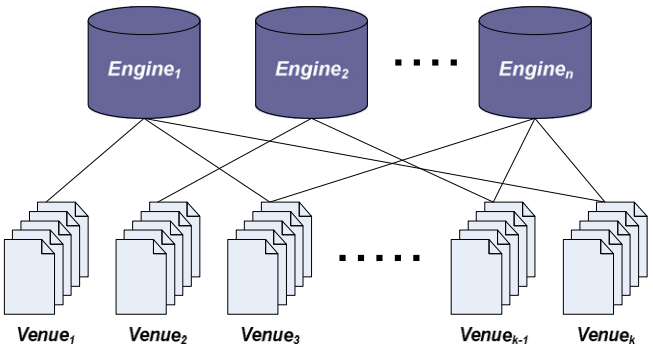

Fig. 1. Search venues and search engines.

subjective evaluation highly relies on their personal knowledge in a 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 and 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 studies that are relevant studies. Fig. 2 shows different search strategies within search universe and the relation with *gold standard* (being explained in this subsection).

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

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

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

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 [29]. 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 the *gold standard*, but may also retrieve many unwanted articles (Fig. 2). A highly *precise* search strategy will retrieve only a small portion of irrelevant articles, but may miss a large number of papers in the *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 medicine, clinical research and social science [11,29]. Nevertheless, it is not possible to have '*gold standard*' for most SLRs in SE. Accordingly, this paper introduces the concept of '*quasi-gold standard*' that is a set of known studies from related publication venues on a 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.

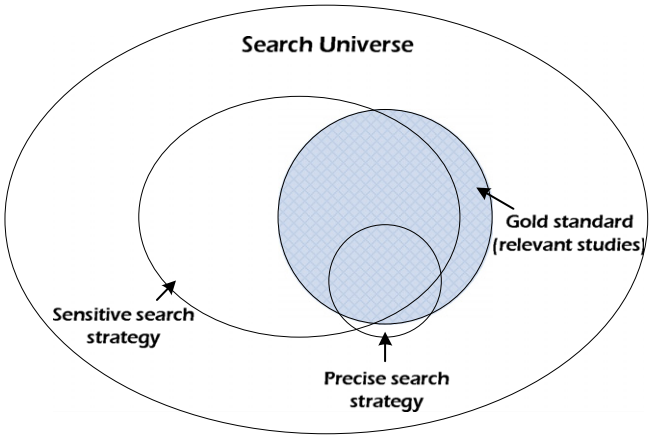

Fig. 2. Search sensitivity, precision, and gold standard.

Table 1
Search engines and venues.

Rank	Search engine	# of SLRs	% of SLRs
(a) *Search engines used more than once*			
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
(b) *Search venues used more than once*			
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

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 search reported in the tertiary study [22] with the updated records by the end of 2008. This up-to-date SLR search identified 38 SLRs (including systematic mapping studies). The search results consists of 68% (26 out of 38) reported SLRs using automated search; 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 [18].

Methodologically, the manual and automated searches were independently used in the existing SLRs in SE. Even for the SLRs using the both, they designed their manual and automated search respectively and simply combined the results from the both methods. These SLRs neither discover any relationships between the search methods (i.e. automated and manual), nor established the methodological linkage between and integration of them.

2.3.2. Search engines (digital libraries) and search venues

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

In addition, some SLRs employed more specific sources for their literature search, such as BESTweb [17] and university library services, which are dedicated to one subject in SE but sometimes not accessible for external researchers.

3. QGS-based systematic search approach

Based on the concept of quasi-gold standard (*QGS*), this section constructs a systematic, repeatable, and practical literature search approach for SE, which provides a mechanism 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 a complement to the existing SLR guidelines, in support of retrieval of relevant studies. It recommends that an optimum search strategy should be an effective integration of manual search and automated search, which are able to support each other (as illustrated in Fig. 3).

3.1.1. QGS: quasi-gold standard

In terms of our observation [32] (that is also confirmed by the results from the first case study), most of the 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, which highly depend on reviewers' knowledge on a subject matter, 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 the related venues, e.g., domain-specific conferences 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 (time span). 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 many 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).

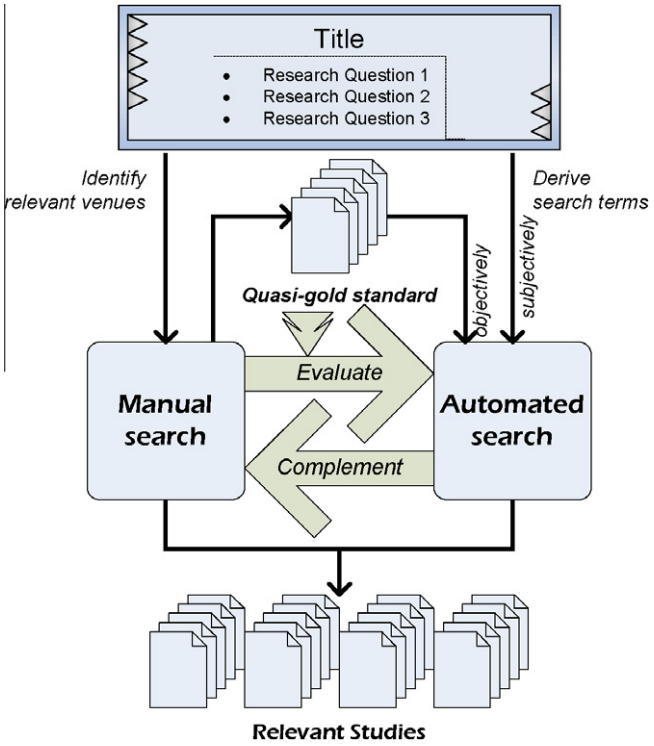

Fig. 3. Mechanism underpinning the systematic search approach.

Fig. 3 shows the mechanism underpinning the proposed search approach. The results (papers) 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 most of the relevant studies in a relatively rigorous form.

3.1.2. Approach overview

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

The design of search string can be in a subjective or objective approach. In subjective method, the search strings are devised by researchers according to their knowledge in the subject matter (like many previous SLRs), then tested by the 'quasi-gold standard'. The objective approach elicits keywords for search automatically from articles in the QGS through word frequency or content analysis tools. They are connected by logic operators and input to automated search, and search results will be combined with the QGS once they are assessed as '*acceptable*' in evaluation.

3.2. The systematic search process

The systematic literature search process proposed in this paper is composed of five steps.

