Search Engine Overlaps : Do they agree or disagree?

John Bailey, Cheng Zhang and David Budgen
University of Durham
Department of Computer Science
South Road DH1 3LE
United Kingdom
david.budgen@durham.ac.uk

Mark Turner
University of Keele
School of Computing and Mathematics
Staffordshire ST5 5BG United Kingdom
m.turner@cs.keele.ac.uk

Stuart Charters
Applied Computing Group
Lincoln University,
Lincoln 7647
New Zealand
charters@lincoln.ac.nz

Abstract

Context: *Secondary studies, such as systematic literature reviews and mapping studies, are an essential element of the evidence-based paradigm. A critical part of the review process is the identification of all relevant research. As such, any researcher intending to conduct a secondary review should be aware of the strengths and weakness of the search engines available.*
Objectives: *Analyse the overlap between search engine results for software engineering studies.*
Methods: *Three independent studies were conducted to evaluate the overlap between multiple search engines for different search areas.*
Results: *The findings indicate that very little overlap was found between the search engines.*
Conclusions of the study: *To complete a systematic review, researchers must use multiple search terms and search engines. The lack of overlap might also be caused by inconsistent keyword selection amongst authors.*

1 Introduction

The goal of the evidence based paradigm is to ensure that important strategical decisions are based on the most reliable technical evidence available. In the context of medicine, Sackett gives the definition " integrating individual clinical expertise with the best available external clinical evidence from systematic research" [8]. Evidence based medicine has become established practice within the UK and many other countries. It helps practitioners and policy makers to base their decisions on proven clinical evidence gathered from objective primary studies, rather than expert opinion alone. Since practioners and policy makers have neither the time nor resources to review every primary study, secondary studies are required. Secondary studies systematically and without bias aggregate and synthesise evidence from many related primary studies. Therefore, one of the key stages in a secondary study is the systematic review of literature.

1.1 Systematic Reviews

Systematic literature reviews aim to remove bias from the literature reviewing process [7]. This is achieved by formalising the search process itself as part of the review protocol. Non-systematic literature reviews risk being prejudiced by a researcher's own preferences and bias. By implementing a rigorous and systematic review and publishing how it was achieved, the credibility of the conclusions drawn are greatly increased [3].

1.2 Review Process

There are five main stages in a systematic review process, and each one is documented in the protocol both for transparency and also to aid future replication.

Searching Researchers develop a strategy for the systematic identification of potentially relevant studies.

Second International Workshop on Realising Evidence-Based Software Engineering (REBSE'07)
0-7695-2962-3/07 $20.00 © 2007 IEEE

Searching involves identifying search engines, keywords and sources of information.

Screening Once a list of candidate studies has been collected inclusion and exclusion criteria are applied. These criteria are established from the review question and are applied to the titles, abstracts and full texts and candidate studies. For example, a review might only include those studies carried out after a certain date.

Data-extraction This involves assessing the quality and extracting data from the studies after applying the exclusion criteria.

Data Analysis or Synthesis the development of a framework for data analysis and identification of key themes.

Reporting and dissemination presentation of the review findings.

The first stage of the process is perhaps the most crucial. Evans [3] observes that one of the problems of identifying studies within education is that "the databases that serve education are far less developed than those found in medical or health care", and this applies equally welll to computing. Without comprehensive databases and efficient search engines the identification of potentially relevant studies is limited. Our work on systematic literature reviews and mapping studies in the domain of software engineering has led us to believe that there is very little overlap between the various search engines within software engineering. If there is little overlap between the search results then there are two possible explanations: either the databases index different journals or databases, or the search engines are incomplete.

This paper analyses the overlap between the search engine results obtained while performing three independent software engineering studies:

- Technology Acceptance Model Study (TAM study)
- Evidence based OO Software Design Study (OO Design study)
- Software Design Patterns Study (Patterns study)

The first of these was a complete systematic literature review, while the second and third were conducted as 'mapping studies' aiming to assess the scope of available primary studies. A mapping study consists of the first number of steps of a review. Each study was conducted independently by different researchers and the overlap's occuring between the search engines will now be analysed.

