Ms. Patricia E. Kirkwood, University of Arkansas

Patricia Kirkwood is the Engineering and Mathematics Librarian serving more than 3,000 students in these disciplines. She has provided reference and instruction services in every STEM field. Her interest in finding appropriate ways to evaluate resources and services has resulted in studies using citation analysis, use information, interlibrary loan statistics, and publication patterns.
Researcher publication checklists: A quantitative method to compare traditional databases to Web search engines.

As specialized Web crawlers like CiteSeer\textsuperscript{X} and Quertle become more sophisticated, they gain recognition in user communities. Perhaps these Web tools do a better job of gathering references to conference publications and grey literature than do traditional databases. Librarians need to compare traditional databases with Web crawler generated databases when they make collection development decisions and as they choose what tools to highlight during instruction sessions. Traditional databases are often compared by reviewing what journals are indexed in each resource. This type of comparison is not applicable when evaluating Web search engines. First, there is no publication list for databases created by Web crawlers. Second, for engineering related search tools, journal title comparison does not include other important publications like conference articles, government documents, reports, standards, patents and grey literature. One method used to assess Web crawlers is to compare results of subject based searches to the same search in a traditional database. Reviewing the number of items retrieved and the appropriateness of a subset of the citations retrieved provides an assessment of how well a search engine performs. However this method does not show the depth and breadth of content found within a database. This paper explores a method to quantitatively compare the content gathered by Web crawlers to the content provided by traditional databases. A checklist is developed using researchers’ resumes. The researcher publication checklist methodology is evaluated and validated with a test project.

Introduction

Traditional library databases, like INSPEC, Compendex and Web of Science are in competition with Web applications like Google Scholar, Scirus, and CiteSeer\textsuperscript{X}. Though librarians and information literate users know the limits of specific resources, students tend to turn to the Web and use the first available resource. New researchers also tend to use what they are familiar with without evaluating alternatives. In this time of tight budgets and shrinking staff, librarians must make difficult decisions about what products and resources should be purchased and promoted. Librarians must also decide what resources to emphasize when they have the opportunity to get into the classroom. A recent article that examines the information seeking behavior of computer science and engineering faculty at the College of New Jersey states “A third major issue was the need for more subject-specific databases…. [1].” This library was unable to purchase additional databases to meet that expressed need and had to answer the following question: Is CiteSeer\textsuperscript{X} a reasonable substitution for a subscription database in the area of computer science? There is no proven method of assessing Web crawler generated databases, especially in direct comparison to traditional bibliographic databases.

A Web crawler database is a database created using a Web agent, often known as a spider or Web crawler. There are multiple types of Web crawlers and they generate various types of databases. Some are indexes. Others are digital libraries with search engines. Many have
attributes of both. The large Web crawler databases, such as Google, Google Scholar and Scirus seem ubiquitous and have proven their usefulness over time. Smaller specialized databases, such as CiteSeer and Quertle are becoming more prevalent. The methods used to quantitatively compare traditional databases do not allow quantitative comparison with Web crawler generated databases because the information about content inclusion for Web crawler databases is generally not available. To accurately compare traditional databases to Web crawler generated databases, publication specific comparison at the article or item title level is needed. This paper discusses the development of an article checklist that allows for a quantitative comparison of databases developed in starkly dissimilar ways.

**Literature Review**

Nisonger summarizes many methods for analyzing library services and resources. Calling these methods checklists, he details the history and use of checklists by the library community to validate collections, review library services, and evaluate databases. Checklists have been developed from many sources including faculty publications, course syllabi, and textbooks. Each source has been chosen to help evaluate the service or product with real data. Nisonger identifies two studies using the references from journal articles to evaluate database coverage. Neither are studies of engineering related resources. “Typically, a list of journal titles is checked against the vendor’s list of titles theoretically contained in the database [2].”