3.2.1. Step 1: Identify related venues and databases

The literature search process starts with the identification of the relevant publication venues. In SE, many digital libraries are available for automated search, and even more venues for manual search.

Select publication venues for manual search. Research questions for an SLR are motivated by the research in a particular subject

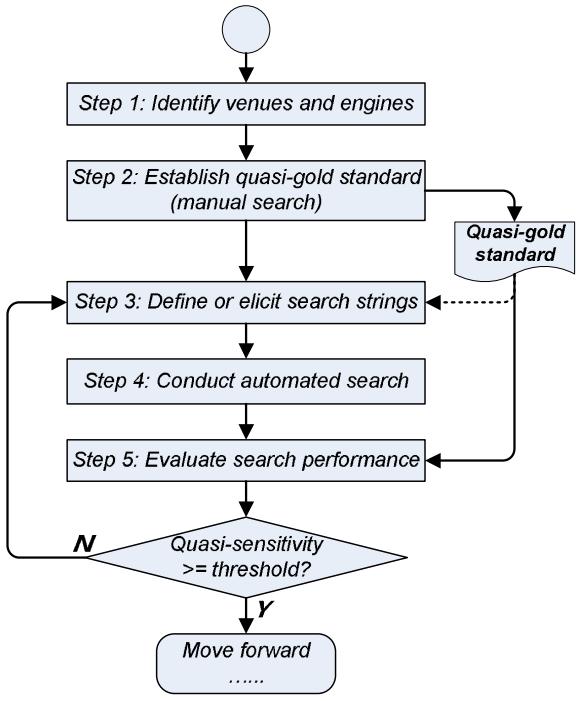

Fig. 4. Workflow of the proposed systematic search process.

matter (domain) in SE. For an experienced and knowledgeable researcher working in this area, the *related* domain-specific venues can be identified without much difficulty. These venues 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 venues 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 venues 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 libraries (databases) for automated search. The selection depends on the distribution of related venues across libraries, the coverage and overlapping among them, and their accessibility to searchers. Whereas, by observing the most existing SLRs, IEEE Xplore and ACM DigitalLibrary 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* relationships between search venues and search engines (Fig. 1), an optimum combination of both should cover a maximum number of search venues with a minimum set of search engines (libraries), in other words, eliminate as much overlap as possible.

3.2.2. Step 2: Establish QGS

The manual search is conducted by screening all papers, one by one, published in the selected publication venues (e.g., proceedings and journals) during a given time. The *title-abstract-keywords* fields of a paper are often first checked. The inclusion and exclusion criteria should be explicitly defined in advance. As recommended in the guidelines [24], 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 papers within the indicated venues 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 papers are used to form the QGS.

As QGS is *venue-* and *period-specific*, the venues selected in Step 1 can also be grouped by digital libraries (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 library.

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 used automated search in a subjective form. The reviewers defined their search strings based on their domain knowledge and past experiences. Though the strings they chose could be evaluated later by QGS, in the subjective approach, it would be inspected by experts in the subject matter to reduce the number of possible iterations and further save effort. Fig. 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 in QGS through next round of automated search.

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

Some textual analysis packages, such as SimStat and WordStat [3], 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 and their relations with exception of some stop words which are deliberately excluded [29] (e.g., 'the' and 'of').

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

3.2.4. Step 4: Conduct automated search

In this step, the digital libraries are searched using the strings, 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 DigitalLibrary [14]), 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 papers retrieved from different search engines also need to be identified and removed in this step.

3.2.5. Step 5: Evaluate search performance

The search results need to be evaluated against QGS for ensuring 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 (Eq. 1) at this stage. Whereas, our systematic search approach uses the *quasi-gold standard* (from the manually selected venues) to measure *quasi-sensitivity* instead of the search universe (Fig. 2).

Researchers calculate the number of relevant papers retrieved from the selected venues (Step 1) through automated search (Step 4). Obviously, this number must not be greater than the number of papers 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 up to 100% but is often 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 [13], which was inferred from the sensitivity and precision ranges of SLRs in medicine. Based on the scales, we suggest a threshold between 70% and 80% (acceptable) as a reference for sensitivity evaluation of search performance.

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

Table 2
Search strategy scales.

Strategy	Sensitivity (%)	Precision (%)	Comments
High sensitivity	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

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

If the search performance is considered acceptable (quasi-sensitivity $\geqslant$80%), the results from the automated search can be merged with the QGS, 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. However, in practice, the decision of an appropriate threshold need to consider a number of relevant factors to the characteristics of a given research topic (such as the diversity of subject, scope of research questions, and the scattering of the descriptive keywords). The values of the threshold may slightly vary from case to case.

The following two sections investigate the proposed systematic search approach using participant–observer case studies (defined by [30]), in which the literature searches of two published SLRs were replicated and compared. The two case studies chose to implement different search string definition methods: the first case study used the *subjective* string definition; the second implemented the *objective* terms elicitation.

4. Case study 1

The first case study was performed by the first two authors in order to formally trial the systematic search process. A participant–observer case study allowed us to access the case information without any barrier [30]. The original search of a tertiary study in software engineering was replicated in this case study, and the *subjective* search string definition method was applied.

4.1. The original SLR-1

In order to avoid any subjective bias during the search and screening process, the original SLR-1 should be carefully selected as the reference to this case study. Some criteria were applied:

1. The search strategy and search venues and/or databases applied in the original SLR must be explicitly described in the published SLR report.
2. 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.
3. The articles included in the original SLR must be explicitly constrained in definite time frame for an easy replication. Some SLRs with search end date open '*to present*' are excluded here.
4. The publication that reports SLR must include the list of identified papers, which may enable a detailed comparison with the results from the replicated search.

In terms of the above criteria, The SLR-1 by Kitchenham et al. [22] that summarizes and reports the impact of SLRs in software engineering was selected as reference for the case study. This SLR performed a manual search in 13 venues with the explicit time span from Jan 2004 through the middle of 2007. As an SLR is a type of secondary study, their work can be regarded as a tertiary study. It retrieved 34 candidate papers, among which 20 SLRs were finally identified as relevant studies and included in the SLR-1.