2 The Three Studies

There were implementational differences between the TAM study and the other two studies.

2.1 OO Design and Software Patterns Studies

The OO design and software patterns were two distinct studies conducted independently by two separate researchers. The goal of each study was to map the relevant literature in each subject area. After an initial search and examination of the results returned, the best three search terms were selected. The exact search terms for the OO design study were as follows:

1. "object oriented" design "empirical evidence"
2. OO empirical design
3. "software design" OO experimental

while for the software patterns they were:

software design patterns experiment

2. software patterns empirical
3. software design patterns study

These six search terms were entered into the following six search engines:

- ACM Digital Library (A) - ACM journals, newsletter articles and conference proceedings.
- IEEE Xplore (I) - all IEEE online publications.
- Google Scholar (G) - precise details for Google Scholar are not published.
- CiteSeer (C) - Autonomous citation indexing [5]
- ScienceDirect (S) - Elsevier Reference Works
- Web of Science (W) - searches "8,700 of the most prestigious, high impact research journals in the world."

Results from each individual search will be identified using the number associated with the search terms used and the letter associated with the search engine used. For example 1G represents the results from the first search term (in the OO design study) using Google Scholar.

2.1.1 Limitations to Searching

As Table 1 shows, the number of results returned varies considerably between search engines. Google Scholar consistantly returned the most results whereas Science Direct returned the fewest results. ACM, ScienceDirect and Web of Science all required membership and were accessed using an institutional licence. Another limitation to the analysis of overlap were the artificial limits on search results imposed by search engines,. For example, Google Scholar

Second International Workshop on Realising Evidence-Based Software Engineering (REBSE'07)
0-7695-2962-3/07 $20.00 © 2007 IEEE

	1	2	3	4	5	6
A	28,261	32,597	8,136	122,353	34,905	126,678
I	7	21	0	124	84	241
G	3,030	77,800	6,180	260,000	207,000	860,000
C	26	35	37	118	184	413
S	4	8	1	31	34	97
W	11	26	1	31	81	830

Table 1. Number of results returned by search engines for each search

limits the number of results displayed for every search to 1000. ACM returns only 200 results, IEEE Xplore displays up to 100 and CiteSeer very often limits the number of results to 500 because of heavy system load. However, for the OO and Software Patterns Design studies this was not an issue as CiteSeer did not find over 500 results for any of the six searches.

2.1.2 Data Extraction and Analysis

The extraction of study titles from the search results was semi-automated. Each HTML page from the search results was pasted into a word processor. This automatically converted each HTML page into a rich text document. By maintaining the HTML presentational formatting it was possible to select and remove all text with specific formatting. For example, removing all green text of font size 14. If necessary, regular expressions were used to remove any leftover erroneous text. Once the study titles for each search was produced, the task of comparing the results was simple. An automated script was created that compared the study titles of each set of results. Study titles were normalised by removing all white space and non-alphanumeric characters. A study title was considered to match another if the normalised strings were exactly equal or if one title was a subset of another.

2.2 The Technology Acceptance Model Systematic Review

The Technology Acceptance Model (TAM) was proposed by Davis in [1] and [2] as an instrument to predict the likelihood of a new technology being adopted within an organisation. A potential problem is that the TAM assesses the *intention* to use a technology, and the actual use is rarely monitored. The aim of this study was to investigate whether the TAM is a reliable predictor of actual use.

2.2.1 Search Strategy

After conducting a pilot study, the following search strategy was adopted:

("technology acceptance model" AND (usage OR use OR empirical) AND (year >= 1989 AND year <=2006)

The main search was conducted in the following electronic databases: IEEE Xplore, ACM Digital library, Google scholar, CiteSeer library, ScienceDirect, and ISI Web of Science.

