Developing Search Strategies for Detecting Relevant Experiments for Systematic Reviews

Oscar Dieste
Universidad Politécnica de Madrid.
Madrid - Spain
odieste@fi.upm.es

Anna Grimán Padua¹
Universidad Simón Bolívar.
Caracas - Venezuela
agriman@usb.ve

Abstract

Information retrieval is an important problem in any evidence-based discipline. Although Evidence-based Software Engineering (EBSE) is not immune to this fact, this question has not been examined at length. The goal of this paper is to analyse the optimality of search strategies for use in systematic reviews. We tried out 29 search strategies using different terms and combinations of terms. We evaluated their sensitivity and precision with a view to finding an optimum strategy. From this study of search strategies we were able to analyse trends and weaknesses in terminology use in articles reporting experiments.

1. Introduction

The evidence-based concept originated in the field of medicine with what is known as evidence-based medicine (EBM). EBM looks for evidence to help solve clinical problems [13]. A key aspect in EBM is the combination of results. Results that can be combined to gather evidence can be found in specialized databases like MEDLINE, AIDSLINE, CANCERLIT, Dissertation Abstracts Online, and TOXLINE.

The concept of evidence-based software engineering (EBSE) emerged through an analogy with EBM. EBSE involves the performance of systematic reviews as a means of gathering evidence [8]. A key aim of systematic literature review is to combine empirical results related to a technology. Systematic literature review identifies, evaluates and interprets all available relevant research that can answer a specific research question. Systematic reviews, then, are secondary studies that are produced with the help of primary studies [7].

Systematic reviews identify relevant empirical studies on the basis of a search strategy [3]. It is vital for the strategy to detect as much relevant material as possible. Leaving relevant results out of a systematic review could lead to the generation of inaccurate evidence.

Because of the rapid growth of knowledge, the information required for systematic reviews is becoming more and more difficult to identify and retrieve. This problem has been dealt with at length in the field of medicine, where the importance of taking a systematic approach to information retrieval has been stressed as a way of gathering unbiased information and of estimating costs and effort [10].

Developing search strategies that retrieve as much relevant information as possible, while keeping cost and effort low is critical for generating evidence. Such search strategies are known as optimal strategies. The goal of our work has been to analyse the optimality of a number of search strategies for systematic reviews. This analysis may be of use in the early stages of systematic reviews of empirical studies.

To be able to evaluate optimality, we have established a gold standard: the empirical studies gathered through a manual and exhaustive review conducted in [12]. We have tried out 29 strategies and compared the results against the gold standard to evaluate their optimality. By evaluating the results of the different strategies we have been able to derive recommendations on searching as part of systematic reviews. As a spin-off of this research, we have identified trends and weaknesses in the standardization of the terms used in reported experiments.

Section 2 of this article presents some definitions related to search strategies and search optimality. Section 3 describes how the gold standard and the

¹ Anna Grimán is also a postgraduate student at the Technical University of Madrid.
search universe were established. Section 4 analyses the search strategies against the established standard. Section 5 sets out search recommendations. Section 6 analyses the trends in term usage in published experiments. Section 7 discusses the importance of establishing a proper search universe. Finally, section 8 presents our conclusions.

2. Optimal search strategy

A search’s optimality is based on its sensitivity and its precision [2]. A search’s sensitivity is its ability to identify all of the relevant material. On the other hand, precision is the amount of relevant material there is among the material retrieved by the search. In other words, precision is the strategy’s ability to detect no or few irrelevant articles. The diagram in Figure 1 shows the relationship between these two concepts.

In the context of this work, relevant material will be defined as any articles that report Software Engineering (SE) experiments.

![Diagram of search sensitivity and precision](image)

**Figure 1. Search sensitivity and precision**

In Figure 1, D is the search universe; (A+B) represent the search results; A is the set of relevant articles retrieved by the strategy; B is the set of irrelevant articles retrieved by the strategy; C represents the set of relevant articles not detected by the search strategy. Sensitivity and precision can be formulated as a ratio between A, B and C: Sensitivity = A/ (A+C); Precision = A/(A+B).

The smaller C (relevant articles that the search missed) is, the higher the sensitivity score will be. If C is zero (all relevant articles have been detected), search strategy sensitivity is 100%. A search with low sensitivity detects very few relevant articles. A systematic review conducted on the basis of such a search will miss rather a lot of relevant experiments. An incomplete systematic review will lead to unreliable evidence. The smaller B (irrelevant articles gathered in the search) is, the greater precision will be. When B is zero (no irrelevant material is detected), search strategy precision is 100%. A search strategy with low precision will lead to a lot of irrelevant articles being retrieved. This might be seen as less of a problem than the search with low sensitivity. Note, however, that it will take a huge manual review effort to identify and reject irrelevant articles.