4.2. Search implementation

4.2.1. Search venues and libraries

At the manual search stage, we chose the venues (journals and conferences) related to Empirical Software Engineering (ESE) and EBSE. By carefully considering the publication venues available in SE community with reference to the rankings [2,1], 9 of them were selected by the first two authors for this study (Table 3). Note that the selected venues for manual search in this case study are different to the original SLR somehow for two reasons: (1) though the replicated search strategy is designated for the same research questions, the authors may have slightly different recognition of the '*related*' venues from the original researchers of the SLR-1; (2) the purpose of the manual search here is to establish the QGS, rather to strive to capture as many relevant papers as possible. Therefore, some originally used venues were ignored at the manual search stage, and two additional venues, EASE and ESEM, were added to the list because of their tight linkages to ESE and EBSE.

The nominated venues can be grouped into five libraries (Table 3), four of which were selected for the automated search due to their popularity in SE research, i.e. IEEE Xplore, ACM DigitalLibrary, ScienceDirect and SpringerLink. Note that other libraries can also be employed for automated search, but given this selection, the QGS is only valid for evaluating the searches through them.

4.2.2. QGS and automated search

In this case study, the target papers should be '*systematic reviews in software engineering*'. Accordingly, we slightly refined the inclusion and exclusion criteria reported in the original SLR-1 [22]. Two authors screened all papers published in the venues from 2004 to 2008 during the manual search independently until reached joint agreements on all included SLRs. In total, 21 published SLRs were retrieved and 20 of them were used for building the QGS. One paper published by BCS was excluded from the QGS since it is impossible to be retrieved by the above selected libraries for the QGS (BCS exclusive), but was included in the final relevant studies. Table 3 shows the venues and their numbers of relevant papers (by 2007 and 2008 respectively).

The case study implemented automated search by following the *subjective* definition approach, in which the search strings were constructed based on the authors' knowledge about EBSE, and their observation of the papers (SLRs) 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`)

Table 3
Selected venues for manual search in CS-1.

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

into the fields of *title-abstract-keywords* through the above libraries. The search strings then were coded to fit the syntax requirements and capability of each search engine.

4.2.3. Evaluation and refinement

Table 4 summarizes the number of studies retrieved by each search engine with the initial and refined search strings. For example, there are 12 relevant studies retrieved by IEEE Xplore, five in the QGS. In total, 13 papers in QGS were retrieved through the initial automated search. In terms of the sample size of QGS, the *quasi-sensitivity* was calculated to be 65% (13/20), which is below the desirable threshold (70–80%) suggested in Section 3.2.5. As required, the search process had to go back to improve the string.

By carefully checking the SLRs included in the 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 literature review* was just introduced in SE. Some authors claimed their review studies using other terms (e.g., '*survey*'). So we refined the string as follows and performed the automated search again.

```
(software AND (systematic OR controlled
OR structured OR exhaustive OR comparative) AND
(review OR survey OR "literature search"))
```

The revised automated search was able to capture 17 studies from the QGS (Table 4), which increased the *quasi-sensitivity* up to 85% (i.e. '*acceptable*'). Fig. 5 illustrates the result compositions of the initial and final search, and the contributions by the manual search (QGS) and the automated search. By combining the papers from manual search, the systematic search process finally identified 38 SLRs for the tertiary study. A full list of the identified SLRs is available in [32].

4.3. Performance comparison

Although the similar inclusion and exclusion criteria were employed in both the original and the replicated searches in this case study, we excluded several '*relevant*' papers that were selected in the original SLR-1 during the manual search and selection due to the deviation caused by how strictly the inclusion/exclusion criteria were followed. For example, [31] was excluded from our study as its random sampling of ICSE papers is not repeatable.

Because of the slight disagreement on identifying SLRs between the original and the replicated searches, it is not appropriate to directly compare the numbers of identified studies from them. Rather we focus on the comparison of performance between the implementations of different search strategies. Table 5 summarizes the numbers of SLR 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 venues reported in [22] from 2004 till 2008. Two more SLRs could be found when screening their specified venues (more than our manual search venues). The row with 'automated only' shows the automated search performance without evaluation and refinement. The bottom row presents the results through the QGS-based systematic search process. Fig. 6 shows a graphic comparison between the results by different search strategies.

Table 4
Results from automated search in CS-1.

	Initial search			Final search		
Search engine	# Retrieved	# QGS	# Relevant	# Retrieved	# QGS	# Relevant
IEEE Xplore	146	5	12	270	8	15
ACM DigitalLibrary	34	1	5	160	1	6
ScienceDirect	31	6	6	82	7	7
SpringerLink	42	1	5	145	1	6
Overall	**253**	**13**	**28**	**657**	**17**	**34**

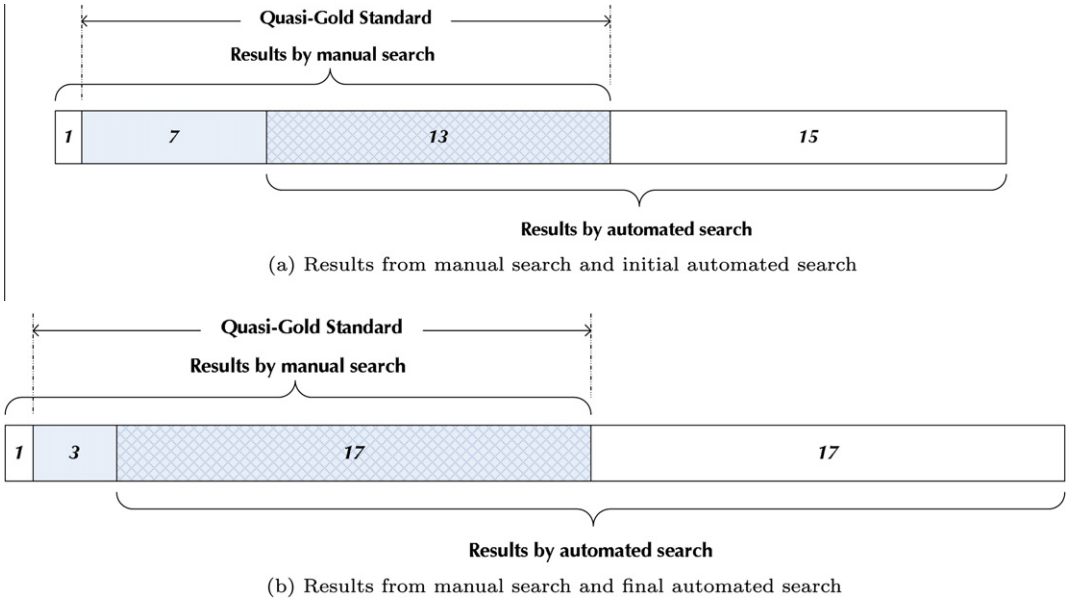

Fig. 5. Compositions of initial and final search results in CS-1 (2004–2008).

