2018

Understanding the Human Element in Search Algorithms and Discovering How It Affects Search Results

Susan Nevelow Mart
University of Colorado Law School

Note from the Editors: *As part of our ongoing series of research spotlights, this issue features the work of Susan Nevelow Mart, Associate Professor and Director of the Law Library at the University of Colorado-Boulder. Here she reviews for PPIRS members the fascinating results of her search algorithm comparisons in legal databases. The project she describes has received wide attention, including a featured article in the March 2018 issue of ABA Journal. Her full article can be found at "The Algorithm as a Human Artifact: Implications for Legal {Re} Search*," 109 LAW LIBR. J. 387 (2017), available at http://scholar.law.colorado.edu/articles/755/.

Understanding the Human Element in Search Algorithms and Discovering How It Affects Search Results

Susan Nevelow Mart (University of Colorado-Boulder)

Your Search Algorithm Was Created by Humans

If you search online, you are relying on a team of people you never met. The results you see when you hit the *submit* button are governed by the choices those people made when the algorithm was designed. Algorithms just follow the rules. When designing an algorithm for an academic or legal research database, the teams that create the algorithms are trying to solve the same age-old computer communication problem: what documents in the system will help the researcher solve their research problem? The teams designing the algorithms all have the same goal, so does it really matter that different teams of humans created the algorithms for each research database?

As it turns out, the human element in algorithms matters a lot. I recently conducted a study comparing the top ten results of 50 legal searches in six different legal databases. The study looked at Casetext, Fastcase, Google Scholar, Lexis Advance, Ravel (now part of Lexis Advance), and Westlaw. The study limited the database for each search to reported cases in a specific jurisdiction. Because that pool of information is nearly identical, using jurisdictional limits allows true comparisons of the work each algorithm is performing when it processes the search. These results would be transferable to any academic database, if the searches were entered into similarly limited parts of the database. For example, a database of a specific journal title's articles from 1980 to 2017 should have the very same information in it, regardless of whether the articles are searched in JSTOR or Ebscohost.

The results of the study certainly indicate that every group of humans will solve the same problem in a very distinctive way. An average of 40 percent of the top ten results in each database were *unique* to that database. Only a few cases turned up in all six databases. Every database has a point of view, offering unique responses to a legal problem that no other database provides. That is because each database makes different choices about how to process terms in a search.

What Choices Govern Research Algorithms?

While researchers don't know precisely how a specific algorithm works, we do know about some of the options the engineers work with when they create algorithms for legal research. Following are some of the biases (which are preferences in a computer system) that can make a difference:

Terms: How does the algorithm treat the number of terms in the search? If a search has five words in it, will the algorithm require all the words to be in a document, or only some?

Proximity: How close to the words in the search have to be to each other?

Stemming/Other Search Grammar: Humans decide which terms are stemmed, which legal phrases the algorithm recognizes without quotation marks, and if and when legal phrases are added to the search without researcher input.

Network/Citation Analysis: Does the algorithm rely on citation analysis to boost results?

Classification/Content Analysis: Does the system boost results by mining its own classification system or by mining other legal content in the database?

Prioritization: Relevance ranking is one form of prioritizing that emphasizes certain things (like the things in this list) at the expense of others.

Filtering: Including or excluding information according to specific rules or criteria.

Once decisions about how to implement these elements are coded into the algorithm, searches are automatically executed, and researchers have little insight into why certain results are returned. More insight into the search process would improve researchers' ability to get good results. Providing that information to researchers is known as algorithmic accountability. Of course, database providers do have FAQs about searching. The information is just not that detailed.