2.2.2 Limitations in Searching

One of the difficulties encountered was that the search engines each used their own interfaces, meaning it was impossible to use the search string in the exact form given in section 2.2.1. For example, IEEE Xplore enables complex logical combination, allowing the search string to be used exactly as outlined, whilst Google Scholar and the ACM digital library use interfaces organised around separate textboxes for each section of the search string.

A further problem was encountered due to the way that CiteSeer treats Boolean searches. By default the CiteSeer interface treats all searches as Boolean and therefore looks for the whole phrase entered. However, this search returned no results, and a non-Boolean search returned many irrelevant results. A search of the CiteSeer database using the Google interface returned around 45 results. Therefore, for this review the CiteSeer database was searched through the Google interface.

2.2.3 Data Extraction and Analysis

The extraction of study titles from the search results and matching of study titles between search engines was conducted as detailed in Section 2.1.2.

3 How well did the search engines agree?

3.1 OO Design Study and Software Patterns Study Overlap

The results this are presented in Tables 2 and 3.

3.2 TAM Study Results

Two research assistants searched each of the electronic databases detailed in section 2.2.1 and conducted an initial screening based on titles, keywords, and abstracts. A more detailed screening process was conducted on the remaining papers by applying a set of criteria developed from the research questions. This resulted in 71 papers being included in the final systematic review, with four papers being discovered by following references in the included papers. The following examines the overlap in the

x	1A	2A	3A	1I	2I	3I	1G	2G	3G	1C	2C	3C	1S	2S	3S	1W	2W	3W
1A	x																	
2A	32	x																
3A	8	36	x															
1I	2	1	0	x														
2I	0	2	0	3	x													
3I	0	0	0	0	0	x												
1G	88	16	7	5	7	0	x											
2G	19	40	13	3	19	0	168	x										
3G	8	11	35	4	9	0	98	145	x									
1C	0	0	0	0	0	0	3	0	1	x								
2C	0	1	0	0	0	0	2	6	1	3	x							
3C	1	0	1	0	0	0	3	3	3	0	1	x						
1S	0	0	0	0	0	0	0	0	0	0	0	0	x					
2S	0	0	0	0	0	0	2	4	1	0	0	0	1	x				
3S	0	0	0	0	0	0	0	1	1	0	0	0	0	0	x			
1W	1	1	0	1	1	0	2	1	0	0	0	0	3	1	0	x		
2W	1	1	0	1	5	0	7	15	6	0	0	0	1	7	0	5	x	
3W	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	0	x

Table 2. Overlap for OO design study

search engine results for the 71 included papers.

Google Scholar In total, 39 of the 71 included papers were found via a search of Google Scholar. The following lists the number of papers that were found in Google Scholar that were also found in another of the five search engines.

- ACM: 2
- CiteSeer: 2
- IEEE Xplore: 6
- Science Direct: 4
- Web of Science: 9

Several papers that were found in Google Scholar were actually found in multiple databases:

- ACM and Web of Science: 2
- CiteSeer and ACM: 1
- CiteSeer and IEEE Xplore: 1
- CiteSeer and Web of Science: 1
- IEEE Xplore and Web of Science: 1
- Science Direct and Web of Science: 14

Web of Science. 43 of the 71 included papers were found through searching the Web of Science database, 12 of which were *only* found in Web of Science and were not listed in the results of any other database. In addition, four were found in both the Web of Science database and also via a search of Science Direct. Many papers that were found in the Web of Science database were also found in Google Scholar (see above).

CiteSeer. One extra paper was found in the CiteSeer database that was not found in any other database, out of a total of six relevant papers highlighted by the search. All of the other papers found by CiteSeer were also found by Google Scholar, with some also being found by IEEE Xplore, Web of Science or the ACM digital library (see above).

Science Direct. Seven papers were found through a search of Science Direct that did not appear to be found by searching other databases, with 29 relevant papers being found in total.

