Reference-based search strategies in systematic reviews

Runeson, Per; Skoglund, Mats

Published in:
[Host publication title missing]

2009

Reference-based search strategies in systematic reviews

Mats Skoglund and Per Runeson
Department of Computer Science, Lund University, Box 118, SE-221 00 Lund, Sweden
{mats.skoglund, per.runeson}@cs.lth.se

Abstract

In systematic reviews, the number of articles found by search strings tend to be very large. In order to limit the number of articles to handle manually, we investigate a search strategy based on references between papers. We first identify a "take-off paper" which is the starting point for the search and then we follow the references from that paper. We also investigate "cardinal papers", i.e. papers that are referenced by many authors, and let the references to those papers guide the selection in the systematic review. We evaluate the search strategies on three published systematic reviews. The results vary greatly between the three studied systematic reviews, from 88% reduction to 92% extension of the original paper set.

1. INTRODUCTION

Systematic reviews have gained popularity recently within software engineering. Thanks to Kitchenham's *et al.* seminal work on adapting general guidelines to software engineering (EBSE-2007-01), results from scattered empirical studies have been gathered into coherent knowledge. Identification of primary studies to include in systematic reviews should be a highly structured process where as much as possible of relevant research is retrieved. This process includes the development of a search strategy, where a retrieval method is defined. The method is typically based on running search strings in electronic publication databases, such as ACM Digital library, IEEEXplore, ISI Web of knowledge or Google scholar.

An example of the use of search strings is in the systematic review conducted by Dybå *et al*. (Dybå2007) reviewing studies on agile software development. They searched eight different electronic sources using search strings such as:<agile> AND <software>, <extreme programming>. The search strings are typically based on the research question(s) for the systematic review and may include different aspects of the field of interest such as context, type of research, population, etc.

A search strategy based on search strings usually results in a large set of primary study candidates, often thousands (Engström2008), of various quality that must be "cleaned" or "filtered" manually. For example, Dybå *et al.* (Dybå2007) filtered their initial search result set on publication types by excluding for example editorials, prefaces, summaries of tutorials. The initial result set is then further cleaned by excluding papers based on their titles, keywords, contents in abstracts and by studying the full text to get a final search result set containing the primary studies to analyze in the systematic review. The cleaning process is often a time consuming process, leading to a large proportion of the initial search result set being discarded. It is also observed that electronic publication databases are not designed to support systematic reviews (Brereton2007, Dybå2007).

Attempts have been made to improve the search for primary studies. For example, Dieste *et al*. investigated how different terminology may be used in the search strings in systematic reviews (Dieste2008).

In this paper, we propose and evaluate a search strategy for primary studies in systematic reviews, based on citation analysis. In traditional literature search, references are used to find relevant papers. Our hypothesis is that the semantic information in references between papers can be used to more efficiently identify the set of primary studies for the systematic review. Specifically, we expect this to be true for systematic reviews on precise and focused research questions.

In Section 2, we define evaluation criteria for search strategies. Section 3 defines the proposed reference-based search strategy and Section 4 presents the method for evaluation of the proposed approach on three cases of previously published reviews. Section 5 presents in detail the application of the referenced-based approach to the three studies and Section 6 analyses the results. In Section 7 we extended the proposed method and evaluated the outcome. Finally, Section 8 concludes the paper.

2. EVALUATION OF SEARCH STRATEGIES

To be able to conduct a good systematic review, the search strategy should give an initial search result set containing a large portion of the relevant primary studies in the field. This goal may be approached by using a search strategy including as much as possible of all existing research. Unfortunately, this approach also includes a

large number of irrelevant materials in the result set which must be excluded in the manual cleaning later. Another approach may instead be to perform a very narrow search to only include relevant studies to try to reduce the time that must be devoted to cleaning, an approach that as a consequence may miss important research. A good search strategy should include much relevant research but also exclude much of the irrelevant research.
The measure of the degree of included relevant research in a search result set is sometimes referred to as *recall* (Rijsberger1979). Recall is defined as the proportion of retrieved relevant material among all available relevant material, expressed as a percentage number. An optimal search strategy has a recall of 100%, meaning that all available relevant material is included in the result set.
The measure of a search strategy's ability to exclude irrelevant material may be expressed as *precision* (Rijsberger1979), which is defined as the proportion of relevant material among the retrieved material. This may also be expressed as a percentage number, where 100% precision means that all of the retrieved material is relevant.
The optimal search strategy has 100% recall and 100% precision. However, it is unlikely that a search strategy gives 100% recall and/or 100% precision. Thus, we must in most cases satisfy with a "good enough" search strategy, resulting in not too much relevant material is missed and that a manageable volume of irrelevant material is included. Different search strings may result in different recall and precision. A search strategy using search strings may be optimized by for example adding search terms or combining several search strings with the goal of increasing recall, precision or both.
Dieste, Grimán and Juristo have studied the optimality of search strategies using search strings for systematic reviews in software engineering (Dieste2008). In this study a number of search strings were executed and evaluated in terms of recall and precision. To calculate recall it is required to know how many relevant papers there is in a research field. But since this is impossible to know, Dieste *et al*. instead calculated recall using a *gold standard* (or test collection (Bailey2003)). The gold standard used was the result from the systematic review conducted by Sjøberg *et al*. (Sjøberg2005) which Dieste *et al*. compared their search result set with to compute recall.