Although there are no explicit values for typifying the different strategies, we have inferred them from the sensitivity and precision ranges in [4] and [13] (as shown in Table 1) with a view to evaluating the results of our searches.

<table>
<thead>
<tr>
<th>Strategy type</th>
<th>Sensitivity range</th>
<th>Precision range</th>
<th>Goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>High sensitivity</td>
<td>85-99%</td>
<td>7-15%</td>
<td>Maximum sensitivity despite poor precision</td>
</tr>
<tr>
<td>High precision</td>
<td>40-58%</td>
<td>25-60%</td>
<td>Maximum precision despite poor sensitivity</td>
</tr>
<tr>
<td>Optimum</td>
<td>80-99%</td>
<td>20-25%</td>
<td>Maximization of both ranges (sensitivity and precision)</td>
</tr>
<tr>
<td>Acceptable</td>
<td>72-80%</td>
<td>15-25%</td>
<td>Good enough sensitivity and precision</td>
</tr>
</tbody>
</table>

Table 1. Search strategy scales

There will always be a trade-off between sensitivity and precision in any search strategy because irrelevant documents are more likely to be retrieved the higher sensitivity is. Optimal search can be defined as a search that strikes a balance between high sensitivity and high precision [10]. Depending on the goals of the systematic review and available resources, we might want to use a high sensitivity search strategy or a high precision strategy. But, generally, the goal will be to get an optimal, or at least acceptable, search strategy.

3. Establishing the gold standard and search universe

To be able to calculate sensitivity and precision, we need to know how much relevant material there is in a set of publications, i.e. (A+C) in Figure 1. However, this parameter is impossible to establish beforehand. Therefore, we need to use a gold standard to calculate the search’s precision and sensitivity.

This paper looks at the part of the search defined by the type of SE empirical study. At this early stage of the research, we limited the scope of the type of empirical study, focusing on the search for experiments. The part of the search describing the empirical study domain (requirements, design, etc.) is outside the scope of this paper. Research into this part of the search is just as necessary, but has been left for the second phase of our research.

Therefore, the gold standard for our work should be a set of manually and exhaustively reviewed SE experiments that have been checked for conformity to
the selection criteria. We found one such benchmark, establishing the set of 103 articles reported in [12] as the gold standard for our work. This set qualifies as a gold standard, because, as reported in [12], one researcher systematically read all the titles and abstracts; if it was unclear from the title or abstract whether the article described a controlled experiment, the same researcher and another person read the full text of the article.

The definition of SE experiment used in [12] was strict, as it considered only studies where the experimental units were individuals or teams conducting one or more SE tasks. By adopting this survey as our gold standard, we are accepting this definition and, therefore, looking for articles that report this type of experiments.

After an exhaustive manual revision, the 103 articles were selected from a universe of 5,453 articles printed in 12 publications of repute: ACM Transactions on Software Engineering Methodology (TOSEM), Empirical Software Engineering (EMSE), IEEE Computer, IEEE International Symposium on Empirical Software Engineering (ISESE), IEEE International Symposium on Software Metrics (METRICS), IEE Software, Information and Software Technology (I&ST), International Conference on Software Engineering (ICSE), Journal of Systems and Software (JSS), Software Maintenance and Evolution (SM&E), Transactions on Software Engineering (TSE), and Software Practice and Experience (SP&E).

Having selected the publications and the time period for exploration, we looked at the different document databases accommodating these publications. We applied test search strategies to select one of the bibliographic repositories. We found that there were important limitations that could affect search strategy sensitivity:

- **Limited bibliographic resources.** Some databases do not cover a broad spectrum of publications but are confined to just publications by one publisher. This applies to IEEEExplore®, Springerlink, ScienceDirect® and ACM Digital Library. This is an impediment for searching because each search strategy has to be applied over again on different search engines for combination at a later date.

- **Problems with the search algorithm.** In both IEEEExplore and ACM DL, the search cannot be run on certain fields. For example, IEEEExplore’s Basic Search option searches all fields to retrieve articles. As we will see later, it is important to confine the search to titles and abstracts to achieve satisfactory precision.

- **Failure to recognize plurals.** In some cases, like IEEEExplore (Basic Search option), the search engine did not search both the singular and plural of the specified term. This meant that the plural form of each term had to be entered separately.

- **Incomplete article abstracts or full texts.** In some cases, like IEEEExplore and ACM DL, the abstract or the full text cannot always be accessed because they are sometimes not linked to the article title. This is an important weakness, as it makes it difficult to retrieve articles and for researchers to check abstracts and texts.