Table 5
Compositions of the different search strategies in CS-1.

Search method	SLRs	Quasi-sensitivity
Manual only	22	n/a
Automated only (w/o evaluation)	28	65%
Systematic	38	85%

5. Case study 2

The second case study was aimed to independently follow and evaluate the proposed systematic search process by replicating the literature search of a recently published domain-specific SLR [28]. The case study was mainly done by a PhD student whose research topic is related to global software engineering. The first two authors developed the case study protocol and acted as supervisor and checker in this case. Different from the first case study described in Section 4, the *objective* method to elicit relevant terms and phrases was applied for developing search strings in this case study.

5.1. The original SLR-2

This case study conforms to the SLR selection criteria addressed in Section 4.1. To minimize the potential effect of domain knowledge, a recent SLR-2 on empirical evidence in global software engineering (GSE) [28] was selected for replicating the search process in this case study. The original SLR-2 was designed to answer the following two research questions:

RQ.1 "*What is the state-of-the-art in empirical studies of GSE?*"
RQ.2 "*What is the strength of the empirical evidence reflected in the empirical GSE literature?*"

Another main difference between the original SLR chosen in our case studies is the search methods used. Unlike the SLR-1 that used manual search only, the SLR-2 used automated search only to retrieve the relevant studies. The search strings used in SLR-2 were based on the experience from pilot searches and consisted of a Boolean expression like ((A1 OR A2 OR A3 OR A4) AND (B1 OR B2 OR B3 OR B4)), where

A1:	global software development	B1:	empirical
A2:	global software engineering	B2:	industrial
A3:	distributed software development	B3:	experiment
A4:	distributed software engineering	B4:	case study

The original search started in November 2007. Full-text field was searched using the above query through seven digital libraries including Compendex, IEEE Xplore, SpringerLink, ISI Web of Knowledge, ScienceDirect, Wiley InterScience, and ACM DigitalLibrary. The authors of the original SLR-2 [28] intentionally decided to exclude the papers published before 2000 from their search. A list of 59 relevant studies is included in the appendix of the SLR-2 [28].

5.2. Search implementation

The student was given the SLR guidelines [24], the QGS-based search process [33], and the original SLR-2 [28] as the learning material prior to implementing the search step. The student was allowed to consult other published SLRs in SE as examples. The first two authors were responsible for answering his questions related to SLR methodology and the systematic search process. After learning the SLR methodology, the student developed the search protocol for replicating the search process of the original SLR by following the systematic search process presented in this paper.

5.2.1. Search venues and libraries

In order to establish the QGS, the student under the supervision chose the venues (high quality conferences and journals in terms of the rankings [1,2]) for manual search that are related to *empirical software engineering*, *global software engineering*, and *generic software engineering* (summarized in Table 6). In terms of their publishers, these venues can be grouped into 5 libraries: IEEE, ACM, Elsevier (ScienceDirect), Springer, and Wiley.

5.2.2. QGS and automated search

By following the study inclusion/exclusion criteria reported in [28], the manual search of the literature venues in Table 6 identified 52 relevant papers published between 2000 and 2007. These 52 papers formed the QGS for this case study.

In order to investigate the alternative approaches to defining search string (Section 3.2.3), this case study applied the *objective* method of eliciting search terms, in which the high-frequency terms and their relations are revealed by textual analysis tool based on the sample data (e.g., titles, abstracts, and keywords) of the papers from the QGS.

The text analysis tool used in this case study was WordStat 6.1 [3] that is able to provide a graphic representation of the frequently occurring terms and their relations. In this case, the student chose the *term frequency* (TF) and *inverse document frequency* (IDF) from the algorithms offered by WordStat. The statistical analysis, *Jaccard's similarity coefficient* used in this case, enables a researcher to determine the importance of a term or phrase to a collection of documents by comparing the similarity and diversity of the sample sets.

After removing the irrelevant words, the terms with a frequency factor of 30 or higher remained as the candidate terms for constructing the search strings. Fig. 7 shows these terms in two dendograms: the top one (a) assesses the importance of the original terms (from SLR-2) in the QGS; the bottom one (b) was obtained by crossing out the words '*software*' and '*development*' in order to have a better visible representation, in terms of recurrence, of

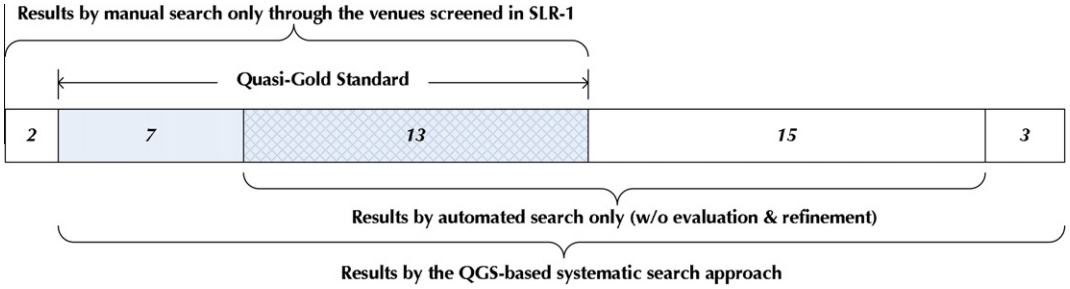

Fig. 6. Comparison and composition of the different search strategies in CS-1.

Table 6
Selected venues for manual search in CS-2.

Venue	Library/publisher	Papers (2000–2007)
ICGSE	IEEE	13
SPIP	Wiley	11
CSCW	ACM	7
IST	Elsevier	6
ICSE	IEEE/ACM	5
ISESE/ESEM	IEEE/ACM	3
TSE	IEEE	2
JSS	Elsevier	2
FSE	ACM	1
EMSE	Springer	1
ICSP	Springer	1
Total		52

the terms and phrases frequency in the QGS. Compared to Fig. 7a, the terms relevant to the research topic (such as '*distributed*' and '*collaboration*') became more visible in Fig. 7b.

5.2.3. Evaluation and refinement

The student first tried to adopt the original search string used in the SLR-2, and retrieved 20 papers in QGS. Divided by the sample size of the QGS (52), the quasi-sensitivity of the original query is 40% (21/52), which is much lower than the recommended threshold (e.g., 70–80%) in Section 3.2.5.