3. REFERENCE LISTS AS SEARCH STRATEGY

In this study, a search strategy for systematic reviews, based on reference lists in papers, has been developed and evaluated. Using reference lists to identify relevant papers is not a new idea, it has been used in information science for several years, e.g. through concepts such as co-citation and bibliographic coupling. Co-citation is based on the assumption that two papers are related if they are cited together in other papers. The concept has been used to measure relationships between authors, documents and journals in order to describe the mainstream in a field (Tsay2003). Bibliographic coupling was presented by Kessler as a measure on affinity between papers (Kessler1963). Bibliographic coupling is based on the assumption that two documents that cite a common source are related. However, as pointed out by Martyn, two documents may not reference the same piece of information in the source but bibliographic coupling may still serve as an indication of a relation between papers (Martyn1964).

3.1. Reference lists in systematic reviews

Using reference lists to retrieve material for systematic reviews is suggested in Kitchenham's *et al.* guideline for performing systematic reviews in software engineering (EBSE-2007-01). The guideline states that it is not sufficient to solely use search strings to hunt for relevant material but it must also be complemented by scanning reference lists. Reference lists has also been used practically in systematic reviews as a part of a search strategy, see e.g. (Jørgensen2007). But as far as we know, reference lists has not been used and evaluated as the main search strategy in systematic reviews in software engineering before.

3.2. A reference list search strategy

The reference list search strategy evaluated in this work is based on four components, which are defined below: (1) A take-off paper, (2) papers referenced by the take-off paper in two steps, (3) identification of cardinal papers, (4) papers from external sources referencing cardinal papers.
We define the *initial search result set* to refer to the set of primary studies that is the result of applying a search strategy, before filtering. This initial search result set corresponds to the set of papers retrieved by e.g. running a search string on electronic sources using a search string based search strategy or by applying our reference-based search strategy. The initial search result set may then be further filtered using e.g. keyword or abstract filtering or manual filtering based on the papers full text, to get a *final search result set* to analyze in the systematic review. The remaining of this section describes the reference list search strategy evaluated in this work.

3.2.1. Take-off paper

The starting point for using the reference list search strategy is a paper on the topic that is to be systematically reviewed. The paper, called the *take-off paper*, should be considered relevant to such a degree that it will be

included in the final search result set later. Researchers conducting a systematic review should easily be able to select such a relevant paper that can serve as a take-off paper, based on their pre-understanding of the research question. The selected take-off paper is added to the initial search result set (and will by definition also be included in the final search result set).

3.2.2. Papers referenced from the take-off paper

Papers referenced from the take-off paper directly and indirectly are added to the initial search result set in two steps, step 1 and step 2.

Step 1: All references listed in the take-off paper's reference list are assumed to be relevant to the systematic review subject and are added to the initial search result set.

Step 2: For each of the references in the initial search result set after performing step 1, except for the take-off paper, the same procedure is repeated to add to the initial search result set all papers referenced from the papers referenced from the take-off paper. Some filtering may be performed in this step to collect references only from papers from step 1 that are from high quality sources, or to exclude reference lists in books, manuals, tutorials, etc, see Figure 1.

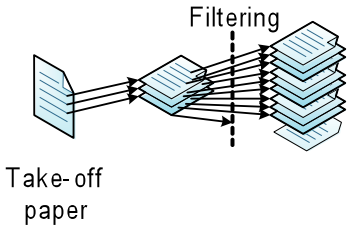

FIGURE 1. The take-off paper and the references from the take-off paper in two steps are added to the search result set.