Grzeszkiewicz and Hawbaker research goes a step further. They determine if the full text of articles are available in full text aggregators as advertised. They use the table of contents from journals to create their article level checklist [3]. Newton and Jennlyn compare two traditional library databases with an article based checklist in 2010. A checklist is one of several methods used to compare two databases in medieval history. Given the importance of non-journal essays for this field, the researchers felt a list of journals covered is not a sufficient assessment of database coverage. They generate an item based checklist using Web of Science and book citations from references in dissertations identified through ProQuest Dissertation Abstracts [4].

2005 is a banner year for non-comparative reviews of web search engines. Brophy and Wleklinski both provide reviews of Google [5, 6]. Notess and Felter provid similar reviews of Google Scholar and Scirus [7, 8]. These are general reviews of specific resources and do not offer comparative data.

Several reviews of traditional science databases are available. In 1991, Hightower and Schwarwalder compare engineering databases several platforms with journal title checklists [9]. This is one of the earliest empirical comparisons of traditional databases. Soremark’s [10] uses the results of subject searches to compare of Medline and Embase. Jaguszewski [11] compares current awareness services with a journal checklist in 1995. Though Brier and Lebbin do not focus on science databases, their article detailing a method for assessing databases based on a library’s journal holdings addresses the need to provide analysis of databases based on local
experience. Their checklist is journal title based [12]. Structural biology databases are compared using author searches of 11 scientists in 2002 [13]. Salisbury and Noguera review bioengineering databases with a journal title checklist [14]. Zhou and Leydesdorff provide interesting insights into the issue of citation patterns using Web of Science and China Scientific and Technical Papers and Citations Database [16]. Nagaraja uses a journal checklist to compare traditional databases in pharmacy in 2008[18].

Comparisons between traditional science and engineering databases and Web crawler produced database are rare. In 2007, the Web crawler database Google Scholar is compared to CAS (Chemical Abstract Service) with subject and author searching [15]. Meier offers a comparison between traditional engineering databases and Google Scholar using subject searching [17].

A comparison between tradition science and engineering databases and Web crawler produced databases coverage at the article level has not been published. In order to provide this type of analysis, a checklist must be developed with a representative sample of publications at the item level. The publication checklist should include items of interest to the community that use the database to locate information for their research needs. Following is one option for developing an article or item level checklist, using local researchers’ resumes. The methodology is testing using a small checklist with several traditional databases and the CiteSeerX Web crawler generated database. Details of that test are included later in the paper.

A Proposed Methodology for the review of databases using a publication checklist.

1. Develop a checklist based on items published by the members of the community that will use the databases. Harvest items from their publication lists, vitas or resumes for the checklist. In most engineering disciplines the checklist will include articles published in conferences, scholarly journals and trade magazines as well as technical reports, patents and other grey literature.

2. Organize the items in the checklist by publication type. Depending on the databases being reviewed specific categories of publications may be removed from the active checklist. Carefully document the decision to remove items and retain all items removed for later review.

3. Review the active checklist to determine if it is representative of the publications of the user community and provides appropriate items to complete the evaluation. Questions that may be asked:

- Are the items in the checklist published in venues used by and on topics of interest to target group? For example, if the checklist includes articles on the education of elementary students and the main target group is research scientists in national laboratories, it may be appropriate to remove these items from the checklist.
• Are the publications on the checklist representative of the type of materials covered in the databases being reviewed? A checklist of journal articles is inappropriate to review a database that focuses on standards or patents. However, if the purpose of the study is to review dissimilar products to determine where the publications of a specific population of researchers could be found, then items from multiple formats should be retained in the active checklist.

4. Examine the checklist to determine if it will provide sufficient quantitative information for statistical analysis. Questions that should be asked include:

• Is the checklist large enough to be to generate a useful comparison?
• Is the checklist of brief enough to provide the necessary comparison in a timely fashion?
• Does the checklist contain unintentional biases as a result of authors’ publication patterns?
• Does the checklist have a subject bias inappropriate for the study?
• Does the checklist show a preference to specific publication formats?
• Are specific journals, conferences, or publishers over represented in the checklist?