We decided to use the SCOPUS™ document database. It was developed by Elsevier and it is available on www.scopus.com. The reason for this selection was that SCOPUS had fewer of the above snags. SCOPUS™ includes 10 of the 12 publications specified in the gold standard. We removed ISESE and METRICS from the set of 103 papers due to search engine limitations. This meant that we were left with a gold standard of 90 articles reporting SE experiments.

Now that we have established the gold standard, the search universe and the bibliographic database that we are going to use, let us move on to run and evaluate the different search strategies.

## 4. Search strategy analysis

The first strategy that we tried out was denoted by the type of empirical study that we wanted to retrieve: controlled experiments. Thus, we used the search term *experiment*. We searched titles and abstracts to remove irrelevant documents that might occasionally contain the word *experiment* in the main body of the text but were not primarily focused on reporting an experiment.

For this first strategy, which we called STR1 (see Table 2), the strategy has a sensitivity of 76.6% and a precision of 19.6%. The sensitivity level is acceptable; however, precision needs to be improved. Note that the title and abstract of 283 articles would have to be read before being rejected as irrelevant. This can be a problem in systematic reviews with limited resources, but not necessarily, for example, if there are more than enough resources (people and time) to do this job.

We compared the results of this search strategy against the search for the same term in all the document fields. Our aim was to check that the strategy of searching titles and abstracts was appropriate.

The strategy of searching all the fields has a sensitivity of 82.2%, which is quite an acceptable value, and a precision of 12%, which is very low. An extra 260 articles need to be reviewed exhaustively to get just an extra five relevant experiments. This is a high cost for any review process. We consider that this result confirms that searching titles and abstracts rather than the full text is a better strategy. The next step is to optimize strategy STR1 by increasing its precision and, if possible, also its sensitivity.
To do this, we ran further searches using terms that appeared recurrently in titles and abstracts in gold standard articles that were not retrieved by *experiment*.

One of the most frequent terms was *empirical study* (see Table 2). STR2’s sensitivity and precision are 10%. These scores are rather low, which means that STR2 should not be used as a separate strategy for detecting experiments. It might perhaps be worthwhile combining this strategy with *experiment*.

This combination is represented by strategy STR3 (see Table 3). The sensitivity and precision scores for STR3 are 84.4% and 17.7%, respectively. With respect to *experiment*, these scores amount to a 7.8% increase in sensitivity and a 1.9% decrease in precision. In other words, sensitivity is better than for STR1 but precision is not, as we get seven more relevant articles in return for an extra 70 irrelevant articles. Therefore, the term *experiment* should be added to the term *empirical study* only if you are looking for a highly sensitive search strategy, but not if you want to optimize STR1.

Other recurrent terms within the set of experiments belonging to the gold standard not retrieved by the above strategies are: *experimental study, empirical evaluation, experimentation, experimental comparison, experimental analysis, experimental evidence, experimental setting* and *empirical data*. Let’s look at how each of these terms behaves both separately and combined with other strategies. The use of these terms will not retrieve 100% of the articles in the gold standard because some of the articles did not include any of the synonyms of or terms closely related to *experiment*. However, they can help to maximize the sensitivity and precision of STR1.

Therefore, the next step was to try out *experimental study*, as STR4 (see Table 2). The result was a rather low sensitivity for STR4 at under 3%. Therefore, it cannot be classed as a good strategy for detecting experiments. However, STR4’s precision is fairly high at 37.5%.

Let’s combine *experimental study* and *experiment* as STR5 (see Table 3). There is a 1.1% (77.7% for STR5) increase in sensitivity and precision is unchanged (at 19.6%). This means that *experimental study* adds a few relevant experiments to *experiment*, whereas it retrieves no other irrelevant articles, which is a definite improvement.

In pursuit of improved sensitivity and precision levels, these results led us to analyse another strategy (STR6) based on the combination of *experiment, empirical study* and *experimental study* (see Table 4). The sensitivity and precision scores are 85.5% and 17.8%, respectively. Compared with the gain achieved by combining just *experiment* and *experimental study* (STR5), STR6 brings with it a 7.8% gain in sensitivity, plus a 1.8% loss in precision. This can interpreted as STR6 improving on STR5 but at a cost in terms of precision. Compared with *experiment*, it provides greater sensitivity than STR5 in return for a drop in precision (see Tables 3 and 4).