Papers already in the initial search result set are not added again, thus there are no duplicates. However, to be able to identify cardinal papers later (see below), the number of times each paper is referenced by another paper is recorded. The search result set may be represented by a simple table with information about the papers and reference counts, see an example in Table 1.

Id	Paper	Reference count
1	<The title of the take-off paper>, …	1
…	…	…
n-1	<The title of another paper>, …	8
n	<The title of the last paper>…	3

TABLE 1. An example of a representation of the initial search result set using a table containing paper Id, data about each paper and the number of times each paper is referenced by other papers.

3.2.3. Identification of cardinal papers

Some papers are referenced more than others and we assume those papers are more likely to be referenced also from the relevant papers. We call these particularly referenced papers for *cardinal papers*. In the initial search result set example represented by Table 1, the paper with Id=n-1 is the topmost cardinal paper and is used to collect more potentially relevant papers. In this study we also use variants of this definition of cardinal papers to further explore the reference list strategy in various directions.

3.2.4. Papers in external sources referencing cardinal papers

There may exist papers that are not themselves in the initial search result set, but refer to papers that are in the result set. The last component of the reference list search strategy is to add to the initial search result set papers from external sources, which refer to the cardinal papers. These external papers may be identified by the aid of features available in some digital libraries such as in ISI Web of Science or Google scholar where papers that refer to a specific paper may be listed.

To summarize, the initial search result set contains: the take-off paper, the papers referenced from the take-off paper, the papers referenced from the papers referenced from the take-off paper and papers referencing the cardinal papers, se Figure 2. The initial search result set may then be further filtered on e.g. keywords, abstract, full text etc. to get a final search result set of studies to analyze in the systematic review.

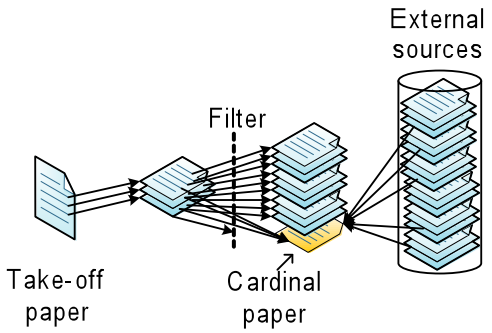

FIGURE 2. The complete initial search result set consists of the take-off paper, papers referenced from the take-off paper in two steps and papers referencing cardinal papers from external sources.

4. EVALUATION OF THE REFERENCE LIST SEARCH STRATEGY

We have studied the effectiveness of the reference list search strategy described above by applying it in to three cases of previously published systematic reviews. The results of the reference list strategy were then evaluated by relating the results to the results of the original search strategies used in the systematic reviews, see Sections 5 and 6. The reference list strategy was also further explored in various directions with the goal to further optimize the strategy, see Section 7.

4.1 Systematic reviews used as evaluation cases

The three systematic reviews used as evaluation cases are referred to as *Engström study*, *Dybå study* and *Sjøberg study,* respectively, based on the first authors of the papers reporting each of the three systematic reviews. The reviews are selected to represent different types of systematic reviews.

4.1.1. Engström study

In this systematic review, empirical evaluations of specific techniques, namely regression test selection techniques, were studied (Engström2008). The search strategy was search string based and the search strings were applied on seven different electronic sources. Technical reports, workshop reports and work in progress were excluded. The reported initial search result set contained 2 923 papers. The initial search result set was filtered on titles, abstracts and full text to reach a final search result set of 28 relevant papers included in the systematic review analysis. Thus the precision for the original search strategy can be calculated to 28/2923x100 = 0.96%.

4.1.2. Dybå study

Empirical studies of agile software development were studied (Dybå2008), i.e. a development method. The search strategy was search string based and the search terms were applied on eleven electronic sources. The reported initial search result set contained 1 996 papers. The result was filtered on titles, abstracts and full text and excluded editorials, prefaces, article summaries, interviews, news, reviews, correspondence, discussions, comments, reader's letters, and summaries of tutorials, workshops, panels and poster sessions. The final search result set contained 36 relevant papers that were included in the analysis, giving a precision of 36/1996x100=1.8%.

4.1.3. Sjøberg study

Experiments in software engineering were reviewed (Sjøberg2005), i.e. studies using a specific methodology. All articles in 10 volumes of twelve journals and conference proceedings were searched and the reported initial search result set contained 5 453 papers. The search result set was filtered on titles and abstracts and on full text to exclude editorials, prefaces, article summaries, interviews, news, reviews, correspondence, discussions, comments, reader's letters, and summaries of tutorials, workshops, panels and poster sessions. The final search result set contained 103 papers that were analyzed, which gives a precision of 103/5453x100=1.9%.