IEEE Xplore. Thirteen papers were found in the IEEE Xplore database, five of which were not discovered in other databases. All of the other eight papers were found in at least Google Scholar, but also Web of Science or CiteSeer (see above).

ACM Digital Library. All of the five included papers that were found in the ACM digital library were also found in Google Scholar or other databases (see above).

Second International Workshop on Realising Evidence-Based Software Engineering (REBSE'07)
0-7695-2962-3/07 $20.00

	4A	5A	6A	4I	5I	6I	4G	5G	6G	4C	5C	6C	4S	5S	6S	4W	5W	6W
4A	x																	
5A	29	x																
6A	56	42	x															
4I	3	3	3	x														
5I	0	2	1	4	x													
6I	3	3	5	12	5	x												
4G	32	30	23	11	5	15	x											
5G	8	53	21	5	18	8	136	x										
6G	26	27	66	8	9	26	274	230	x									
4C	1	0	0	0	0	0	5	1	3	x								
5C	0	1	0	0	0	0	2	5	3	23	x							
6C	2	0	1	0	0	0	6	3	21	52	43	x						
4S	0	0	0	1	0	0	1	0	2	0	0	0	x					
5S	0	0	0	0	0	0	1	4	1	0	1	0	1	x				
6S	0	0	0	0	0	0	1	0	6	0	0	0	10	3	x			
4W	3	0	1	6	1	0	6	2	3	1	0	0	4	1	1	x		
5W	1	1	1	1	6	1	7	17	7	0	1	0	0	16	2	5	x	
6W	13	4	11	9	3	9	26	17	81	1	0	3	8	5	46	31	30	x

Table 3. Overlap for software patterns study

4 Discussion

4.1 Search Results Size

The number of papers identified by each search varied greatly depending on the search engine and the search terms used. In the OO design study the ACM had the greatest recall for the 1st and 3rd search while Google Scholar returned more papers in the 2nd search. However, Google Scholar had by far the greatest recall, followed by ACM in the software patterns study.

In total, search engines returned many more papers in the software pattern study (1,613,204 results) than in the OO design study (156,181 results). From the number of papers found it is clear that the search terms used in the software patterns study were quite general. Search term 6 was especially so, as the search terms included 'study' and 'software' which could apply across topics and so offered by far the largest recall, especially in the multidisciplinary Google Scholar. The number of results returned for the TAM study were not recorded. However, of the 71 verified papers used in the study the majority (43) of papers were found using the Web of Science with Google Scholar finding slightly fewer (39) papers.

4.2 Overlap

Table 2 shows there was relatively little overlap between the search results in the OO design study. Most overlap occurred between searches using the same search engine. One interesting exception was the overlap between 2W and 2S. Out of the 8 papers found by 2S 7 (87.5%) were also found by 2W. The only other overlap over 50% was between 2W and 2G (58%). In both the OO design and software patterns studies, the greatest number of papers shared between search engines was found between the ACM and Google Scholar searches. However, this was partly due to the larger recall and also that Google Scholar appears to have access to the ACM digital library. For the OO design study intra-search engine overlap was very low. The highest overlap was 18% between the 2A and 3A searches followed by 16.8% between 1G and 2G (16.8%) and finally 2G and 3G having 14.5% of papers in common. The level of intra-search engine overlap was slightly higher in the software patterns study. 27.5% of papers found in 6G were also found in 4G and 23% of 6G papers were also found in 5G.

One explanation for this lack of agreement between the search engines is the lack of standardisation within the software community for the creation of meta-data. This is in part caused by the lack of any up-to-date centrally maintained keyword repository. For example, neither "design patterns" nor "empirical" appear in either the ACM or IEEE taxonomy of software engineering. As such authors are forced to generate their own keywords and so variation is bound to occur. Another explanation might be that authors do not consider that their work might be used in a secondary study. As such, classification and providing appropriate meta-data is not a priority.

