2018

Results May Vary

Susan Nevelow Mart
University of Colorado Law School

Results May Vary

Which database a researcher uses makes a difference

BY SUSAN NEVELOW MART

WHEN A LAWYER SEARCHES IN A LEGAL DATABASE, THAT SINGLE SEARCH BOX IS LIKE A LURE: Put in your search terms and rely on the excellence of the search algorithms to catch the right fish.

At first glance, the various legal research databases seem similar. For instance, they all promote their natural language searching, so when the keywords go into the search box, researchers expect relevant results. The lawyer would also expect the results to be somewhat similar no matter which legal database a lawyer uses. After all, the algorithms are all trying to solve the same problem: translating a specific query into relevant results.

The reality is much different. In a comparison of six legal databases—Casetext, Fastcase, Google Scholar, Lexis Advance, Ravel and Westlaw—when researchers entered the identical search in the same jurisdictional database of reported cases, there was hardly any overlap in the top 10 cases returned in the results. Only 7 percent of the cases were in all six databases, and 40 percent of the cases each database returned in the results set were unique to that database. It turns out that when you give six groups of humans the same problem to solve, the results are a testament to the variability of human problem-solving. If your starting point for research is a keyword search, the divergent results in each of these six databases will frame the rest of your research in a very different way.

SEEING IS BELIEVING

It is easy to forget that the algorithms returning search results are completely human constructs. Those humans made choices about how the algorithms will work. And those choices become the biases and assumptions that are built into research systems. Bias for algorithms simply means a preference in a computer system. While researchers don't know the choices the humans made, we can know the variables that are at work in creating legal research algorithms.

Search grammar: Which terms are automatically stemmed (returned to their root form) and which are not, which synonyms are automatically added, which legal phrases are recognized without quotation marks, how numbers are treated, and how the number of word occurrences in a document determine results—these are examples of search grammar.

Term count: If your search has six words and only five words are in a document, the algorithm can be set to include or exclude the five-term document.

Proximity: The algorithm is preset to determine how close search terms have to be to each other to be returned in the top results.

Machine learning: The programmers decide whether to include instructions that allow the algorithm to "learn" from the data in the database and make predictions.

Prioritization: Relevance ranking is one form of prioritizing that emphasizes certain things at the expense of others. U.S. Supreme Court cases, newer cases or well-cited cases may get a relevance boost.

Network analysis: The extent to which the algorithm uses citation analysis to find and order results is a human choice.

SHUTTERSTOCK

Classification and content analysis: Database providers with full classification systems and access to secondary sources to mine may be programming their algorithms to utilize that value-added content.

Filtering: Decisions about what content to include and exclude from a database affect results. These decisions may be based on copyright or other access issues.

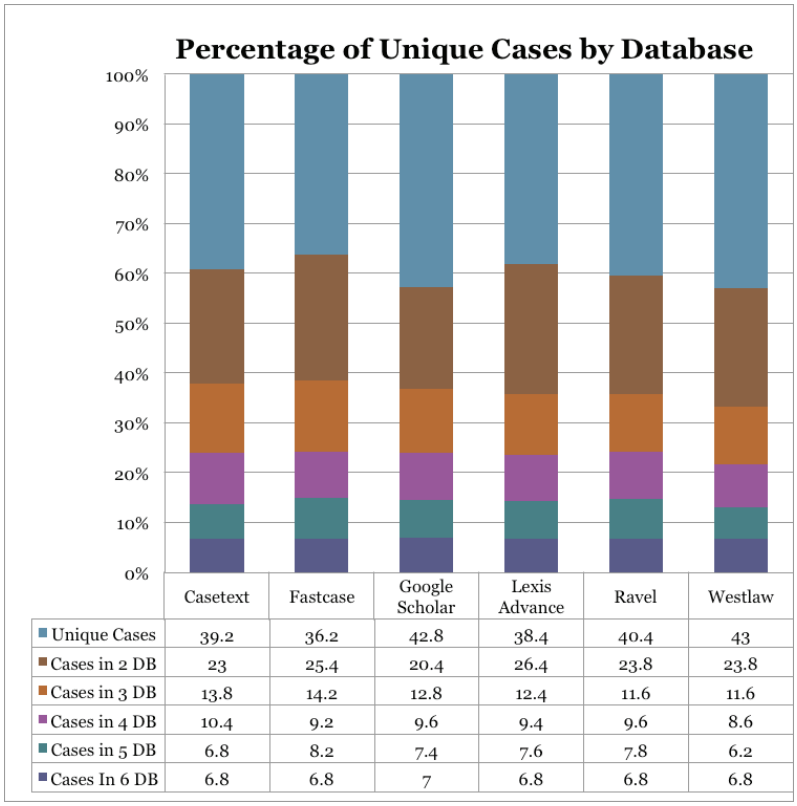
Once these decisions have been made and the code has been implemented, legal researchers don't know how those human choices are affecting search results. But the choices matter to what a researcher sees in the results set. Code is law, as Lawrence Lessig famously said in his 1999 book, *Code and Other Laws of Cyberspace*.

HOW ALGORITHMS WORK

I originally noticed that when I compared a single search in more than one database, the results varied widely. I used these one-off comparisons to illustrate to my students that algorithms differ, and that over-reliance on keyword searching might not be the best search strategy.