4.2. Evaluation criteria

One reason to use an alternative search strategy is to achieve a smaller initial search result set to reduce the amount of work needed for manual filtering. Thus, in this evaluation we compare the number of papers in the initial search result set retrieved by the reference list strategy with the original initial search result set retrieved by the search strategies employed in the systematic reviews, as conducted by Dieste *et al.* (Dieste2008). However, if relevant material should not be missed when reducing the initial search result set, the precision must increase. Thus, we also compare the precision of the reference list search strategy with the original strategies.

As mentioned above, recall is defined as the proportion of retrieved relevant material among all available relevant material. But since it is impossible to know how much relevant material exists, true recall cannot be calculated for either the reference list search strategy or the original search strategies. Instead, we calculate recall in the same manner as Dieste *et al.* do in their evaluation of search strings. We calculate a relative recall by relating the number of the papers in the final search result set of the systematic review to the size of the initial search result set from applying the reference list search strategy. This relative recall indicates the reference list strategy's ability to retrieve the same relevant material as the original strategies.

5. DATA COLLECTION USING REFERENCE LIST STRATEGY

As described above, the search result set of the reference list search strategy consists of the four components: (1) a take-off paper, (2) papers referenced from the take-off paper in two steps, (3) identification of cardinal papers, and (4) papers from external sources referencing cardinal papers. To use the reference list strategy in practice the four components must be further detailed. This section describes the implementation of the search strategy and the data collection in this work.

5.1. Take-off paper

One paper from the final search result set in each of the systematic reviews was selected as take-off papers. Since an older paper is less likely to reference a newer paper than the other way around, we selected the most recent papers as take-off papers and added them as the first item in their initial search result set. The take-off papers used for each of the systematic reviews are shown in Table 2.

Systematic review	Selected take-off paper
Engström study	Applying Regression Test Selection for COTS-based Applications, by Jiang Zheng, Brian Robinson and Laurie Williams, published in Proceedings of International Conference of Software Engineering (ICSE-06), Shanghai, China, 2006 (Zheng2006).
Dybå study	The role of knowledge creation in adopting extreme programming model: an empirical study, by Bouchaib Bahli, El Sayed Abou Zeid, published in Proceedings of ITI 3rd International Conference on Information and Communications Technology: Enabling Technologies for the New Knowledge Society, 2005 (Bahli2005).
Sjøberg study	How much information is needed for usage-based reading? A series of experiments, by Thomas Thelin, Per Runeson, Claes Wohlin, Thomas Olsson, Carina Andersson, published in Proceedings of the International Symposium on Empirical Software Engineering, Nara, Japan, 2002 (Thelin2002).

TABLE 2. The most recent paper for each systematic review was selected as take-off papers.

5.2. Papers referenced from the take-off paper

In step 1, each entry in the take-off paper's reference lists was added to the initial search result set. The initial search result set was represented by a database table created in Microsoft Access, similar to the example presented in Table 1, and the references in the take-off paper's reference lists were added without filtering.

In step 2, we analyzed and filtered the items (references) added to the database table in step 1. For each item that was not a book, tutorial, standard, specification or inaccessible, the corresponding paper's reference lists were retrieved and their references were added to the initial search result set, also without filtering. Items already present in the database table were updated to keep track of the number of references to each item. The results from the data collection in steps 1 and 2 are displayed in Table 3.

Systematic review	Engström study	Dybå study	Sjøberg study
No entries in take-off paper's reference lists (step 1)	33	46	22
No of items after filtering	23	31	14
No of reference entries in step 2	265	1180	254
No of unique items added in the step 2	237	1161	236
Total no of items: take-off+step 1 +step 2	265+33+1=271	1161+46+1=1208	236+22+1=259

TABLE 3. The number of reference list entries found added in step 1 and step 2.

The total number of items represents the number of publications that after step 2 must be processed (manually) to reach a final search result set for analysis in a systematic review. Items in step 1 is not filtered before they are added to the initial search result set to be able have all types of publications as cardinal papers later.

5.3. Identification of cardinal papers

The cardinal papers are those papers in the initial search result set after step 2 having the most number of references to them from other papers in the result set, see Table 4. In the data base table we created, the number of references to each paper in the initial search result set was recorded, thus identifying cardinal papers.