Looking Into the Search Process

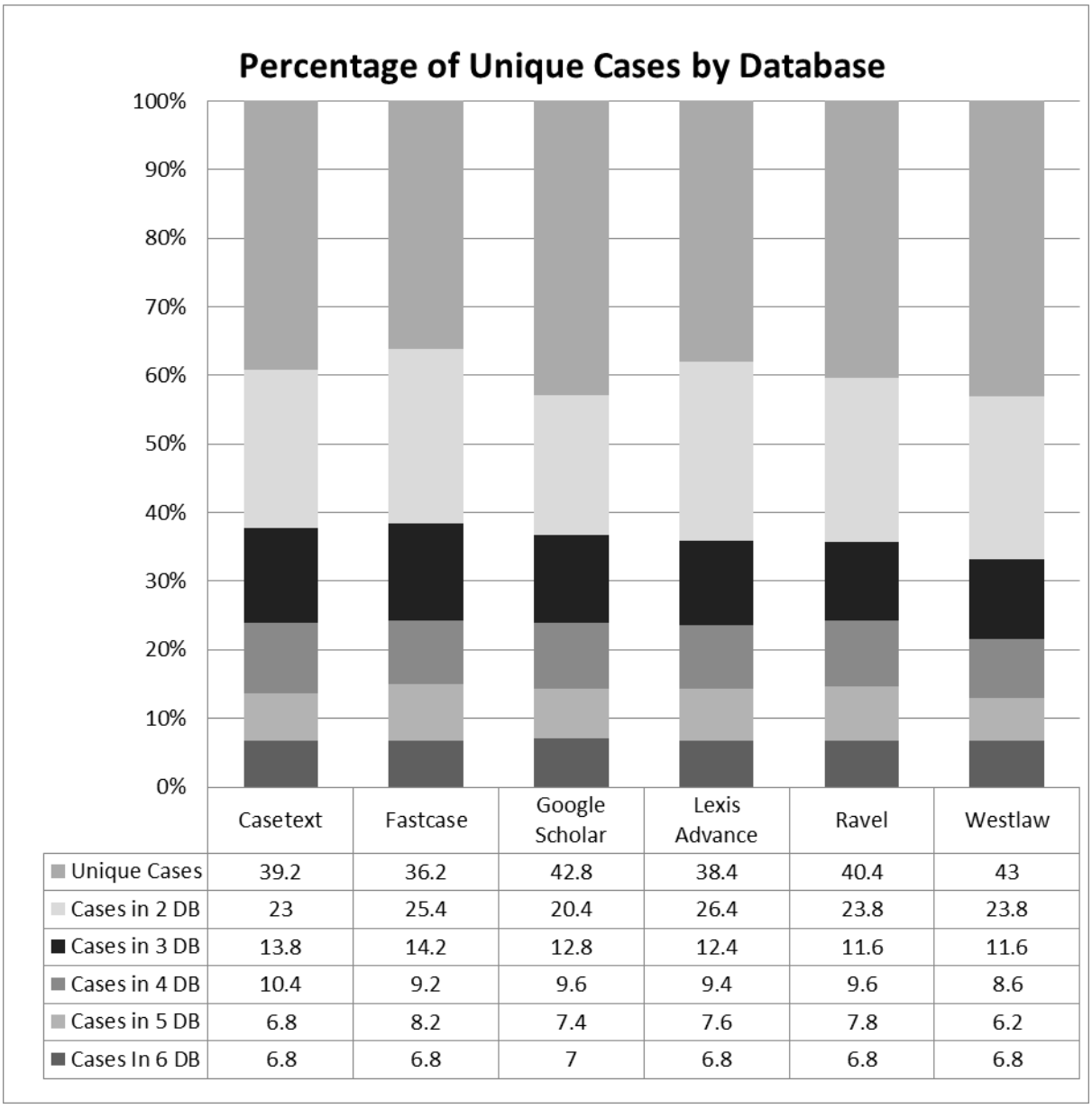
For each of the 50 searches in the study, the research assistants searched in one specific jurisdictional database. Within that jurisdiction, each search needed to return at least ten results in each of the six legal databases, so that there were ten cases to compare from each search. Limiting the results to the top ten made the comparison manageable – only 3000 cases to review! And looking at the top ten is pretty much what modern researchers do. In addition, as researchers, we expect the top results to be the best results. Advertising by legal database providers supports this expectation.

Uniqueness in Search Results

Computer scientists might expect that six different algorithms would solve the same problem in somewhat different ways. In this study, since each algorithm was attempting to bring back results that matched the expectations of a legal researcher with the same objectives, with the same terms, and the same cases to mine, researchers expect to find some similarity in the search results. Both groups would be surprised at the results illustrated in the chart on the next page.