STR7 analyses the behaviour of the term *empirical evaluation* (see Table 2). STR7’s sensitivity score is rather low, 2%, whereas its precision is quite high, 7.7%. This is another strategy that should not be used on its own. Combining *empirical evaluation* with *experiment* as STR8 (see Table 3) yields a 2.2% increase in the sensitivity of *experiment*. However, the precision of this strategy is cut by 0.7%. This helps to maximize the sensitivity of STR1, but does not optimize this strategy.

Combining *empirical evaluation* with *experiment, empirical study* and *experimental study* (STR6) as STR9 (see Table 4), we find that STR9 improves STR6’s sensitivity by 1.1% with a 0.5% drop in precision. At the same time, it represents a sensitivity gain of 10% over *experiment*, which is a very notable increase, in exchange for a 2.3% loss in precision.

The next term we analysed (as STR10) was *experimentation* (see Table 2). Again this is a strategy with rather low sensitivity and precision, 3.3% and 8.8%, respectively.

These scores mean that *experimentation* should not be used on its own. Its combination with *experiment* (as STR11) increases STR1’s sensitivity by 1.1% in return for a 0.9% drop in precision (see Table 3).

---

**Table 2. One-term search strategy results and properties**

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Search term</th>
<th>Results</th>
<th>Sensitivity (S)</th>
<th>Precision (P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>STR1</td>
<td>experiment</td>
<td>352 retrieved, 69 matched</td>
<td>76.6%</td>
<td>19.6%</td>
</tr>
<tr>
<td>STR2</td>
<td>empirical study</td>
<td>87 retrieved, 9 matched</td>
<td>10%</td>
<td>10%</td>
</tr>
<tr>
<td>STR4</td>
<td>experimental study</td>
<td>8 retrieved, 3 matched</td>
<td>3%</td>
<td>37.5%</td>
</tr>
<tr>
<td>STR7</td>
<td>empirical evaluation</td>
<td>26 retrieved, 2 matched</td>
<td>2%</td>
<td>7.7%</td>
</tr>
<tr>
<td>STR10</td>
<td>experimentation</td>
<td>34 retrieved, 3 matched</td>
<td>3.3%</td>
<td>8.8%</td>
</tr>
<tr>
<td>STR13</td>
<td>experimental comparison</td>
<td>9 retrieved, 7 matched</td>
<td>7.7%</td>
<td>77.7%</td>
</tr>
<tr>
<td>STR16</td>
<td>experimental analysis</td>
<td>2 retrieved, 1 matched</td>
<td>1.1%</td>
<td>50%</td>
</tr>
<tr>
<td>STR19</td>
<td>experimental evidence</td>
<td>2 retrieved, 1 matched</td>
<td>1.1%</td>
<td>50%</td>
</tr>
<tr>
<td>STR22</td>
<td>experimental setting</td>
<td>2 retrieved, 1 matched</td>
<td>1.1%</td>
<td>50%</td>
</tr>
<tr>
<td>STR25</td>
<td>empirical data</td>
<td>21 retrieved, 1 matched</td>
<td>1.1%</td>
<td>4.7%</td>
</tr>
</tbody>
</table>

---
This means that *experimentation* provides one more relevant article and an extra 31 irrelevant articles. Its suitability for improving *experiment* is therefore questionable (see Table 3).

The result of combining all the previous terms as STR12 would be a sensitivity of 87.7% and a precision of 16.7% (see Table 4). This new strategy provides a 1.1% increase in sensitivity over the combination of *experiment*, *empirical study*, *experimental study* and *empirical evaluation*, with a 0.6% drop in precision.

Nevertheless, the fall in precision is still rather large (greater than 0.5%), meaning that *experimentation*’s contribution to the other strategies is arguable. Table 4 also shows while this new strategy increases the sensitivity of *experiment* by 11.1%, this improvement is offset by a near 3% drop in precision.

Next we examined the term *experimental comparison* as the strategy labelled STR13 (see Table 2). STR13 is a strategy with very low sensitivity (7.7%) and high precision (77.7%). Therefore, this strategy should not be used on its own. However, its combination with *experiment* as STR14 increases STR1’s sensitivity by 2.2% and precision by 0.4%. This is a definite improvement on STR1 (see Table 3). If we combine *experimental comparison* with *experiment*, *empirical study*, *experimental study*, *empirical evaluation* and *experimentation* (STR12) as STR15, there is a 2.3% increase in STR12’s sensitivity, which reaches 90%, as well as a 0.4% increase in its precision (17.1%) (see Table 4). Compared with *experiment*, this is a 13.4% gain in sensitivity in return for a 2.5% drop in precision.