Systematic review	Cardinal paper	No of references
Engström study	A Study of Integration Testing and Software Regression at the Integration Level, by Hareton K. N. Leung and Lee White, published in the Proceedings of Conference on Software Maintenance, 1990.	14
Dybå study	User Acceptance of Computer Technology: A Comparison of Two Theoretical Models, by Fred D. Davis, Richard P. Bagozzi, Paul R. Warshaw, published in Management Science Vol 35 Issue 8, Aug 1989.	6
Sjøberg study	Design and code inspections to reduce errors in program development, by Michael Fagan , published in IBM Systems Journal, Vol 15, No.3, 1976	11

TABLE 4. The cardinal paper for each study and the number of references to them from other papers in the initial search result set after step 2.

5.4. Papers from external sources referencing cardinal papers

The final component of the reference list search strategy is to find papers from external sources referencing the cardinal papers. In this study we used Google scholar since it covers a broad spectrum of sources and has the feature to provide a total count on how many papers referencing the search result. We performed a search on the cardinal paper's title and recorded the number of papers referencing the searched paper as presented by Google Scholar. The result from this search is as follows:

- Engström study: 94 items referencing the cardinal paper
- Dybå study: 2629 items referencing the cardinal paper
- Sjøberg study: 1042 items referencing the cardinal paper

6. ANALYSIS OF THE SEARCH STRATEGY

In this section the results of the reference list search strategy is analyzed and related to the results from the original search strategies reported on in the systematic reviews. With relevant paper we mean a paper that is included in the final search result set analyzed in a systematic review.

6.1. Papers referenced from the take-off paper

For the Engström study, the take-off paper references 6 out of 28 relevant papers, giving a recall of 25% (take-off paper included). Seven of the (after filtering) 23 items added from the take-off paper were relevant papers, giving a precision of 29.2%. The count of 24 papers in the initial search result set after step 1 gives a 99.2% reduction compared to the 2 923 papers in the original initial search result set. For the Dybå study no relevant paper was referenced from the take-off paper and in the Sjøberg study 4 relevant papers were found in the take-off paper's reference list. In step 2, both relevant papers already added in step 1, as well as additional relevant papers, were found. Number of papers found, recall, precision and reduction for both steps 1 and 2 are shown in Table 5.

Systematic review	Engström study	Dybå study	Sjøberg study
#relevant papers in step 1 (incl. take-off)	7	1	5
Recall step 1	25.0% (7/28)	2.8% (1/36)	4.9% (5/103)
Precision step 1 (after filtering)	29.2% (7/24)	3.1% (1/32)	33.3% (5/15)
Reduction step 1 (after filtering)	99.2% (2923->24)	98.2% (1996->32)	99.7% (5453->15)
#relevant papers in step 2	12	1	24
#unique relevant papers in step 2	7	1	21
Recall step 1 + step 2	50.0% (14/28)	5.6% (2/36)	25.2% (26/103)
Precision step 1 + step 2	5.4% (14/271)	0.2% (2/1208)	10.0% (26/259)
Reduction step 1 + step 2	90.7% (2923->271)	39.5% (1996->1208)	95.3% (5453->259)

TABLE 5. The number of relevant papers found in step 1 (including the take-off paper) and step 2.[1]

6.2. Papers from external sources referencing cardinal papers

After performing steps 1 and 2, the last component of the strategy is to include all papers from external sources referencing the cardinal papers. This was, as mentioned above, done by using the Google scholar feature that can

[1] With (*X*->*Y*) we mean that the original search strategy included *X* number of papers in the initial search result set and the reference list search strategy included *Y* number of papers.

produce a count on how many papers that refers to the searched paper. However, this count number may include papers already included in the initial search result which was not accounted for and thus the total recall, precision and reduction reported may actually be conservative.
To calculate precision the number of relevant papers referring to a cardinal paper must be obtained. These numbers was obtained by examining each relevant paper's reference list manually and record the number of relevant papers referencing cardinal papers. It is assumed that the relevant papers found to reference cardinal papers through the manual search are also included in the reference list produced by Google scholar, but this has not been asserted.

Systematic review	**Engström study**	**Dybå study**	**Sjøberg study**
Recall (external papers)	60.7% (17/28)	0.0% (0/36)	27.2% (28/103)
Precision (external papers)	18.1% (17/94)	0.0% (0/2629)	2.7% (28/1042)
Reduction (external papers)	96.8% (2923->94)	-46.7% (1996->2629)	80.9% (5453->1042)
Total recall	82.1% (23/28)	5.6% (2/36)	35.9% (37/103)
Total precision	6.3% (23/365)	0.05% (2/3837)	2.8% (37/1301)
Total reduction	87.9% (2923->365)	-92.2% (1996->3837)	76.1% (5453->1301)
Original precision	0.96%	1.8%	1.9%

TABLE 6. The recall precision and reduction for the use of cardinal papers are shown both separately and together with the step 1 and step 2.