The percentage of unique cases is very high, as the top bar shows. An average of 40 percent of the cases in the top ten results are unique to one database and an average of 25 percent of the cases show up in two of the six databases. The percentages go way down from there.

If you just compare the cases in Lexis Advance and Westlaw, only 28 percent of the cases appear in both databases. That means that 72 percent of the cases returned in the top ten results in each database are unique. Of course, one hopes that no one's research process would end with one search and ten results!



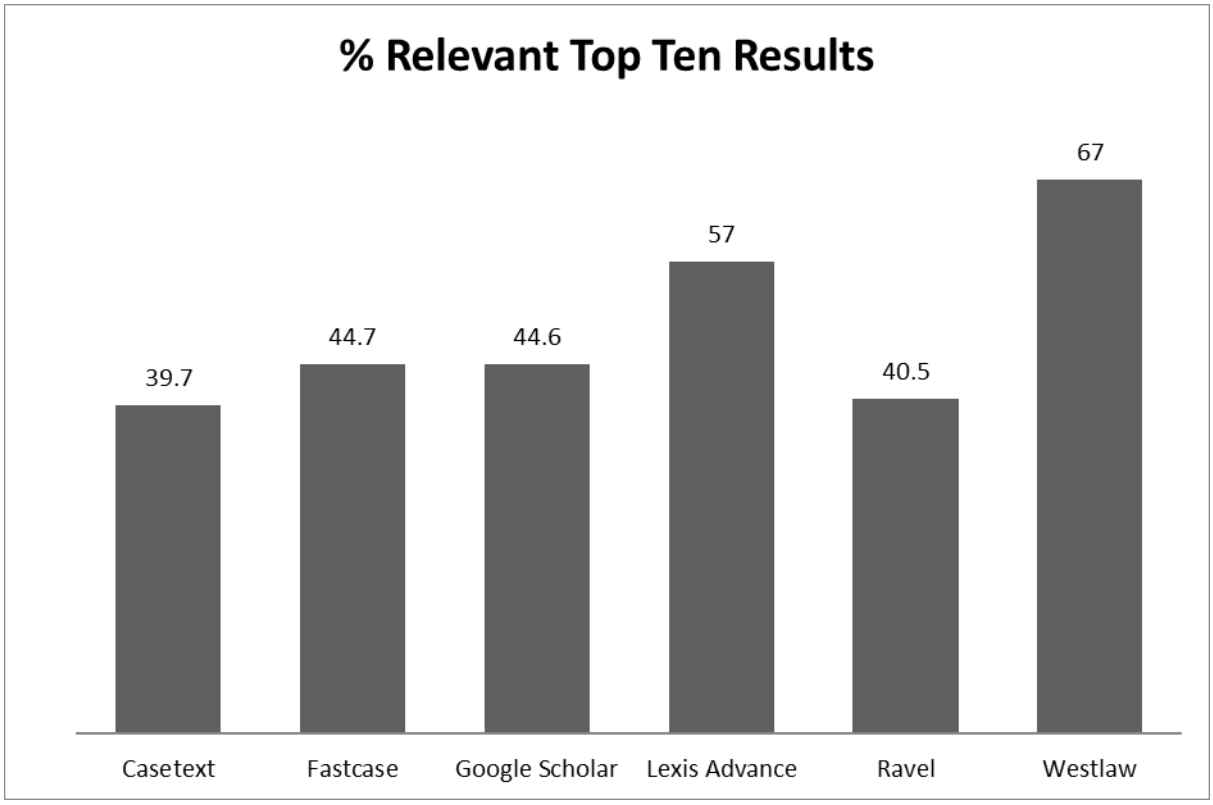
	Casetext	Fastcase	Google Scholar	Lexis Advance	Ravel	Westlaw
Unique Cases	39.2	36.2	42.8	38.4	40.4	43
Cases in 2 DB	23	25.4	20.4	26.4	23.8	23.8
Cases in 3 DB	13.8	14.2	12.8	12.4	11.6	11.6
Cases in 4 DB	10.4	9.2	9.6	9.4	9.6	8.6
Cases in 5 DB	6.8	8.2	7.4	7.6	7.8	6.2
Cases In 6 DB	6.8	6.8	7	6.8	6.8	6.8

What About the Top Ten Results? Are They Relevant?

