# Assessing a Technique for Mining Topic Trends in Bibliographic Databases

**Keith R. Nelms, PhD, PE**
Walker School of Business • Piedmont College • Demorest, GA USA
706-778-8500 • knelms@piedmont.edu

*This presentation briefly reviews a scripted apparatus developed for automating complex bibliographic database searches then scraping results data from the database's web interface. Operating characteristics of the apparatus, capabilities, weaknesses, and precautions are presented. Apparatus output is displayed and discussed. Implications for bibliometrics are considered.*

Commercial bibliographic databases inherently contain a wealth of data about publication activity in a variety of disciplines. Analysis of this data can provide insight into trends and relationships within disciplinary literature.[1,2] For example, bibliometric measures are used to assess individual research productivity by analyzing data from author searches in bibliographic and citation databases.[3] Other bibliometric insights are possible with more complex searches.[4,5] Time and effort required for data harvesting can be a significant barrier to sophisticated bibliometric research.

As part of a family of research tools,[6] the author developed a scripted apparatus[7] to automate tedious, repetitive, and complex bibliographic database research. The apparatus can run multiple queries on common bibliographic databases then "web scrape" the resulting data for analysis. One use of this apparatus is to identify publication trends. Insights can also be gained by comparing publication trends across different publication types (e.g., newspapers, peer-review journals, research reports).

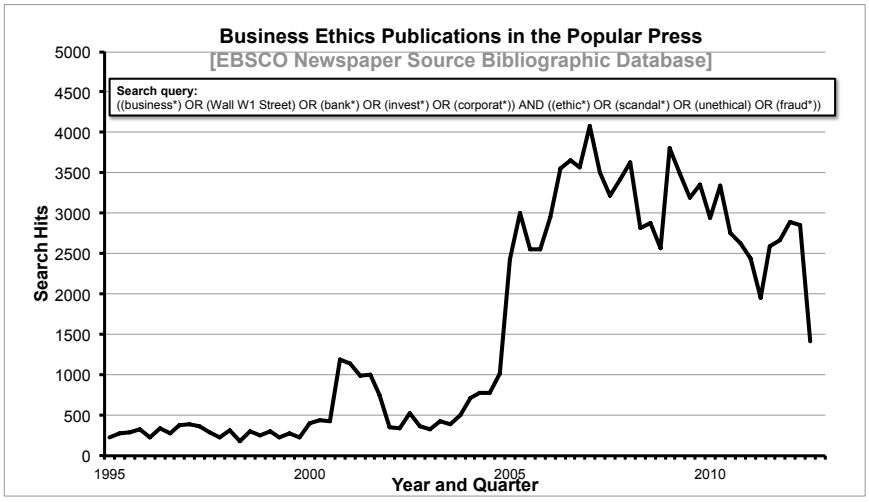
This apparatus makes structured repetitive bibliometric research more practical and efficient. Effective deployment of this technique, however, requires awareness of system limitations. Maintenance programming is occasionally required as the apparatus is vulnerable to vendor updates in the bibliographic database, the browser software, the spreadsheet software, the scripting language, and the operating system. In addition to fluency with Boolean operators, proximity operators, wildcard characters, and field limiters, researchers need general understanding of bibliographic database operations to interpret results. Not all topics are amendable to database searches as some keywords have multiple meanings and generate excessive false search results.

Valid research results should be replicable both conceptually (would different experts consider the search query appropriate?) and technologically (does the same query always produce the same data?). Search results should be evaluated for error. Type 1 errors occur when a search does not capture all appropriate articles. Type 2 errors occur when inappropriate citations are returned. Type 3 errors are bibliographic database artifacts that may mislead results (because, say, database thesaurus terms are not consistently applied). Type 4 errors are data entry errors randomly present in the databases. Statistical quality methods are readily available only for Type 2 errors.