Although the frequency analysis of the QGS confirmed some terms employed in the original search of the SLR-2, like '*software development*', '*distributed*', and '*global*' for the first part of the original string, the first dendogram (Fig. 7a) shows some original terms are not strongly connected with each other. For instance, '*software engineering*' seldom appeared as a phrase in the analysis of the QGS, and '*global*' was not closely connected with '*software development*'.

For the second part of the original search string, the terms (i.e. '*empirical*', '*industrial*', '*experiment*', and '*case study*') are not even present. Although the dendogram only shows the combinations of the terms with at least a frequency factor of 30 and was not derived from the entire *gold standard*, we do believe that 11 relevant venues provide a good sample set representing the use of keywords in the papers related to a given research topic.

According to the comparison of the terms and phrases used in SLR-2 as well as identified by text frequency analysis (Table 7), the student decided to lower the restriction of the first part of the original string, as well as to introduce a few more interesting terms to the second part based on the observation and analysis of the QGS. The string for automated search became:

```
("software engineering" OR "software development"
OR "distributed team") AND (global OR
distributed) AND (empirical OR "case study"
OR experiment OR industrial OR interview)
```

By isolating the words '*global*' and '*distributed*' from '*software development*' and '*software engineering*' and including more featuring words denoting *empirical studies*, seven more papers in QGS were retrieved by the refined automated search. Further, considering the capabilities of different search engines in dealing with

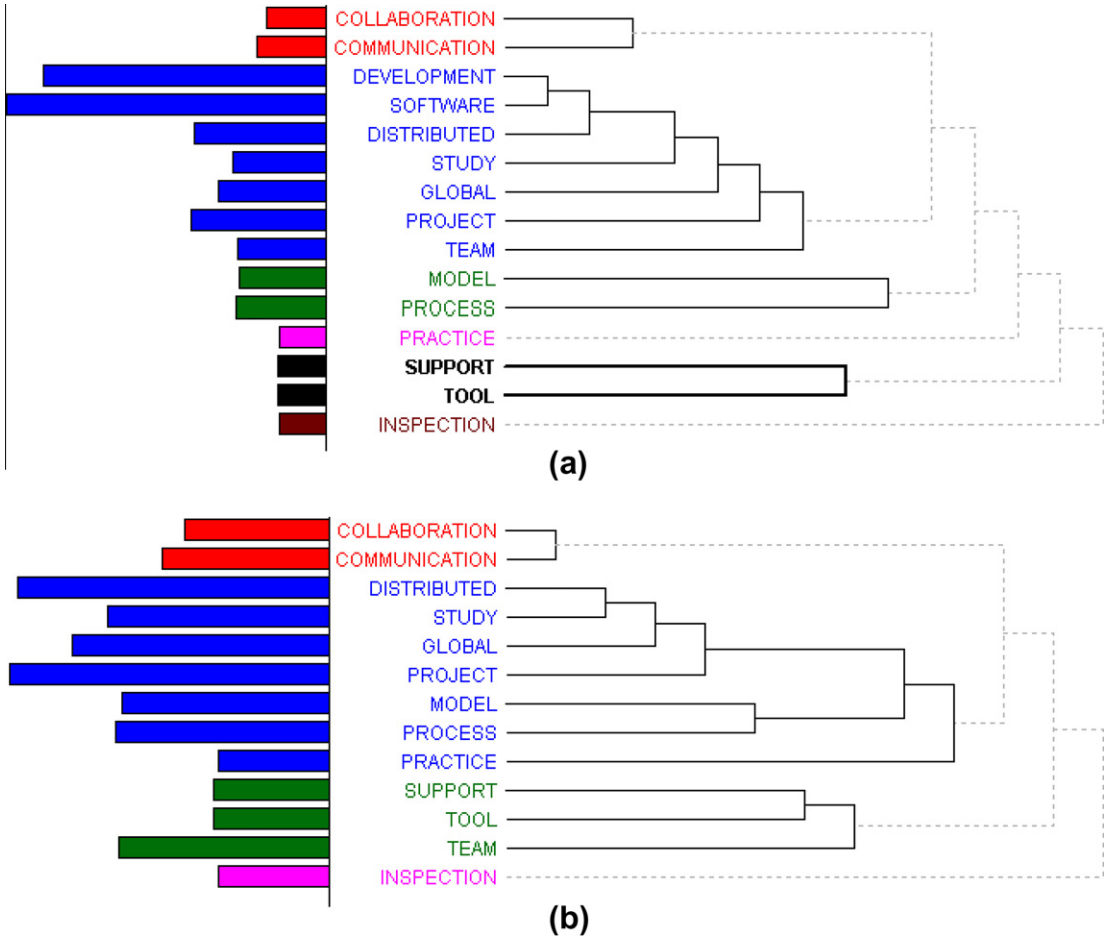

Fig. 7. Terms and phrases with high-frequency.

stemming (e.g., *experimental*, *experiments*, and *experimentation* for '*experiment*'), the string was extended as follows:

```
("software engineering'' OR ''software development"
OR ''distributed team") AND (global OR
distributed) AND (empirical OR ''case study'' OR
"case studies'' OR lesson OR lessons OR experience
OR experiment OR experiments OR experimentation
OR experimental OR experimenting OR industrial
OR interview OR interviews OR survey OR
surveys OR ''test case'' OR ''test cases")
```

The search on the digital libraries by the above query was able to retrieve 38 papers belonging to the QGS through the search engines. Afterward, the student further strived to refine the string with the intention of retrieving more papers in QGS. However, the improvement was very slim. On the contrary, the number of irrelevant papers increased dramatically. Therefore, the student decided to apply the above mentioned search string for the final automated search. Correspondingly, the final quasi-sensitivity increased to 73% (38/52), which results in a fairly acceptable recall with reference to the suggested threshold range (70–80%) in Section 3.2.5.

Finally, it has to be noticed that the new query string is a superset of the original one; hence all the publications retrieved by the original search string in the SLR-2 [28] are also retrieved by the new one.

Table 7
Key phrases and their quasi-sensitivities.

Phrase used in **SLR-2**	Hits	%	Phrase identified from **QGS**	Hits	%
"Global software development"	23	44	"Software development"	39	75
"Distributed software development"	13	25	"Software engineering"	10	19
"Global software engineering"	0	0	Collaboration AND communication	16	30
"Distributed software engineering"	1	2	"Distributed team"	6	11
Empirical	16	30	Practice AND research	6	11
Industrial	1	2	"Tool support"	2	4
Experiment	7	13	"Process model"	3	6
"Case study"	16	30	"Empirical study"	8	15

Table 8
Comparing the original with replicated searches in CS-2.