The results from the three studied systematic reviews vary quite much, from 82.1% recall to 5.6%, from 6.3% precision to 0.05%, and from an 87.9% decrease to a 92.2% *in*crease. The Engström study showed the best results in recall, precision and reduction. The strategy did not work so well for the Dybå study where the reference list strategy included almost twice as many papers as the original strategy and found only 2 relevant papers. For the Sjøberg study the reference list strategy collected only less than one quarter of the number of papers as the original strategy.
Even though the precision was better in two of the three studied systematic reviews, some relevant papers were still missed. For the Engström study only two papers were missed, in the Sjøberg study more than half of the relevant papers were missed and for the Dybå study almost all relevant papers were missed. The final step of collecting external papers referencing the cardinal papers was the step giving best recall, as much as 60.7%.

7. EXTENDING THE SEARCH STRATEGY

The final step of collecting external papers referencing a cardinal paper gave the best recall compared to the other steps. To try to improve the strategy further, we extended the use of cardinal papers and studied the effect on recall, precision and reduction. We extended the use of cardinal papers in three directions: 1) Using additional cardinal papers, i.e. not only the most referenced paper, but also the second most referenced, third most, etc. 2) Identifying cardinal papers based on the number of external references in addition to the number of internal references. 3) Selecting cardinal papers with relevant titles.
Furthermore, the original search strategies all use limited search universes which have the benefit of reducing the initial search result set compared to when using an unlimited search universe. We also investigated the effect on the reference list search strategy of using the same limited search universe as used in the original search strategy.

7.1. Additional cardinal papers

In an attempt to improve recall we investigated the use of more than one cardinal paper for collecting references from external sources. For the Engström study we collected external references (from Google scholar) to the seven most referenced papers. The result is displayed in Table 7.

Cardinal paper	**# internal references**	**# external references**	**# new papers found**
2	12	63	0
3	10	39	0
4a	9	186	2
4b	9	209	0
4c	9	85	0
5a	8	78	0
5b	8	71	0
Total		**731**	**2**

TABLE 7. The number of references and relevant papers found for additional cardinal papers in the Engström study.

By adding in worst case (there might be duplicates) 731 external papers, two additional relevant papers were found, giving an increase of the total recall of 7.1 percentage units to 89.3%, a decrease of the total precision of 4.2 percentage units to 2.3%, and a total reduction of 62.5%, a decrease of 25.1 percentage units. The margin precision increase for this additional use of cardinal papers is low, only 0.3%. However, the total precision is still more than twice as high as with the original strategy after adding 731 more papers to examine and the strategy still requires only less than half the number of papers to investigate.
The use of additional cardinal papers in the Dybå study were also studied. The second most referenced paper was used as an additional cardinal paper. This paper had 2 812 external references to it according to Google scholar, giving in worst case a total initial search result set of 6 633 papers, i.e. more than three times the size of the initial search result set produced by the original search strategy. Two additional relevant papers were found, increasing the total recall with 8.2 percentage units, to a total recall of 13.8%. The total precision increased with 0.03 percentage units to 0.08%, which is an improvement, but still below the performance of the original search strategy.
For the Sjøberg study, the use of two additional cardinal papers was investigated. In worst case 615 more papers must be examined to find five additional relevant papers, see Table 8. The total recall increased by 4.9 percentage units to 40.8%, the total precision decresed with 0.6 percentage units to 2.2%. The total reduction was 65.0%, a decrease with 11.1 percentage units compared to using only one single cardinal paper.

Cardinal paper	# internal references	# external refs	# new papers found
2	10	282	2
3	8	333	3
Total		**615**	**5**

TABLE 8. The number of references and relevant papers found for using additional cardinal papers in the Sjøberg study.

