

Anti-Plagiarism Tools: Scirus v. Google

Melissa Holmberg

Mark McCullough

Introduction

This project is an extension of an earlier research project in which the Google search engine was tested as a tool for detecting plagiarism in online master's theses. The goal of that study was to determine whether Google might offer an inexpensive and efficient alternative to commercial plagiarism detection software and services. The authors selected phrases from electronic master's theses and searched for matching phrases in the Google search engine. The authors sought to find out whether thesis advisors or thesis committee members might be able to use Google as an effective and efficient tool for detecting plagiarism. The earlier study revealed that searching selected texts from Master's theses for 10 minute periods in Google yielded potential occurrences of plagiarism in 27% of the randomly selected theses from various academic disciplines (McCullough and Holmberg). Results from the initial project showed the highest incidence of potential occurrences of plagiarism in science and technology theses (McCullough and Holmberg). Thus, theses used for the current study were randomly selected from a set of science and engineering theses completed in 2003 and available electronically.

While the current study applied the same methodology as the earlier study, the primary goal this time was to compare two search engines as tools for detecting plagiarism. A secondary goal was to further test the method of the earlier study— i.e. determine whether search engines are effective and efficient tools that can be used by theses advisors and other faculty for detecting plagiarism in graduate theses. Several recent articles have profiled extreme cases of plagiarism, suggesting the need for methods

to detect and prevent plagiarism (e.g. Bartlett & Smallwood). The two search engines tested were Google and Scirus—a search engine designated “for scientific information only.” Our hypothesis was that more potential occurrences of plagiarism would be found using the Scirus search engine since it purports to include only scientific resources.

Method

The theses used for the study were extracted from the WorldCat database. Searching the WorldCat database, the authors randomly selected 376, or 20%, of the science and engineering theses published in 2003. Seven theses were excluded: one was a duplicate record and the other six were inaccessible. The sample included ten institutions located in the United States. The authors searched independently for selected phrases from each thesis in one of the search engines. One author searched Google, and the other searched Scirus. Each searcher was allowed to search phrases of his or her choosing for ten minutes. If a match (defined as an undocumented phrase of seven) was found, the searcher stopped the clock to investigate the match. When multiple results were retrieved by the search engine, the searchers selected the first match. If it was not possible to open the first match, then the next one was selected. Many times, the searchers needed to extend the length of the matched phrase beyond seven words or by typing multiple phrases into the search engines. Pages for matches were printed, the time was recorded and matched phrases were highlighted in both the thesis and webpage printouts. Once each searcher had completed searching each thesis, all matched phrases were searched in the other search engine—i.e. matched phrases found in Google were

searched in Scirus and vice versa. The matched phrases were then analyzed. The names of institutions and authors were coded so that identities of both would not be revealed.

Results

Matches, or potential occurrences of plagiarism (POPs), were identified in 46 of the 68 theses searched (67.6%). Theses with matches came from six of the ten universities represented in the sample. Institution I had the largest number of theses in the sample (38) and the highest number of theses with potential occurrences of plagiarism (26). The average time it took searchers to detect a matched phrase was less than 5 minutes. Matches were found more quickly by the Scirus searcher.

(Figure 1)
POPs by Institution ID

Institution	POPs Found	Theses Checked	Percentage POPs
A	0	1	0%
B	9	10	90%
C	1	3	33.33%
D	0	1	0%
E	0	0	0%
F	1	1	100%
G	6	8	75%
H	2	3	66.67%
I	26	38	68.42%
J	0	3	0%
Total	45	68	66.67%

(Figure 2)
Average Time (in Minutes)

Scirus	3.79
Google	4.88
Average of Both	4.42

The Google searcher found matches in 34 of the 68 (50%) theses and the Scirus searcher found matches in 32 of the 68 (47.05 %) theses. Of the 46 theses that yielded

matched phrases, 26 were unique—i.e. detected by only one searcher. The Google searcher found matches in 14 theses in which the Scirus searcher found no matches. The Scirus searcher found matches in 12 theses in which the Google searcher found no matches. Matched phrases from the 26 unique theses were searched in the alternate search engine and 16 of these—eight Google and eight Scirus-- remained unique to one search engine. The type of documents that these phrases most often matched against was journal articles. Journal articles comprised 37.5% of all documents detected by the Scirus searcher but only 8.82% of those matched in Google.