Several simple metrics are useful in evaluating search results data. The most obvious metric is the total number "hits" (found citations) for a search query. Hit rate (number of hits divided by the total number of database citations) reflects a topic's popularity

relative to the overall literature. Hits or hit rates for multiple time periods can be normalized against a single period's value to show percent change over time. It is sometimes visually helpful to use a logarithmic scale for the y-axis.

Researchers should not assume results from automated searches can be used for rigorous statistical metrics common in bibliometrics – automated search results may contain errors not generally present in labor-intensive manual searches. However, search automation provides researchers with a useful tool for identifying trends and uncovering unanticipated questions buried in bibliographic data.



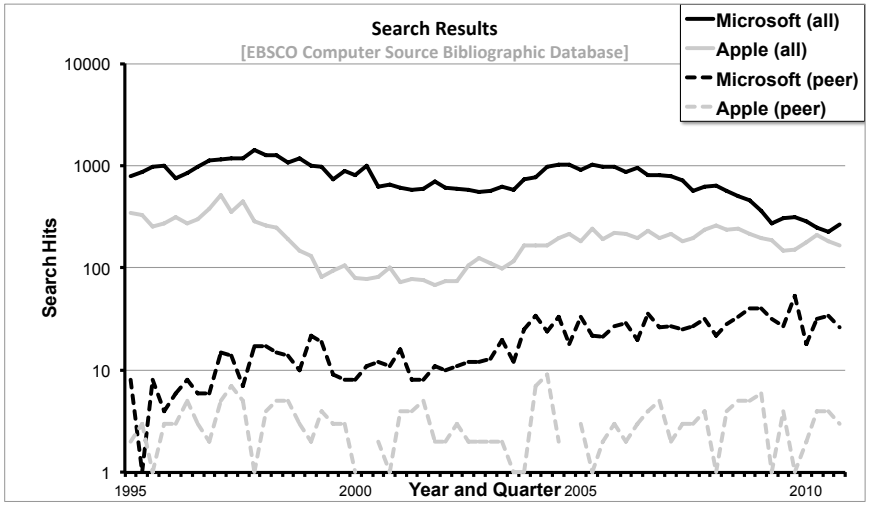

## Figure 1

A complex search query is automatically run in EBSCO's *Newspaper Source* database to study publication trends pertaining to ethical conduct in business.

The research apparatus repeatedly appends the appropriate date range of the form

```
AND (DT >= YYYYMMDD)
AND (DT < YYYYMMDD)
```

to the query shown to generate quarterly data.




## Figure 2

Searches for "Microsoft" and "Apple" show all publications and peer-review publications in EBSCO's *Computer Source* database.

A logarithmic scale makes peer-review results more visible, but lessens the visual impact of dramatic trends. A tremendous *drop* in total Microsoft citations is accompanied by a significant *increase* in Microsoft peer-review citations. This paradox was discovered in this analysis.

## References


[1] Hood, W., Wilson, C. The literature of bibliometrics, scientometrics, and informetrics. *Scientometrics*, 52(2), 291, 2001.

[2] Hicks, D., Tomizawa, H., Saitoh, Y., Kobayashi, S. Evolving indicators: Bibliometric techniques in the evaluation of federally funding research in the United States. *Research Evaluation*, 13(2), 78, 2004.

[3] Ravallion, M., Wagstaff, A. On measuring scholarly influence by citations. *Scientometrics*, 88(1), 321, 2011.

[4] Chang, Y., Huang, M. A study of the evolution of interdisciplinarity in library and information science: Using three bibliometric methods. *Journal of the American Society for Information Science & Technology*, 63(1), 22, 2012.

[5] Wiberley, S. A methodological approach to developing bibliometric models of types of humanities scholarship. *Library Quarterly*, 73(2), 121, 2003.

[6] Nelms, K. Scripting repetitive research tasks, *The Journal of Computing Sciences in Colleges*, 27(2), 199, 2011.

[7] Myer, T., *Apple Automator with AppleScript Bible*, Indianapolis, IN: Wiley Publishing, Inc., 2010.