7.2. Basing cardinal papers on external references

The reference list search strategy relies on the assumption that relevant papers reference cardinal papers. Another assumption is that good cardinal papers are referenced by more papers from steps 1 and 2. A similar assumption is that good cardinal papers are papers frequently referenced also from external papers found in any source. We investigated this assumption in the Engström study by identifying cardinal papers on the basis of external references from any source as well as internal references from steps 1 and 2.
A paper referenced from all sorts of other papers are not necessary a paper within the field of interest for the systematic review. For example, the most referenced publication in the Engström study was the book “The C programming Language”, by Kernighan and Ritchie, a book far outside the scope of the systematic review topic of regression testing. Thus, to find cardinal papers based on external references and having a higher chance of being on the systematic review topic, we decided that the papers selected should have the most number of external references to them *and* with at least more than half of the maximum number of internal references to them.
The papers fulfilling these requirements in the initial search result set were the papers “Analyzing Regression test selection techniques” (Rothermel1996) and “A Safe Efficient Regression Test Selection Technique” (Rothermel1997). Both these papers are within the research field of regression testing and thus should have the potential to be referenced by relevant papers and thus suitable to use as cardinal papers. The results from the use of these two papers as cardinal papers are displayed in Table 9.

Cardinal paper	#external references	# internal references	Recall (28)	#unique papers	Precision
1 (Rothermel1996)	209	9	64,2% (18)	7	8,6%
2 (Rothermel1997)	186	9	53,4% (15)	7	7,8%

TABLE 9. The results for using cardinal papers that are based on the number of external references.

When replacing the results from the use of the standard cardinal papers with the results of the first of these new cardinal papers, we found 21 relevant papers out of the 28. This is a decrease of 7.1 percentage units, to a total recall of 75.0%. The initial search result set’s total size using the first cardinal paper is 480 papers which gives a total precision of 4.4%, a decrease of 1.9 percentage units. When the second new cardinal paper is added, two additional relevant papers are found, thus the recall reaches 82.1%, the same level as with the standard cardinal papers. The precision and reduction is still lower since there are more references to these new cardinal papers than to the standard cardinal paper. The total precision decreases by 2.8 percentage units to 3.4% and the total reduction decreases by 10.3 percentage units to 77.2%.

7.3. Select cardinal paper based on title relevance

We investigated using cardinal papers that were identified on the basis of the relevance of their titles in the Dybå study. The most referenced publication with a relevant title with respect to the systematic review topic, i.e. agile development, was the book "Extreme programming explained" by Beck. We consider this book to be cardinal within the field of agile development and suitable to serve as a cardinal publication.

There were 4 072 references to this book title according to Google Scholar, i.e. more than twice as many as the initial search result set from the original search strategy. However, 25 of the relevant papers refer to this book, resulting in a recall of almost 70%, and a total recall of 72.2%, an increase with 66.7 percentage units. The precision for this search was 0.6%, and the total precision is 0.5%, still worse than the original search strategy.

Using cardinal paper based on relevant title gave a great increase of recall. However, the title used as cardinal publication was a heavily referenced book, and thus suffering from the problem of increasing the search set size and lowering the precision.

7.4. Limiting the search universe

In the Dybå study, the reference list strategy produced a larger initial search result set than the original search strategy did. An explanation for this may be that the original search strategy was limited to a predefined set of electronic sources, a limitation not exploited in the original description of the reference list search strategy. Thus, we analyzed the effect of restricting the search universe for the reference list search strategy by excluding the references that could not be found in any of the sources Dybå *et al*. used in their search. This was performed by searching for each item in the initial search result set in the same set of electronic sources used in the original search strategy and excluding those items that could not be found in any of the sources.

The take-off paper's reference list had 46 entries, of which 11 could not be found in the limited search universe. However, 9 out of these 11 were books or inaccessible and thus already excluded in the second step filtering. The two remaining entries referenced together 94 papers in the second step. From the initial search result set we excluded these two papers and the 94 papers referenced from them. Thus, leaving us with 1 112 papers in the initial search result set from which we removed additional 535 papers from step 2 that were not found in the search universe. This resulted in an initial search result set of 577 papers after the second step, a 71% reduction compared to the original search strategy. Reducing the number of papers with more than half, more than doubles the precision, from poor 0.16% to 0.35%. This is still a low precision and is due to the poor recall in that only one paper was found with the reference list search strategy in this study.

8. CONCLUSIONS

In this paper we have described and evaluated a search strategy for systematic reviews in software engineering based on reference lists. The goal of improving search strategies was to reduce the number of primary studies in the initial search result set that must be filtered manually. The initial search result set should preferably be reduced without sacrificing recall, i.e. without missing relevant papers. Reducing the initial search result set size without also decreasing recall requires the precision to increase, i.e. a larger portion of the whole initial result set should be relevant papers.