	Relevant papers	Quasi-sensitivity (%)
Original search	59	40
Replicated (systematic) search	150	73

5.3. Performance comparison

Table 8 summarizes the overall search results and performance evaluation of both search strategies. On one side, the original SLR-2 applied automated search only and identified 59 relevant papers (40% of the papers in the QGS). On the other side, the new search string, developed and refined based on the analysis of the terms used in the QGS, identified 73% of the papers in the QGS resulting in 83% performance increase (i.e. from 40% to 73%) in terms of the number of retrieved papers found in the QGS (quasi-sensitivity).

Table 9, which synthesizes some of the search phase data, shows that after the screening of 2404 papers 150 publications were selected as relevant. Therefore, considering the two phases, the total number of papers selected is 164: 52 papers coming from the manual search and 150 from the automated one (38 overlapping, illustrated in Fig. 8). As a subset of the results by the replicated search in this case study, the original search of the SLR-2 could find only 36% (59/164) of the relevant papers identified through the systematic search process presented in this paper.

Note that after restructuring and refining the search query, the search with the new string is able to capture all the papers retrieved from the QGS by the original string of the SLR-2. It is also noticeable that CSCW (ACM Conference on Computer Support Cooperative Work) contributes a significant number of relevant studies (7%) to the QGS, however no relevant paper from this venue was reported in the search of the original SLR-2 [28].

Table 9
Results from automated searches in CS-2.

Search engine	# **QGS**	Original search			Final search		
		# Retrieved	# in QGS	# Relevant	# Retrieved	# in QGS	# Relevant
IEEE Xplore	23	–	10	23	231	20	75
ACM DigitalLibrary	8	–	0	7	92	2	19
ScienceDirect	8	–	3	9	63	5	11
SpringerLink	2	–	0	10	140	1	20
Wiley InterScience	11	–	8	5	399	10	14
Web of knowledge and compendex		–	n/a	5	1479	n/a	11
Overall	**52**	387	**21**	**59**	2404	**38**	**150**

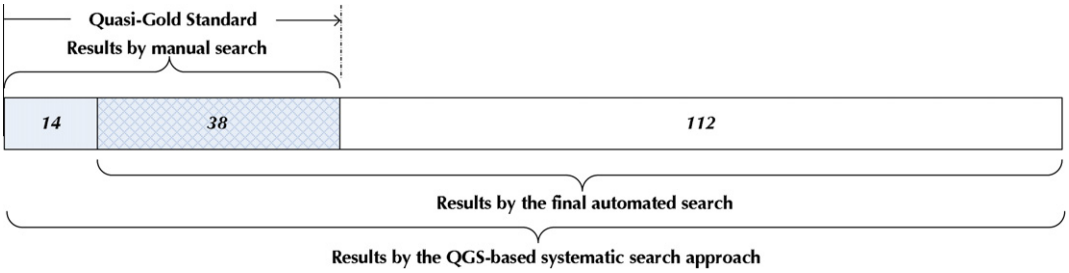

Fig. 8. Composition of the systematic search results in CS-2 (2000–2007).

6. Discussion

The research reported in this paper was motivated by an important need for improving the search process of conducting SLRs. Like many other practitioners of EBSE [23], the case studies reported in this paper have also illustrated the limitations of applying automated or manual search methods alone. A manual search can consume a huge amount of effort when scanning a large number of literature venues. On the other hand, the performance of automated search is highly dependent upon the quality of search strings used. It is quite common that researchers have to continuously refine the search strings through several trials with different search engines. While some of the previously reported SLRs have applied both automated and manual searches, almost all of them simply merged the search results achieved from both kinds of search methods. Our first case study has demonstrated that neither the automated search with intuitive search terms nor the manual search in limited number of literature venues could retrieve an "*acceptable*" number of relevant papers alone.

In contrast, the QGS-based systematic search approach aims to achieve the tradeoff between them while targeting an acceptable performance (in terms of sensitivity in the reported case studies). It not only combines the results from the two search methods together, but establishes linkage between them for leveraging the strengthens of both kinds of search methods (i.e. automated and manual). Our systematic search approach also provides a mechanism of quantitatively determining when a researcher can stop the iterative refinement of search strings for automatic search method. We also assert that the presented approach can help capture considerable number of studies identified in QGS with a reasonable amount of effort.

The performance improvement of the systematic search process is confirmed with different search settings used in the two case studies. The SLR-1 applied manual search only, while the SLR-2 used automated search. The improvements to the original searches were observed in both cases when adopting the systematic search process. In particular, a large number of new relevant papers were identified to complement the papers included in the original SLR-2. Although many more papers retrieved by the search engines had to be screened in the second case, we believe a more comprehensive collection of the relevant papers for SLR makes the required extra effort worthwhile. Additionally, the two reported case studies demonstrate the benefits and limitations of the *subjective* and *objective* methods for identifying the relevant terms and phrases, and developing search strings.

If a researcher can find a set of secondary studies, which may have been identified and screened by some other researcher during a review of a related topic, those studies can be added to the QGS in order to reduce the effort required for the manual search during the development of QGS. For instance, a few previous SLRs [15,18,20,21] directly used studies identified by [26] as their full set of primary studies. As another example, the results from the mapping study [19] can be directly used to build QGS for more specific SLRs on software cost estimation. In such cases, the results may need to be tailored in terms of *subject* and *time* that conform to the scope of the new SLR.

As the aim of an SLR is to find as many primary studies relevant to the research questions as possible [24], '*sensitivity*' is usually the top priority in defining search strategies in most SLRs. Accordingly, we have focused on this metric when describing the case studies. Though another metric '*precision*' has received relatively less attention in the initial set of evaluation case studies reported in this paper, we consider that "*precision*" is also important to measure the productivity of a search strategy. We are undertaking a series of case studies which would consider both metrics to further assess the presented systematic search process.

Based on our observation of the existing SLRs in SE [32], automated search often 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. [10] investigated the possible influence of the quality of abstract to SLRs by experiments, and concluded that *structured abstracts* can improve the likelihood of understanding and identification of studies. We assert that structured abstracts are also likely to further improve the search accuracy.