5. Determine if every publication type should be searched in every database. Note the reasons for each decision.

• To save time, local conferences, unpublished materials, and non-standard publications included in the checklist may be removed or coded as “unexpected” for a given database being reviewed.
• Websites, online government publications outside normal distribution channels, corporate information, statistical databases and similar non-standard resources can also be removed prior to the search or coded as “unexpected” in a given database.
• Web crawler generated databases may choose to limit documents included to items available online as full-text. If information about the online full-text status of items in the checklist is available it could be used to enhance the study as well as focus search efforts. When Web crawler databases limit to full-text items, items not available on the Web in full-text could be coded as “unexpected”. However, as discussed in more depth later, full-text status is difficult to determine.

6. Use the checklist to find content within the databases.

• Code each item as found or not found.
- Items on the checklist that should not be expected in a given database may be recorded as “not expected”. This category may or may not be needed depending on the question being asked, the source of the checklist, and the nature of the databases being reviewed.

7. Review all items not found to determine if an alternative search strategy is needed. An item should not be coded as “not found” until at least two types of searches are performed and an expert searcher reviews the results.

- Items not found should be reviewed for typographical or spelling errors.

- Journal and series abbreviations (e.g. LNCS; Lecture Notes in Computer Science) should be expanded.

- Incomplete and incorrect citations need to be searched with correct information. Government publications and conference materials are especially problematic.

- Alternative publication possibilities should be considered. For example, conference articles may be published within a journal. The database entry may not include the conference information. The checklist citation may not include journal information. The author and title will match but the publication details differ. An expert should decide whether these are the same publication.

8. Review the results to determine if there is an unintended bias.

**Methodology Validation**

The proposed methodology was developed and modified using a test case. The test case was performed in 2011 and additional quality checks of the results completed in early 2012. The results of this test should not be considered as a complete and valid review of the databases involved. The methodology presented above is the final methodology developed after the validation study was completed.

The databases compared in the test study were INSPEC, Compendex, Web of Science, and CiteSeerX. INSPEC, Compendex, and Web of Science are traditional databases that contain citations to literature in computer science, information technology and other subject areas. INSPEC and Compendex index articles published in journals and conferences. Both also have records to some complete conferences. Web of Science indexes articles from journals with high citation rates. CiteSeer (original name) was developed in 1997 at the NEC Research Institute, Princeton, New Jersey. The service moved to the Pennsylvania State University's College of Information Sciences and Technology in 2003 and was renamed as CiteSeerX. It is currently used as a research platform by computer scientists interested in autonomous citation indexing, full-text indexing, Web agent effectiveness when harvesting documents, and autonomous metadata generation. In 2012 a new version, CiteSeerX Beta, was launched.
The Web agent for CiteSeerX locates full text papers posted on the Web. It is a focused crawler with a predetermined topical focus of computer science and information science. It does not crawl or index publishers’ Websites. Most of the documents included in the CiteSeerX database are journal papers, conference proceedings, thesis papers and books. CiteSeerX had 1.5 million documents in its database when our search was completed in 2011. But, according to Dr. Giles, [19] another 3.5 million items still needed to be processed and indexed at that time. Since this test study was completed the Web agent (citeseerxbot) used to create the database has been upgraded and revamped. More details about the currently available CiteSeerX Beta database can be found at: http://csxstatic.ist.psu.edu/about. The results of this methodology validation cannot be considered a formal review of the current CiteSeerX Beta database.

A checklist based on articles published by scientists from the Computer Science and Computer Engineering department of the University of Arkansas was combined with the publication lists from NEC Research Institute authors. Dr. Giles’ publications were added to the checklist as well. The checklist included articles from conferences, journals and trade magazines. The publication checklist also included books, presentations, patents, and government reports. The checklist was limited to items that were published between 2000 and 2010. Some articles had 2010 publication dates but were excluded when there was no indication that publication occurred by the time searching was completed. A few publications in subject areas not covered by these databases were removed from the checklist.