In two of the three studied systematic reviews the reference list strategy had better performance in terms of precision than the original search strategies. However, relevant papers were still missed to various degrees, from less than 20% missed papers in the Engström study to 95% missed in the Dybå study with the Sjøberg study in between. Extending the strategy with cardinal papers did not lead to any great increase on the recall. It contributed more to reduce the precision and increasing the initial search result than to find more relevant papers.

The reference-based search strategy seems to work best for the study on a specific software engineering topic, in the Engström case, regression test selection. For the Dybå study, better results would have been expected, since this also reviews a specific type of development process. Apparently, the references to agile methods are so many and diverse that they are not strict enough guides for the systematic review. The search strategy performed better than expected for the Sjøberg study, being a research methodology focused review. However, since about one third of the experiments were on inspections, and the take-off paper was on inspections, this might have improved the results for this study.

In summary, the proposed reference-based search strategy increased the precision of the systematic review, without sacrificing the recall too much for the technically focused systematic review on regression testing. However, for the more general agile methods systematic review, and the review of experiments, results were not satisfactory. We propose further evaluation of the strategy, and also acknowledge the need for tool support in extracting the reference information used in the search strategy.

ACKNOWLEDGEMENTS

The work is funded by The Swedish Governmental Agency for Innovation Systems under grant 2005-02483 and the Swedish Research Council under grant 622-2004-552 for a senior researcher position in software engineering.

REFERENCES

Bahli B. and Zeid, E.S.A. (2005) The role of knowledge creation in adopting extreme programming model: an empirical study. Proceedings of ITI 3rd International Conference on Information and Communications Technology: Enabling Technologies for the New Knowledge Society, Cairo, Egypt, 5-6 December, pp. 75-87.

Bailey P., Craswell N. and Hawking D. (2003) Engineering a multi-purpose test collection for web retrieval experiments. Information Processing & Management, **39**, 853-871.

Brereton P., Kitchenham B.A., Budgen D., Turner M. and Khalil M. (2007) Lessons from applying the systematic literature review process within the software engineering domain. Journal of Systems and Software, **80**, 571-583.

Dieste O., Grimán A. and Juristo N. (2008) Developing search strategies for detecting relevant experiments. Empirical Software Engineering. DOI: 10.1007/s10664-008-9091-7.

Dybå T. and Dingsøyr T. (2008) Empirical studies of agile software development: A systematic review. Information and Software Technology, **50**, 833-859.

Dybå T., Dingsøyr T. and Hanssen G. (2007) Applying systematic reviews to diverse study types: An Experience Report. Proceedings of First International Symposium on Empirical Software Engineering and Measurement, Madrid, Spain, pp. 225-234.

EBSE-2007-01 (2007) Guidelines for performing systematic literature reviews in software engineering - Version 2.3. Keele University and University of Durham, UK.

Engström E., Skoglund M. and Runeson P. (2008) Empirical Evaluations of Regression Test Selection Techniques: A Systematic Review. Proceedings of International Symposium on Empirical Software Engineering and Measurement, Kaiserslauten, Germany, pp. 22-31.

Jørgensen A. and Sheppard M (2007) A systematic review of software development cost estimation studies. IEEE Transactions of software engineering, **33**, 33-53.

Kessler M. (1963) Bibliographic coupling between scientific papers. American Documentation, **14**, 10.

Martyn J. (1964) Bibliographic coupling. Journal of documentation, **20**, 236.

Rothermel G and Harrold M. (1996) Analyzing Regression test selection techniques. IEEE Transactions on Software Engineering, **22**, 529-551.

Rothermel G. and Harrold M. (1997) A Safe Efficient Regression Test Selection Technique. ACM Transactions on Software Engineering and Methodology, **6**, 173-210.

Sjøberg D., Hannay J., Hansen O., Kampenes V., Karahasanovic, Liborg N. and Rekdal A. (2005) A survey of controlled experiments in software engineering, IEEE Transactions of software engineering. **31**, 733-753.

Thelin, T., Runeson, P., Wohlin, C., Olsson, T. and Andersson, C. (2002) How much information is needed for usage-based reading? A series of experiments. Proceedings of International Symposium on Empirical Software Engineering, Nara, Japan, pp. 127-138.

Tsay M., Xu H. and Wu C. (2003) Journal co-citation analysis of semiconductor literature. Scientometrics, **57**, 7-25.

Van Rijsbergen C. J. (1979) Information Retrieval, Department of Computing Science. Glasgow: University of Glasgow.

Zheng J., Robinson B., Williams L. and Smiley K. (2006) Applying Regression Test Selection for COTS-based Applications. Proceedings of International Conference on Software Engineering, Shanghai, China. pp. 512-522.