7. Threats to validity

Aiming to demonstrate and evaluate the different implementations of search string development of the systematic search approach, we conducted two case studies on different topics instead of a single case study. However, the cases were not randomly selected due to the considerations (criteria) discussed in Section 4.1 as well as the resource available to us when planning them.

Due to the focus of this paper, only the study search and selection steps of the original SLRs were replicated in the case studies. According to our observation of existing SLRs in SE, we also notice that for some SLRs the study exclusion may still take place in *data extraction* activity because normally the full-text of paper is not checked in the prior study screening, which means the *sensitivity* and *precision* might be also related to other activities in an SLR. However, we believe the portion of the studies excluded in *data extraction* is much smaller compared to *search* and *selection* steps. In practice, these residual irrelevant studies are difficult to remove no matter whether or not a systematic search process is applied. Hence, their influence on the findings of the case studies is quite limited.

Our two case studies observed the literature search processes being performed by both experienced researchers and research student, which we believe are typical practitioners of SLR methodology in software engineering [4].

The CS-2 was mainly performed by a research student. Although he was able to consult with the experienced researchers, the most final decisions were made by himself. The student was provided with the original SLR-2 as background material at the beginning of the CS-2. We notice that it may influence the student's performance of the search process. However, even if the student had been intentionally blinded from the SLR-2 (as we originally planned), because the SLR-2 had already been published, he would have been able to find and read the original study in terms of the information like research questions. Hence we decided to avoid such uncertainty in the case study. On the other hand, the original SLR-2 has been published by a quality journal, which implies the quality of their review (including their literature search) is acceptable in the research community. Accordingly, we believe that the comparison of the both searches is justifiable.

8. Related work

Systematic literature reviewers in software engineering are aware of the importance of literature search, as well as the challenges involved in searching relevant studies when applying SLR methodology in different sub-disciplines of software engineering and computer science. Many SE researches have reported various kinds of difficulties during the "*searching relevant studies*" step of SLR methodology [4]. Several experienced systematic literature reviewers have also discussed the issues related to literature search in SE. In the following paragraphs, we brief the work of

other researchers who have tried to propose different approaches to improving the literature search step of SLR in SE.

Brereton et al. identified several issues of electronic (i.e. automated) search derived from their experience in conducting SLRs [9]. Based on their extensive experience with SLR methodology, they conclude that researchers must select and justify a search strategy that is appropriate for their research questions. They also advised against retrieving the primary studies from a single venue. Bailey et al. [6] analyzed the overlapping phenomenon between results returned from different search engines in SE.

Dieste et al. [13,12] investigated the optimal search strategies using the combination of alternative search strings for automated search in SLR. They used the studies identified in another SLR [26] to establish the 'gold standard' for calculating the search sensitivity. However, it is almost impossible for most of the researchers in SE to have access to such pre-made 'gold standard' during the planning stage of their intended SLRs. In other words, a '*gold standard*' in this case provides no help to search strategy evaluation, or to ensure the retrieval quality of relevant studies in SLRs.

Kitchenham et al. reported an participant-observer case study [23] that investigated several impact factors for the literature search in SLRs: search breadth, gray literature, and performance of manual and automated search. The results support the idea that a restricted manual search targeting an appropriate set of venues may help to avoid the omission of good quality papers. Despite the fact that most systematic reviewers appeared to prefer automated search method in their SLRs (68% reported SLRs by 2008 using automated search), the finding reported by Kitchenham and her colleagues confirms the value of manual search for SLRs in SE.

As an alternative to search engine based search strategy, a reference list driven search method can be another option for retrieving relevant studies. This method was innovated with the concepts of co-citation and bibliographic coupling, and has been proposed in SE [27]. However, it is well-known fact that most of the major digital libraries in SE are not designed for supporting this kind of search. Accordingly, it can be very time-consuming in terms of manually retrieving studies from reference lists. Thus this search method appears to be not practical enough unless being supported by major digital libraries in SE. As suggested by the SLR guidelines [24], this method can be used as a supplementary venue for a full SLR.

Although the above mentioned studies discuss a number of aspects of literature search for SLRs and propose a few potential approaches, we are not aware of any instructive and practical literature search approach, which is able to provide systematic and rigorous process for integrating different search methods, defining the relevant search terms and phrases, and evaluating search queries to improve search performance for SLRs in SE. The presented case studies provide primary evidence to support the practicability of the proposed systematic search approach.

9. Conclusions

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

- Describing an explicit scope of search strategy and its evaluation in searching relevant studies in SE.
- Introducing the concepts of '*quasi-gold standard*' and '*quasi-sensitivity*' for developing and evaluating search strategy for a given SLR.
- Proposing a systematic, evidence-based, and rigorous approach for practical search strategy development, implementation and evaluation.
- Conducting and reporting two case studies that replicated the literature searches of existing SLRs by following the QGS-based systematic search process but applying different search string development methods.

Literature search is the first and critical step of any forms of literature reviews. Although the QGS-based literature search approach is proposed for improving the performance of search processes in SLRs and EBSE, it is also useful in other forms of literature reviews in SE, and benefits the researchers and practitioners who intend to retrieve a relatively comprehensive collection of relevant studies (to a given *subject matter* and *time span*) with reasonable effort.

Currently this approach is being actively and effectively applied in some systematic reviews in SE, such as [34,5]. We will continue the evaluation and improvement of this approach by conducting more case studies with the research interests in both *sensitivity* and *precision* on varying topics in software engineering. In addition, the future methodological work in empirical software engineering community may identify other issues and limitations of the existing SLRs reported in software engineering, and further to suggest practical improvements and enhancements to the guidelines of systematic literature reviews.