(Figure 3)

Unique Theses Matches: Searching Same Phrase in Other Tool

Tool	Total	Found Same	Found Different	Not Found
Scirus	12	2	2	8
Google	14	3	3	8

10 of 68, or 14.71% were missed as a result of the searcher

16 of 68, or 23.53%, were missed as a result of the search engine choice

(Figure 4)

Scirus			Google			Total	
Corporate Website	0 of 32	0%	Corporate Website	4 of 34	11.76%	Corporate Website	4
Dissertation	1 of 32	3%	Dissertation	2 of 34	5.88%	Dissertation	3
Government Documents	4 of 32	12.50%	Government Documents	3 of 34	8.82%	Government Documents	7
Journal Article	12 of 32	37.50%	Journal Article	3 of 34	8.82%	Journal Article	15
Magazine Article	0 of 32	0.00%	Magazine Article	1 of 34	2.94%	Magazine Article	1
Patent	1 of 32	3%	Patent	0 of 34	0%	Patent	1
Preprint Server	0 of 32	0%	Preprint Server	1 of 34	2.94%	Preprint Server	1
Proceedings	0 of 32	0%	Proceedings	4 of 34	11.76%	Proceedings	4
Professional Society	2 of 32	6.25%	Professional Society	1 of 34	2.94%	Professional Society	3
Published Book	1 of 32	3%	Published Book	2 of 34	5.88%	Published Book	3
Thesis	2 of 32	6.25%	Thesis	3 of 34	8.82%	Thesis	5
University Website	5 of 32	15.62%	University Website	3 of 34	8.82%	University Website	8
Unknown	0 of 32	0.00%	Unknown	2 of 34	5.88%	Unknown	2
Unpublished Article	4 of 32	12.50%	Unpublished Article	5 of 34	14.71%	Unpublished Article	9
		100%			99.97%		66

Types of Documents Found Matching Phrases, Sentences, or Paragraphs in Theses

* Totals equal all unique POPs, not total number of theses matched.

Limitations

The study sample was limited to electronically available theses. The time results would likely have increased if printed theses had been examined. In most cases it is faster to cut and paste text directly into the search engines rather than re-keying it as one

would have to do with printed theses. It is unknown whether institutions that disseminate theses on the web are more or less likely to have run plagiarism checks on them, or for that matter, which institutions in our sample had done so. The sample was limited to theses from just ten institutions and over half the theses were pulled from a single institution. Our definition of a match as 7 consecutive words might have led to false matches- instances where theses authors happened to use the same phrases as other web-published documents. Our study looked only at word-for-word matches and did not explore other forms of plagiarism. Some sites retrieved by the search engines were proprietary and only small sections of the text could be viewed. In such instances, the extent of the match was underreported. The study attempted to simulate the search procedures graduate thesis advisors in science and technology might follow, but neither author is a faculty member in the science and, more important, neither served as an advisor for the theses in the sample. Finally, Internet content changes frequently and search results can vary over time. Our methodology provides no controls for the variable nature of the Internet.

Discussion

The results show that Scirus and Google yielded similar results in terms of total matches and suggest that both tools have potential for detecting plagiarism in electronic science and engineering theses. The Scirus searcher found matches more quickly but this could be attributed to the phrases selected, Internet traffic, number of results retrieved (smaller, more targeted sets of results) and other factors. Our hypothesis that Scirus would be a more effective tool was not confirmed. The Scirus searcher did find many more matches against journal articles, however. This was expected since Elsevier titles

are indexed in Scirus. Google, on the other hand identified more matches against government documents and corporate websites. The difference in the types of documents retrieved by the two search engines suggests that using both search engines would maximize effectiveness.