A similar analysis was run for the terms *experimental analysis* (STR16), *experimental evidence* (STR19), *experimental setting* (STR22) and *empirical data* (STR25). When calculating their sensitivity and precision separately (see Table 2), we found that *experimental analysis*, *experimental evidence* and *experimental setting* have a relatively low sensitivity (1.1%) and a precision of 50%. As a result these terms are not used on their own. Unlike the above terms, the term *empirical data* has very low scores for both properties.

If we try combining each of these terms with *experiment* (see Table 3), we find that they lead to a small increase in sensitivity (1.1%). The terms *experimental evidence* and *experimental setting* also generate modest increases in precision (0.1%), slightly improved upon by *experimental analysis* (0.2%). In the case of *empirical data*, on the other hand, the result is a 0.6% loss in precision.

The result of adding each of these terms to the combination of all the above terms is similar (see STR21, STR24, and STR27 in Table 4). All the terms except *empirical data* increase the sensitivity and precision of *experiment* and the other terms. The combination of *empirical data* with the other terms (STR27) fails to increase sensitivity, whereas it does reduce precision. Therefore, *empirical data* adds no value and actually degrades the above combination of all the terms. Consequently, STR27 is rejected for the purposes of this work.

This way we tested all the relevant terms that were synonyms of *experiment*. Table 2 shows the results (number of articles retrieved, number of articles that matched the gold standard) and properties (Sensitivity and Precision) of the single-term strategies. It is clear from this table that *experiment* is the only term to have a good enough sensitivity.

Table 3 shows the properties of the strategies combining each analysed term with the term *experiment*. It also indicates the changes in the
sensitivity ($\Delta S_e$) and precision ($\Delta P_e$) of experiment when combined with another term. Table 4 shows the properties of the strategies aggregating each analysed term. It indicates the changes to sensitivity and precision compared with experiment ($\Delta S_e$, $\Delta P_e$) as well as with the aggregation of all other terms ($\Delta S_a$, $\Delta P_a$).

It is clear from Table 4 that the combination of experiment, empirical study, experimental study, empirical evaluation, experimentation, experimental comparison, experimental analysis, experimental evidence and experimental setting (STR24) is the strategy with the highest sensitivity (93.3%). This is a sizeable gain over experiment; however, the 2% loss in precision means that another 125 articles will have to be reviewed to get an extra 15 relevant articles.

To optimize these results, we decided to run a new strategy, STR28 (see Table 5), which removes the term experimentation, since the sensitivity improvement it provides is questionable because of the precision detriment it provokes as compared with experiment (see Tables 3 and 4).

### Table 4. Properties of aggregated search strategies

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Combination</th>
<th>Search terms</th>
<th>Results</th>
<th>$S$</th>
<th>$P$</th>
<th>$\Delta S_e$</th>
<th>$\Delta P_e$</th>
<th>$\Delta S_a$</th>
<th>$\Delta P_a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>STR6</td>
<td>STR1+STR2+</td>
<td>experiment OR empirical study OR experimental study</td>
<td>432 retrieved, 77 matched</td>
<td>85.5%</td>
<td>17.8%</td>
<td>+8.9</td>
<td>-1.8</td>
<td>+1.1 **</td>
<td>+0.1 **</td>
</tr>
<tr>
<td>STR9</td>
<td>STR1+STR2+</td>
<td>experiment OR empirical study OR experimental study</td>
<td>450 retrieved, 78 matched</td>
<td>86.6%</td>
<td>17.3%</td>
<td>+10</td>
<td>-2.3</td>
<td>+1.1</td>
<td>-0.5</td>
</tr>
<tr>
<td>STR12</td>
<td>STR1+STR2+</td>
<td>experiment OR empirical study OR experimental study OR empirical evaluation</td>
<td>471 retrieved, 79 matched</td>
<td>87.7%</td>
<td>16.7%</td>
<td>+11.1</td>
<td>-2.9</td>
<td>+1.1</td>
<td>-0.6</td>
</tr>
<tr>
<td>STR15</td>
<td>STR1+STR2+</td>
<td>experiment OR empirical study OR experimental study OR empirical evaluation OR experimental comparison</td>
<td>472 retrieved, 81 matched</td>
<td>90%</td>
<td>17.1%</td>
<td>+13.4</td>
<td>-2.5</td>
<td>+2.3</td>
<td>+0.4</td>
</tr>
<tr>
<td>STR18</td>
<td>STR1+STR2+</td>
<td>experiment OR empirical study OR experimental study OR empirical evaluation OR experimental comparison OR experimental analysis</td>
<td>473 retrieved, 82 matched</td>
<td>91.1%</td>
<td>17.3%</td>
<td>+14.5</td>
<td>-2.3</td>
<td>+1.1</td>
<td>+0.2</td>
</tr>
<tr>
<td>STR21</td>
<td>STR1+STR2+</td>
<td>experiment OR empirical study OR experimental study OR empirical evaluation OR experimental comparison OR experimental analysis OR experimental evidence</td>
<td>475 retrieved, 83 matched</td>
<td>92.2%</td>
<td>17.4%</td>
<td>+15.6</td>
<td>-2.2</td>
<td>+1.1</td>
<td>+0.1</td>
</tr>
<tr>
<td>STR24</td>
<td>STR1+STR2+</td>
<td>experiment OR empirical study OR experimental study OR empirical evaluation OR experimental comparison OR experimental analysis OR experimental evidence OR experimental setting</td>
<td>477 retrieved, 84 matched</td>
<td>93.3%</td>
<td>17.6%</td>
<td>+16.7</td>
<td>-2</td>
<td>+1.1</td>
<td>+0.2</td>
</tr>
<tr>
<td>STR27</td>
<td>STR1+STR2+</td>
<td>experiment OR empirical study OR experimental study OR empirical evaluation OR experimental comparison OR experimental analysis OR experimental evidence OR experimental setting OR empirical data</td>
<td>491 retrieved, 84 matched</td>
<td>93.3%</td>
<td>17.1%</td>
<td>+16.7</td>
<td>-2.5</td>
<td>=</td>
<td>-0.5</td>
</tr>
</tbody>
</table>