All journal articles were searched for in the traditional databases and CiteSeerX. Conference articles were not searched in Web of Science as the Conference Proceedings for Web of Science is not available at this institution. Other types of resources, such as book chapters, technical reports and government documents were searched in CiteSeerX but not in traditional databases. Patents were excluded from the search as they were not, according to documentation, included in any of the databases being reviewed.

Each item not found as an exact match with a simple title or author search was searched using additional citation elements like conference title and/or location. Closely related items, such as author/title matches with differences in publication details were carefully reviewed. Items listed in a database as published in an alternate format (journal, conference, or book) but with the same publication year were considered equivalent publications and coded as “found.”

An effort was made to determine if items in badly crafted citations could be located using common solutions. When an item could be better identified, it was then searched in each database with the new information. Items still not located in any database were tracked down using other resources such as WorldCat and Google to determine if they were published in traditional venues or available on the Web.

During this last phase all items not found in a given database were reviewed. A judgment was made as to whether, based on the database description, the item would be “expected” to appear in
that database. For example: publications at regional ASEE conference would not be expected in 
INSPEC or Compendex. The general description of the subject content of database was 
consulted to determine if a published item could be “expected.” The list of journals and 
conferences indexed in a specific database was not consulted.

It was impossible to code items as expected or not expected with CiteSeerX. There was not, at 
the time, a description about the types of materials collected by the Web crawler only a general 
statement about what subject content might be expected in the CiteSeerX database. Dr. Giles, the 
director of the CiteSeerX project, was contacted and his email responses were used to better 
understand the autonomous process used to develop the research platform’s database. The most 
important inclusion parameter was that the full-text of an item must be placed on a server or in a 
repository that allowed CiteSeerX autonomous crawler to harvest the content. The item is then 
subjected to an autonomous review based on subject and other unspecified parameters [19].

The need for the author to post full text in an appropriate venue makes it difficult to determine 
when an item can be “expected” in the CiteSeerX database. Searching the Web to determine if 
full-text is available would not provide an accurate assessment for inclusion in the CiteSeerX 
database because items can be placed on the Web so that CiteSeerX can harvest them but the full-
text is not visible in a general Web search. Conversely, an item placed on the Web in full-text 
can be coded in a manner that prevents a given, or any, Web agent to harvest the items full-text 
or an item’s citation information.

The inclusion of a “not expected” category can explain why a database provides different 
coverage. It can be used to distinguish dissimilar products. However, using it as a decision point 
may not be appropriate. Just because something is “not expected” in a database, doesn’t mean it 
is not an important item to find. The question being asked may require careful analysis of what 
can be “expected” within a database. For example: If researchers need to find articles published 
in the ACM (Association of Computing Machinery) journals and copyright restrictions do not 
allow authors to post the articles so they are routinely harvested by CiteSeerX then the CiteSeerX 
database may not be the best database to search. If, on the other hand, local conference papers 
are important in a research effort, than the INSPEC database cannot be the only database used.

**Validation Test Study Results**

Over 500 journal and conference articles were reviewed during the test study conducted in 2011. 
The article checklists developed for the test study reviewed in three traditional databases and the 
CiteSeerX database. The results of this test study should not be considered valid for the currently 
available CiteSeerX Beta database.

The methodology was modified based on this test to include a step that codes for “expected 
items”. Determining if this step is appropriate for a comparison study would depend on the 
questions being asked. As this study was a test to see if a publication checklist is a valid
quantitative method for this kind of database comparison, both sets of results were retained and reviewed.

Table 1 provides a summary of the test study results without coding for “not expected” items.

<table>
<thead>
<tr>
<th>Articles</th>
<th>INSPEC</th>
<th>Compendex</th>
<th>Web of Science</th>
<th>CiteSeerX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journal</td>
<td>136</td>
<td>75 55%</td>
<td>62 46%</td>
<td>60 44%</td>
</tr>
<tr>
<td>Conference</td>
<td>393</td>
<td>160 41%</td>
<td>138 35%</td>
<td>N/A N/A</td>
</tr>
<tr>
<td>Totals</td>
<td>529</td>
<td>235 44%</td>
<td>202 38%</td>
<td>63 44%</td>
</tr>
</tbody>
</table>

Table 2 provides the results of the search in traditional library databases with a filtered publication checklist.