The incidence of matches found in the study should not be interpreted as actual occurrences of plagiarism. We learned from the earlier study that it is not always easy to determine with certainty whether or not a matched phrase constitutes an occurrence of word-for-word plagiarism. This is why we chose to use the terms “potential occurrences of plagiarism” and “matches” for this study. It is difficult to determine collaborative relationships that students had with professors, whether a student’s thesis was produced before or after the site containing the matched phrase; and whether the student thesis comprised part of a joint endeavor between the academic institution and corporate, government or other research agencies. Actual thesis advisors would have an advantage both in terms of knowing the chronology and collaborative aspects of a student’s thesis research.

The purpose of the project was to compare the effectiveness of the search engines for finding matches of improperly documented phrases—not to determine the extent of plagiarism. However, our results suggest the need for studies on the prevalence of plagiarism in master’s theses in science and engineering. The extent of matches found in the theses in our study is illustrated below (Figure 5). This data was obtained by examining the text surrounding the matched thesis phrases. The figure does not reflect other potential occurrences of plagiarism elsewhere in the thesis. It is important to remember that this data has the same limits as discussed earlier—i.e. it was not always

possible to the full content of the website matches; they include all first matches found by the search engines, etc.

(Figure 5)
Types of Matches Found

	Google	Scirus
Entire paragraph or more	8	3
Multiple sentences	5	3
Entire sentence	5	8
Multiple phrases	5	10
Entire phrase	11	8

Although we did not examine every thesis extensively to determine the full extent of plagiarism, it was possible to identify several in which it appeared very likely that significant plagiarism had occurred. Some examples of what we identified as extremely likely occurrences of plagiarism included:

- The copying of multiple paragraphs of a chapter in a published book that was published in 2001. The authors were faculty at another institution and none were members of the thesis committee. A web search shows that the thesis author is now enrolled in a doctoral program at the same institution where the master's thesis was completed
- The copying of large portions of an earlier published electronic doctoral dissertation from an institution outside the United States
- The inclusion of multiple sentences from a conference paper of the thesis advisor. The conference paper was presented 4 years prior to the completion of the thesis. There was no indication the student was a member of the research team prior to the presentation of the conference paper

The study revealed the need for faculty and universities to consider some of the issues that were encountered during our project. These issues included:

- difficulty in determining whether collaborative relationships existed between theses authors and faculty advisors

- difficulty in determining whether collaborative relationships existed between the thesis author and corporate or governmental agencies
- student contributions to faculty research and the problem of students copying phrases without proper crediting (sometimes with a reference but without quotes; sometimes with neither)
- students using long definitions, formulas, charts, maps, and unique, unfamiliar facts and statistics without crediting a source
- students failing to cite previously published works of their advisors and their own
- problems of working collaboratively with students who have plagiarized and the plagiarized material carries forward into published work of the advisor

The problems surrounding collaborative arrangements might be minimized if all electronic theses included a section such as “collaborative partners” that provided an overview of others individuals, units, agencies and corporations involved in the research or if all publications indicated that the research presented was based on unpublished theses. The problems involving improper citing and the perpetuation of plagiarism into subsequently published works could be minimized if faculty more thoroughly checked the content of theses prior to signing off on them.

Future studies are needed to establish the potential of using Scirus and Google for the purpose of detecting plagiarism, including studies that utilize theses advisors as searchers who are searching actual theses prior to final submission would be useful. Furthermore, studies comparing additional search engines, such as Google Scholar, and commercial databases, such as Science Direct or Web of Science, would be beneficial.

References

- McCullough, Mark and Melissa Holmberg. (in press) Using the Google Search Engine to Detect Word-for-word Plagiarism in Master's Theses: A Preliminary Study. *College Student Journal*.
- Bartlett, Thomas and Scott Smallwood. (Dec. 17, 2004) Special Report: Plagiarism (various pages). *Chronicle of Higher Education*.