** Calculated against STR5
tested strategies will behave as observed here. Even if we relaxed the article selection criteria, the strategies exercised in our paper could be higher than reported here.

Although a highly sensitive search strategy is capable of identifying more relevant material, there is a trade-off between the amount of potentially eligible empirical studies that such a strategy can retrieve and the additional effort required to determine whether or not they are relevant. Therefore, the selection of the best strategy depends on the characteristics and resources of the systematic review.

However, we believe that some general recommendations can be concluded from the above analysis.

Searching by the term experiment can be considered as a strategy with acceptable sensitivity and precision. Searching by all the tested terms (STR24) is a high sensitivity strategy. Note, however, that the strategy’s precision is worse than for STR2. Therefore, you should use this strategy if there is a shortage of resources for your systematic review, as the abstract of 110 extra irrelevant articles will need to be reviewed to locate just 15 relevant articles more than STR2. On the other hand, searching by the synonyms of experiment improves the sensitivity of STR1 considerably, without detracting from precision. This means that combining experiment with its most closely related synonyms (experimental study, experimental comparison, experimental evidence and experimental setting) improves both properties of STR1. Therefore, looking at and comparing the sensitivity and precision of STR29 with the strategies listed in Table 1, we can say that STR29 is an optimal strategy (good sensitivity and precision). None of the other combinations of terms yields good sensitivity without detracting from precision. The above quantitative analysis can be interpreted qualitatively as follows:

- Using just the term experiment to search the titles and abstracts and detect controlled SE experiments is by no means a bad strategy. Quite a lot of experiments are left out (in our case, 25%). On the other hand, the number of irrelevant articles is one of the best (a precision of 19.6% is quite good). Therefore, if what you are looking for is a quick search, do not hesitate to

### 5. Recommendations on search strategies

Depending on reviewers’ interests, the aim could be to maximize the amount of material retrieved (high sensitivity search), maximize relevant material retrieved (high precision search) or optimize the retrieval of relevant material (optimal search).

<table>
<thead>
<tr>
<th>Table 5. Properties of the optimized search strategies</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Strategy</strong></td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>STR28</td>
</tr>
<tr>
<td>STR29</td>
</tr>
</tbody>
</table>

STR28 achieves a sensitivity of 92.2% and a precision of 17.6%; i.e, there is a loss in sensitivity and a 0.5% gain in precision compared with STR24. Table 5 shows that this strategy improves experiment in terms of sensitivity but not in terms of precision.

Looking at these results, we found that the strategy could still not be considered optimal. Analysing Tables 3 and 4, we found that the other terms that detracted from the precision of experiment were empirical study and experimental evaluation. We then tried removing these two terms from STR28.

Sensitivity after combining just experiment and its closest synonyms (experimental study, experimental comparison, experimental analysis, experimental evidence and experimental setting), as STR29, is 83.3% (10% less than for STR24) and precision is 20.7% (3.1% more than STR24). These values can be considered optimum. The improvement in precision also means that 129 articles, of which only 9 reported relevant experiments, no longer have to be reviewed. Table 5 also shows how STR29 improves on experiment.

We should also point out that because of the strict definition of SE experiment used in [12], which we mentioned in section 3, we can assume that the precision scores for the strategies exercised in our paper could be higher than reported here.