In the TAM study, the greatest degree of overlap occurred between Google Scholar, Web of Science, and Science Direct, with 14 of the 71 papers being found in all three databases. These 14 papers represented 35.9% of the Google Scholar results (14/39), 32.6% of the Web of Science results (14/43), and 48.3% of the Science Direct results. The only search engines that did not find any studies that were not found in any other databases were Google Scholar and ACM. Similar to the OO design and software patterns studies, all of the papers found in the ACM digital library were also found in Google Scholar, due to Google Scholar indexing the ACM library. An interesting result is that Google Scholar also has access to IEEE Xplore, but five papers were found by searching through the IEEE Xplore interface that were not found by Google Scholar.

4.3 Experiences

The overriding impression gained from carrying out these studies is that search engines are not designed with systematic literature reviews in mind. They appear to be geared towards a model of an individual purchasing a service rather than mapping current literature.

Inconsistent user interfaces meant that extracting and cleaning up the data from the search results was time consuming and at times awkward.

Limitations placed on the number of results displayed did not allow the full search results to be analysed.

Limitations placed on the number of results displayed on each page meant that copying and pasting paper details was very time consuming.

Another obstacle found when conducting the search for literature was the number of times some search engines were either unavailable because they were busy or crashed.

4.4 Conclusions

This paper has highlighted the lack of overlap and presentational inconsistencies between several important search engines. Therefore, researchers cannot rely on any one search engine, rather they must apply their search across multiple search engines to ensure getting a significant coverage of the literature. However, attempts to use multiple search engines are hindered by inconsistencies in the way search terms are collected and handled. Differences and limitations in the way that search engines present results also cause significant difficulties for any systematic literature survey or mapping study. During the course of this work several search engines have modified their interfaces and the method in which they accept search terms, however there were no modifications to the searches published. This becomes an issue when replication studies are conducted. Another complication is the lack of standardisation of terms used within software engineering. If search engines are to compensate for this, then up-to-date keywords and synonyms must be maintained to ensure relevent papers are not missed.

As a result of this study the following recommendations are made:

- Researchers should familiarise themselves with how each search engine handles search terms. For example, in Google Scholar adding quotation marks allows searching exact match for a whole phrase.
- To avoid redundant searches, researchers should first plan which terms will be applied to which search engines and once completed, the results and timestamp are recorded. This plan can also help spot errors such as spelling mistakes.
- Finally, due to the apparent fragility of some search engines a patient and opportunistic approach must be adopted. If one search engine is temporarily unavailable, then the plan can be used to ensure researchers make the best use of this time.

4.5 Acknowledgements

We thank the members of the Evidence-based Software Engineering (EBSE) project for their discussions and for contributing to the ideas in this paper. This work was funded in part by an award from the Engineering and Physical Sciences Research Council.

References

[1] F. Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.*, 13(3):319–340, 1989.

[2] F. D. Davis, R. P. Bagozzi, and P. R. Warshaw. User acceptance of computer technology: a comparison of two theoretical models. *Manage. Sci.*, 35(8):982–1003, 1989.

[3] J. Evans and P. Benefield. Systematic reviews of educational research: Does the medical model fit? *British Educational Research Journal*, 27(5):527–541, 2001.

[4] B. Kitchenham. Procedures for performing systematic reviews. Technical report, Keele University , National ICT Australia Ltd, 2004.

[5] S. Lawrence, L. G. C, and K. Bollacker. 32(6):67–71, 1999.

[6] C. D. Mulrow. Systematic reviews: Rationale for systematic reviews. *British Medical Journal*, pages 597–599, 1994.

[7] M. Petticrew and H. Roberts. *Systematic Reviews in the Social Sciences: A Practical Guide*. Blackwell Publishing, 2006.

[8] D. L. Sackett, W. M. C. Rosenberg, J. A. M. Gray, R. B. Haynes, and W. S. Richardson. Evidence based medicine: what it is and what it isn't. *British Medical Journal*, pages 71–72, 1996.