The next question to answer was whether not those top ten results actually were relevant. Relevance, especially in the legal context, is a highly debatable subject. So the study needed a definition of relevance that could be understood and shared by all of the research assistants, and that would map to the way lawyers think about legal issues. Here is an example of a search that student research assistants were given:

> *federal official Fourth Amendment violation damages recoverable* (search in the N.D. IL)

Most lawyers can immediately translate that into an actual legal issue: I am looking for cases where federal officials may be liable for damages for violating a person Fourth Amendment rights. This background statement is the framework for the students' determinations of relevance. If a case they were reviewing could be helpful to determining the contours of legal issue in *any way*, the case would go into the pile of cases that *are* or *might* be relevant. This is a very broad view of relevance. So how did the different algorithms perform?



There is clearly a clustering of results here. The oldest databases provide more relevant results. Lexis Advance had 57 percent relevant results and Westlaw had 67 percent relevant results. Casetext, Fastcase, Google Scholar, and Ravel had an average of 42 percent relevant cases.

A Few Other Interesting Findings

Each database provided unique results. Of those unique results, only a percentage were both *unique* and *relevant:*

33 percent of Westlaw's cases
20 percent of Lexis Advance's cases
An average of 12 percent of cases for Casetext, Fastcase, Google Scholar and Ravel.

How old or new the cases are also differs by database. Google Scholar had the highest percentage of older cases; almost 20 percent of the cases were from 1921-1978. Westlaw and Fastcase had the highest number of new cases (~ 67%), with Casetext right behind at 64 percent. Ravel and Lexis Advance had an average of 56 percent newer cases.

The number of cases each database returns from a search is quite different. The median number of cases in the results ranged from over 1,00 cases for Lexis Advance to 70 results for Fastcase.. Westlaw, Ravel, and Casetext returned just over 100 results. Google Scholar returned 180 results, and Fastcase returned 70 results.

Time is critical to this study, which is a snapshot of the results with the algorithms as they were when the searches were performed. Database providers are constantly changing their algorithms. Although you could run the exact same searches in the exact same databases, the cases would be *very* different. And not just because new cases have been added. I know, because I have tried this. The numbers shift somewhat, but the differences remain.

Algorithmic World Views

We now know several things about searching that we did not know before. One is the older databases (Lexis Advance and Westlaw) return more cases that are relevant and unique. These databases mine

complex classification systems and secondary sources, each of them very different. However, both of the classification systems have a very 19th century view of the law. The newer entrants into the legal research market may be offering, in their 40 percent of unique cases, results that are not affected by that 19th world view.

Final Thoughts

The important takeaways for researchers and teachers are that every algorithm is *very* different and every database has its own point of view. Researchers need to understand that the variability in results requires multiple searches with multiple terms and in multiple resources. Redundancy in searching is necessary to ensure you are getting a good set of relevant results. Researchers cannot rely on the black box of the algorithm and be satisfied with their initial results.

ACRL Preconference at 2018 ALA Annual Conference: Big Easy RoadShow

Join ACRL in New Orleans for the full-day preconference *Assessment in Action: Demonstrating and Communicating Library Contributions to Student Learning and Success*, an ACRL RoadShow offered in conjunction with the 2018 ALA Annual Conference on Friday, June 22, 2018.

Higher education institutions of all types are facing intensified attention to assessment and accountability issues. Academic libraries are increasingly connecting with colleagues and campus stakeholders to design and implement assessment that documents their contributions to institutional priorities. In this day-long preconference on strategic and sustainable assessment, participants will identify institutional priorities and campus partners, design an assessment project grounded in action research, and prepare a plan for communicating the project results. This preconference is based on the highly successful ACRL Assessment in Action program curriculum.

Complete details, including a full program description, learning outcomes, and registration materials, are available online.