This does not affect the results that we have achieved, as this definition was applied globally to compare the strategies. Therefore, we can say that, even if we relaxed the article selection criteria, the tested strategies will behave as observed here.
use this strategy.

- To reduce the number of relevant papers that the term *experiment* does not find, you should add close and accepted synonyms that authors may have used instead of that term. The most commonly used synonyms in the gold standard articles that do not detract from the precision achieved by *experiment* are: *experimental study*, *experimental comparison*, *experimental analysis*, *experimental evidence*, and *experimental setting*. Combining these terms, we get good sensitivity (83.3%) and precision (20.7%) scores, overlooking 16.7% of the articles in the gold standard.

- If you add other more general terms like *experimentation*, *empirical study*, and *empirical evaluation* to these synonyms of the term *experiment*, you detect many more of the gold standard experiments (93.3%). On the other hand, you lose out in terms of an increase in the number of irrelevant articles.

- You should not use any of these synonyms, not even *empirical study*, alone as they omit the huge majority of experiments.

Looking at the diversity of terms used in the reviewed articles on experiments, it is worthwhile analysing their evolution over time.

### 6. Analysis of the trends in term usage in published experiments

We focused on the analysis of the most recurrent terms (*experiment*, *empirical study*, *experimental study*, *empirical evaluation* and *experimental comparison*). Figure 2 shows this trend based on the experiments of the gold standard.

![Figure 2. Trends in term use](image)

Note that the predominant term is *experiment*. Nevertheless, there is no confirmed trend, i.e. the myriad of terms used to refer to experiments since 1993 cannot be said to be improving. For example, *experiment* was the only term used in 1995 (100% of the articles reporting experiments), but it was only used in 63.6% of the articles reporting experiments in 1997. The trend appeared to be reversing in 2001 (the term was used in 84.6%). However, the term *experiment* was used in the title or abstract of only 57.1% of the experiments published in 2002, whereas *empirical study* and other terms (*experimental analysis*, *experimental evidence*, *experimental setting*, and *empirical data*) appeared in the others.

On the other hand, the use of the term *experimental comparison* is neither especially significant (with the exception of 1993) nor very constant over the years. Accordingly, we can say that this term hardly ever appears or has fallen into disuse. *Experimental study* appears only in 2000, when it was used by 4% of the articles reporting experiments and then disappeared. This is evidence of a fairly erratic use in the analysed period, despite it being one of *experiment*’s most closely related synonyms.

With respect to *empirical study*, we find that it is a term with a modest, but constant presence over the years. In other words, although it is a broader term than *experiment*, it is its most commonly used synonym. *Empirical evaluation* is a term used sporadically by some authors. Finally, a set of other miscellaneous terms (*experimental analysis*, *experimental evidence*, *experimental setting*, and *empirical data*) account for from 9 to 22% of all terms.

These data show that there has been no standard use of terminology between 1993 and 2002. This has a direct impact on the results of the search strategies that we considered, as it would be necessary to use an unmanageable number of terms to get a search with 100% sensitivity. At the same time, this would cause a considerable drop in precision, calling for unwarranted effort in the experiment selection stage. Note that the non-standardization of terminology is a big impediment to effective searching and, consequently, to systematic reviews.

Another important aspect that we wanted to analyse concerns how indicative the titles and abstracts of the articles are. We analysed how many authors have used the term *experiment* in the title.

In our gold standard set, we found that only 18% of the relevant articles used this term in the title. However, a large proportion (55.6%) of authors did use the term *experiment* in the abstract, even if it didn’t appear in the title. As a whole, this term was explicitly mentioned in either the title or the abstract of 73.6% gold standard experiments. The remainder (25.6%) used either one of the synonyms analysed in this paper or no reference term in either the title or in the abstract.

This analysis should lend strength to the recommendation encouraging experimenters conducting systematic reviews to adopt the practice of
searching a combination of title and abstract and not confining searches to the article title.

Analysing the use of terms in titles and abstracts against time (see Figure 3), we found that the situation has not improved over the years. Although standardization initiatives and recommendations have been emerging since 1998, [9, 11] in [5], they do not appear to be having much of an effect. This confirms the need for community efforts working towards this type of guidelines to be put forward and adopted.

![Figure 3. Ratio between articles with term experiment in title and abstract](image)

### 7. Importance of the search universe

Another point that requires analysis if we aim to maximize the number of relevant articles retrieved by a strategy is the search universe. As mentioned earlier, a question that influences the retrieval of relevant experiments for systematic reviews is which sources (journals, conferences, etc.) to search using the selected strategies. To minimize the selection effort, these searches should be run on sources of interest, composed of publications whose scope includes SE empirical studies.