<table>
<thead>
<tr>
<th>Expected Articles</th>
<th>INSPEC</th>
<th>Compendex</th>
<th>Web of Science</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journals</td>
<td>85</td>
<td>75 88%</td>
<td>62 73%</td>
</tr>
<tr>
<td>Conferences</td>
<td>211</td>
<td>160 76%</td>
<td>138 65%</td>
</tr>
<tr>
<td>Total Expected</td>
<td>296</td>
<td>235 79%</td>
<td>202 68%</td>
</tr>
</tbody>
</table>

Less than half of the journal and conference articles on the author’s publication lists could be expected in the traditional databases. The items that would most likely not appear in traditional databases were local conferences papers, position papers, workshops, and patents. This information, in itself, could be a useful study result. If a small institution faculty does not publish in venues that are indexed in traditional databases then the librarian could poll faculty to determine if they use traditional databases to locate resources for research and teaching. An analysis of the materials cited by faculty could also be done to provide an additional decision point about database purchases and instruction.

**Discussion of Possible Bias**

Checking for biases that might occur when using a researcher publication checklist is critical. The publications from each researcher were reviewed to determine if individual publication patterns impact results. Table 3 shows how critical this review can be. The publications of Dr. Giles, current director of the CiteSeerX project, were included in this test study as a quality
The CiteSeerX database included 12 of the 16 articles he co-authored in the noted time period. The 75% found rate is much higher than the average (13%) of found articles.

Table 3: Author results: C. Giles; director of CiteSeerX project

<table>
<thead>
<tr>
<th>Articles</th>
<th>Articles Found</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>INSPEC</td>
</tr>
<tr>
<td>Journals</td>
<td>4</td>
</tr>
<tr>
<td>Conferences</td>
<td>12</td>
</tr>
<tr>
<td>Total</td>
<td>16</td>
</tr>
</tbody>
</table>

Similar result occurred with the publications of Dr. S. Gauch, a researcher involved with the CiteSeerX project. Dr. S. Gauch is also a faculty member at the University of Arkansas. Eleven out of 33 articles on her publication list were included in CiteSeerX database (see Table 4). This was 3 times higher than the average. Though this frequency could be within the normal distribution curve, given the results for Dr. Giles noted above, it indicates a possible bias against CiteSeerX with the checklist developed.

Table 4: Author: Gauch S, CiteSeerX researcher

<table>
<thead>
<tr>
<th>Articles</th>
<th>Articles Found</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>INSPEC</td>
</tr>
<tr>
<td>Journals</td>
<td>7</td>
</tr>
<tr>
<td>Conferences</td>
<td>27</td>
</tr>
<tr>
<td>Total</td>
<td>33</td>
</tr>
</tbody>
</table>

The test study checklist was developed using the publications of two organizations, the University of Arkansas and NEC Research Institute. Do the authors from NEC Research Institute (12.5%, N=128 found in CiteSeerX) and the University of Arkansas routinely post their publications so that CiteSeerX can harvest them? Both Giles and Gauch are aware of the parameter for inclusion into CiteSeerX and they may be more likely to post materials appropriately. If a publication checklist were developed using the resumes of faculty from other institutions would the results change? If the publications included in a checklist are not routinely placed on the Web where the crawler can find them, does that introduce a bias into the study of a Web crawler generated database? Or is that a valid review? These are questions that might have different answers depending on the purpose of the comparison study. In this case, I would recommend the checklist developed for this test study be revised to include at least 100 items known to be posted in a manner that would allow CiteSeerX to find them. At that point the question “Is CiteSeerX of equal content to the traditional databases?” could be legitimately answered.
Another source of bias that should be explored with any publication checklist is the possibility of bias introduced by a specific researcher’s publication pattern. Dr. Giles co-authored 5 papers in the 18th ACM Conference on Information and Knowledge Management. This pattern is not an unusual for a research group, especially if the conference is held at an accessible venue and exciting results from a research group need to be reported. Unfortunately, for the INSPEC retrieval rate, this conference content was not indexed. The inclusion error will be reported to INSPEC; however, this publication pattern and the assumed error on the part of INSPEC lowered the retrieval rate by 50% for conference articles in INSPEC.