Acknowledgments

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

References

[1] The excellence in research for Australia ranked conference list. Australian Research Council, 2007 & 2010. <http://www.arc.gov.au/era/default.htm>.
[2] The excellence in research for australia ranked journal list. Australian Research Council, 2007 & 2010. <http://www.arc.gov.au/era/default.htm>.
[3] Simstat v.2.5 and wordstat v.6.1. Provalia Research, May 2010. <http://www.provalisresearch.com/>.
[4] M. AliBabar, H. Zhang, Systematic literature reviews in software engineering: preliminary results from interviews with researchers, in: Proceedings of the 3rd International Symposium on Empirical Software Engineering and Measurement (ESEM'09), Lask Buena Vista, FL, October 2009. IEEE Computer Society, pp. 346–355.
[5] X. Bai, L. Huang, H. Zhang. On scoping stakeholders and artifacts in software process, in: Proceedings of International Conference on Software Process 2010 (ICSP 2010), vol. LNCS 6195, Paderborn, Germany, July 2010, Spinger-Verlag, pp. 39–51.
[6] J. Bailey, C. Zhang, D. Budgen, M. Turner, S. Charters, Search engine overlaps: Do they agree or disagree? in: Proceedings of 2nd International Workshop on Realising Evidence-Based Software Engineering (REBSE'07), Minneapolis, MN, USA, May 2007, IEEE Computer Society.
[7] J. Biolchini, P.G. Mian, A.C.C. Natali, G.H. Travassos. Systematic Review in Software Engineering, Technical Report, Universidade Federal do Rio de Janeiro, 2005.
[8] J. Boynton, J. Glanville, D. McDaid, C. Lefebvre, Identifying systematic reviews in medline: developing an objective approach to search strategy design, Journal of Information Science 24 (3) (1998) 137–154.
[9] P. Brereton, B.A. Kitchenham, D. Budgen, M. Turner, M. Khalil, Lessons from applying the systematic literature review process within the software engineering domain, Journal of Systems and Software 80 (1) (2007) 571–583.
[10] D. Budgen, B.A. Kitchenham, S.M. Charters, M. Turner, P. Brereton, S.G. Linkman, Presenting software engineering results using structured abstracts: a randomised experiment, Empirical Software Engineering 13 (4) (2008) 435–468.

[11] K. Dickersin, R. Scherer, C. Lefebvre, Systematic reviews: identifying relevant studies for systematic reviews, British Medical Journal 309 (6964) (1994) 1286–1291.
[12] O. Dieste, A. Griman, N. Juristo, Developing search strategies for detecting relevant experiments, Empirical Software Engineering 14 (5) (2009) 513–539.
[13] O. Dieste, A.G. Padua. 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 2007, IEEE Computer Society, pp. 215–224.
[14] T. Dyba, T. Dingsoyr, G.K. Hanssen. 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 2007. IEEE Computer Society, pp. 225–234.
[15] T. Dyba, V.B. Kampenes, D.I. Sjoberg, A systematic review of statistical power in software engineering experiments, Information and Software Technology 48 (8) (2006) 745–755.
[16] T. Dyba, B. Kitchenham, M. Jorgensen, Evidence-based software engineering for practitioners, IEEE Software 22 (1) (2005) 158–165.
[17] S. Grimstad, M. Jorgensen, K. Molokken-Ostvold, Software effort estimation terminology: the tower of babel, Information and Software Technology 48 (4) (2006) 302–310.
[18] J.E. Hannay, D.I. Sjoberg, T. Dyba, A systematic review of theory use in software engineering experiments, IEEE Transactions on Software Engineering 33 (2) (2007) 87–107.
[19] M. Jorgensen, M. Shepperd, A systematic review of software development cost estimation studies, IEEE Transactions on Software Engineering 33 (1) (2007) 33–53.
[20] V.B. Kampenes, T. Dyba, J.E. Hannay, D.I. Sjoberg, A systematic review of effect size in software engineering experiments, Information and Software Technology 49 (11–12) (2007) 1073–1086.
[21] V.B. Kampenes, T. Dyba, J.E. Hannay, D.I.K. Sjoerg, A systematic review of quasi-experiments in software engineering, Information and Software Technology 51 (1) (2009) 71–82.
[22] B. Kitchenham, O.P. Brereton, D. Budgen, M. Turner, J. Bailey, S. Linkman, Systematic literature reviews in software engineering: a systematic literature review, Information and Software Technology 51 (1) (2009) 7–15.
[23] B. Kitchenham, P. Brereton, M. Turner, M. Niazi, S. Linkman, R. Pretorius, D. Budgen. The impact of limited search procedures for systematic literature reviews – a participant–observer case study, in: Proceedings of the 3rd International Symposium on Empirical Software Engineering and Measurement (ESEM'09), Lask Buena Vista, FL, October 2009. IEEE Computer Society, pp. 336–345.
[24] B. Kitchenham, S. Charters, Guidelines for Performing Systematic Literature Reviews in Software Engineering (version 2.3), Technical Report, Keele University and University of Durham, 2007.
[25] B. Kitchenham, T. Dyba, M. Jorgensen, Evidence-based software engineering, in: Proceedings of 26th International Conference on Software Engineering (ICSE'04), Edinburgh, Scotland, UK, May 2004. IEEE Computer Society, pp. 273–284.
[26] D.I. Sjoberg, J.E. Hannay, O. Hansen, V.B. Kampenes, A. Karahasanovic, N.-K. Liborg, A.C. Rekdal, A survey of controlled experiments in software engineering, IEEE Transactions on Software Engineering 31 (9) (2005) 733–753.
[27] M. Skoglund, P. Runeson. Reference-based search strategies in systematic reviews, in: Proceedings of 13th International Conference on Evaluation and Assessment in Software Engineering (EASE'09), Durham, England, April 2009. BCS.
[28] D. Smite, C. Wohlin, T. Gorschek, R. Feldt, Empirical evidence in global software engineering: a systematic review, Empirical Software Engineering 15 (1) (2010) 91–118.
[29] V. White, J. Glanville, C. Lefebvre, T. Sheldon, A statistical approach to designing search filters to find systematic reviews: objectivity enhances accuracy, Journal of Information Science 27 (6) (2001) 357–370.
[30] R.K. Yin, Case Study Research: Design and Methods, fourth ed. Sage Publication, 2009.
[31] C. Zannier, G. Melnik, F. Maurer. On the success of empirical studies in the international conference on software engineering, in: Proceedings of 28th International Conference on Software Engineering (ICSE'06), Shanghai, China, May 2006. ACM, pp. 341–350.
[32] H. Zhang, M. AliBabar, Adopting Systematic Reviews in Software Engineering: An Evidence-based Report. Technical report, Lero Software Engineering Research Centre, 2009.
[33] H. Zhang, M. AliBabar, On searching relevant studies in software engineering, in: Proceedings of 14th International Conference on Evaluation and Assessment in Software Engineering (EASE'10), Keele, England, April 2010. BCS.
[34] H. Zhang, B.A. Kitchenham, D. Pfahl, Software process simulation modeling: an extended systematic review, in: Proceedings of International Conference on Software Process 2010 (ICSP 2010), vol. LNCS 6195, Paderborn, Germany, July 2010. Springer-Verlag, pp. 309–320.