Sometimes a search may be confined to highly relevant publications or publications with a particular impact factor, as applies, for example to the review [12] that we have used as a gold standard. We wanted to study whether it is right to set a constraint of this type or whether such a limitation omits a sizeable number of relevant articles. To analyse this point, we compared the gold standard used in this paper against two systematic reviews, [6] and [1], that did not place any limitation on the search based on the quality of sources.

The first review [6] looked at 20 experiments in the testing area over a period of 25 years. Of these, 16 belonged to the 1993-2002 period and only 8 were part of the search universe of the gold standard. Therefore, 50% of relevant experiments for this systematic review were published in other journals or conferences.

In the second case [1], the review included a total of 30 experiments in the field of requirements elicitation, of which 10 belonged to the 1993-2002 period and reported controlled experiments. However, none of them were found in the search universe used as a gold standard. In this case, most of the experiments were published in journals and conferences such as the Journal of Economic Psychology, Journal of Management Information Systems, Knowledge Acquisition, Human Factors, Expert Systems, Information Systems Research, Journal of Experimental Psychology: Applied, International Conference on Requirements Engineering, and International Conference on Automated Software Engineering.

If we assume that there are at least 121 relevant articles for the 1993-2002 time period —the sum of the 18 articles covered by the above two reviews but omitted by the gold standard, plus the 103 articles belonging to the gold standard—, at least 15% of these relevant articles (represented by the 18 articles) are to be found in other sources. But we know that the searches [6] and [1] aimed to detect controlled experiments in the fields of testing and requirements elicitation, respectively. Additionally, one of the reviews [1] at least applied exclusion criteria: only individual elicitation techniques were considered and empirical studies that reported the use of just one elicitation technique or that compared an elicitation technique with some traditional way of doing things were omitted. Therefore, we can expect there to be more SE experiments (than the 18 considered) outside the universe used in the gold standard.

Consequently, the set of relevant articles that are overlooked if only publications of repute are used will be greater than 15% of the universe of relevant articles.

This leads us to make the following recommendations:

- If a systematic review (and the evidence it is to produce) is to be exhaustive and should find all existing experiments in certain domains, the search should not be limited to publications of repute. It appears that if the search is confined in this manner, a sizeable number of relevant experiments will not be found.

- Depending on the specific topic or technology of interest (such as elicitation techniques), publications from other areas should be explored (in the case of elicitation, areas like information systems, economics, artificial intelligence, quality, etc.).

### 8. Conclusions

We find that optimizing a search strategy for use to retrieve relevant SE experiments is not a
straightforward task. One way of increasing sensitivity is to include a lot of terms related to experiment, but an important problem that has to be tackled is the loss of precision as sensitivity increases. In other fields searches can, thanks to the discipline’s maturity, use very specific terms to maximize a strategy’s sensitivity up to 100%, but this approach does not manage to produce high-sensitivity searches in EBSE. Supplementing the term experiment with closely related terms (e.g. experimental comparison and experimental study, etc.), we can get better scores for sensitivity without any loss in precision. However, adding terms indiscriminately in an effort to maximize search sensitivity would lead to a drop in precision.

Notice that although these results are based only on SCOPUS, we have validated the strategies in other repositories (like IEEEXplore and ACM DL) and they yield similar results.

With respect to the terms used in searches for EBSE systematic reviews, the term experiment has very acceptable scores as a search strategy. Therefore, it is a term that will produce good results if used on its own. This does not apply to other synonyms of experiment. The analysis of the term experiment revealed that it is the term used predominantly, but authors do not use it scrupulously in the titles of their publications. Therefore, searches should be applied to both fields and not confined to just the title. This situation highlights the need to establish guidelines within the SE community instructing authors to use more indicative titles and abstracts. Also, our study alerts experimenters to the negative side of confining searches for systematic reviews to sources of repute or with high impact factors, as a considerable number of relevant articles could slip through the net. For some SE topics, the search should even explore publications in related fields like information systems, psychology, economics, quality, artificial intelligence, etc.

Regarding the limitations of our study, we only considered SE experiments; other types of empirical studies were not considered (i.e. case studies). On the other hand, the part of the search related to the topic has been left for the second phase of our research. Finally, the search universe is only a small subset of all SE-related publications; therefore further research is needed to extrapolate the strategies’ properties to other universes.

9. Acknowledgements

We would like to thank Prof. Natalia Juristo for her valuable collaboration in this research. Thanks also go to Dag Sjøberg and Jo Hannay, of Simula Research Laboratory, for providing the references to the 103 articles that were used to formalize the gold standard used in this research.

10. References