The possible biases noted by this test study show how critical it is to review the publication checklist for patterns that might skew results. A publication checklist developed in the manner described is not a random sample. Therefore it must be reviewed for unintentional biases that must be considered when evaluating the results.

**Conclusion**

The use of a publication checklist to evaluate database coverage, including the coverage of Web crawler generated databases, is viable. A publication checklist can be used to determine if appropriate citations are in databases that are created using disparate means. A study using publication checklists also points out weaknesses in any given database and helps determine which of several databases might be the most useful within a specific field. The test publication checklist of 500 articles clearly shows that INSPEC, Compendex and Web of Science have similar retrieval rates for literature in the field of computer science and computer engineering and that the version of CiteSeer\textsuperscript{X} available at the time of the study did not cover the publications as completely. More in-depth analysis of the results could show how much duplication exists in these databases and whether all databases are needed as tools within a given user community.

The development of a publication checklist for a Web crawler generated databases has some inherent difficulties. Web crawler generated databases are constantly shifting, and what is included or excluded is not always clearly stated. The time taken to do a review of this depth may be problematic because a Web crawler generated database may change dramatically in a short period of time. However, Web crawler databases should still be rigorously reviewed before they are marketed by the library as quality resources. At the very least, a comparison at the article or item title level will allow librarians discuss with the user the pros and cons of each database. A publication checklist is a viable method to use for this type of study.

Additional benefits may occur when the publication checklist is developed using faculty resumes. For example, the results of this test study may be presented to the faculty and university administration as one of the reasons for the development of a local online repository for faculty publications or to encourage faculty to deposit articles with open access repositories within their disciplines. The information could also be used as a starting point for a discussion on open access and other copyright related issues.
Future Research

It was hoped that the checklist developed for this study would be able to answer the question “Can CiteSeerX be a substitute for Compendex, INSPEC, or Web of Science when purchasing or teaching about article databases?” Although the test study showed the publication checklists method was viable, the study could not answer the question definitively for two reasons. First, CiteSeerX was in a transitional period when the study was conducted. A large number of articles harvested had not been processed. Because those items may have been processed since, the results of the test study are no longer applicable. Second, to avoid any possibility of bias, it would be wise to create a more robust checklist by adding titles from resources known to be posted on the Web in a manner that allows the Web crawler for this product to find them. Combining the results of the test study with a second checklist should provide sufficient data to reach a better informed conclusion about the use of CiteSeerX Beta database as a tool for students and researchers.

The checklist developed for the comparison study showed many of the publications of researchers are not indexed by traditional databases. Most of these items would not have been expected in traditional databases as they were not published in journals or conferences. However, some were published in traditional resources like books or as conference on the Web. A next step might be to check all the items not found in the traditional database and/or CiteSeerX against the more general Web crawler databases, such as Google Scholar and Scirus.

Are the items not found in any resource important enough to merit inclusion in some finding aid? Are non-traditional and “unexpected” items of enough importance that some type of tool, other than a general Web search, should be developed to harvest and organize these items? Should traditional database vendors expand on what they index? Are the Web crawler databases doing enough to help engineers find publications from local venues (e.g. regional conference publications, workshops), government agencies (e.g. technical reports, policy papers) and other grey literature? These are all questions a publication checklist can be used to help answer.

Currently librarians use citation analysis, journal checklists, and subject retrieval results to evaluate databases. Publication checklists are another method to quantitatively evaluate the depth of coverage in traditional and Web crawler created databases.

References