I also noticed that if I ran the same search a year later, the results still varied widely and different cases turned up in the results. One would expect new cases to show up, but older cases turned up as well. Algorithms are fluid, not static. Since one-off searches do not prove that much, I thought it would be interesting to run the experiment on a larger scale and see what happened. I crafted 50 different searches and had law student research assistants look at the top 10 results.

How unique are the search results? When you search in most databases, there is no way to determine what documents are actually in the database and which documents are excluded. In legal databases, jurisdictional and coverage limits allow you to know exactly which set of documents is being searched. If one searches a database of reported cases in the 6th U.S. Circuit Court of Appeals at Cincinnati, every database provider has the same documents, plus or minus a few cases from 1925 to 1933.

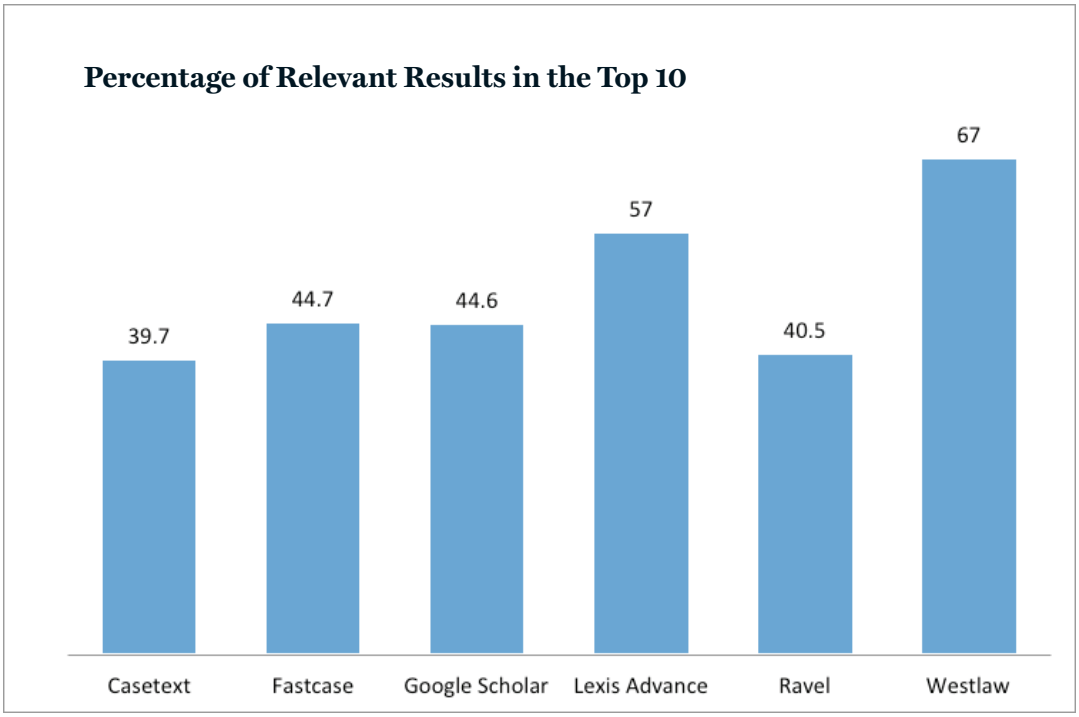


	Casetext	Fastcase	Google Scholar	Lexis Advance	Ravel	Westlaw
Unique Cases	39.2	36.2	42.8	38.4	40.4	43
Cases in 2 DB	23	25.4	20.4	26.4	23.8	23.8
Cases in 3 DB	13.8	14.2	12.8	12.4	11.6	11.6
Cases in 4 DB	10.4	9.2	9.6	9.4	9.6	8.6
Cases in 5 DB	6.8	8.2	7.4	7.6	7.8	6.2
Cases In 6 DB	6.8	6.8	7	6.8	6.8	6.8

Computer scientists would expect some variability in search algorithms, even if lawyers do not have the same expectations. Here, however, each vendor's group working on the research algorithm has an identical goal: to translate the words describing legal concepts into relevant documents. One of the hypotheses of the study was that, as the number of searches expanded, the overall results returned by the algorithms from each database provider would be similar. The top 10 cases ought to be somewhat similar. That hypothesis did not turn out to be true, as shown in the chart above, "Percentage of Unique Cases by Database."

The blue bar at the top shows the percentage of unique cases in each database. An average of **40 percent** of the cases in the top 10 results are unique to one database. Nearly **25 percent** of the cases only show up in two of the databases. The numbers drop quickly after that, and only **7 percent** of the cases show up in five or six of the databases. When the comparison was limited to the two oldest database providers, Lexis Advance and Westlaw, there was only **28 percent** overlap. That means that **72 percent** of the top 10 cases are unique to each provider.

Starting with a keyword search is just one way to frame a research problem. Legal research is a process that has always required redundancy in searching. The rise of algorithms has not changed that. Researchers need to use multiple searches, of multiple types, in multiple resources. But if a researcher starts with a keyword search, each legal database provider is going to offer a different set of results and, therefore, a different frame for the



next steps in the research process. This means that where you start your search matters.

RESEARCHERS WANT RELEVANT RESULTS

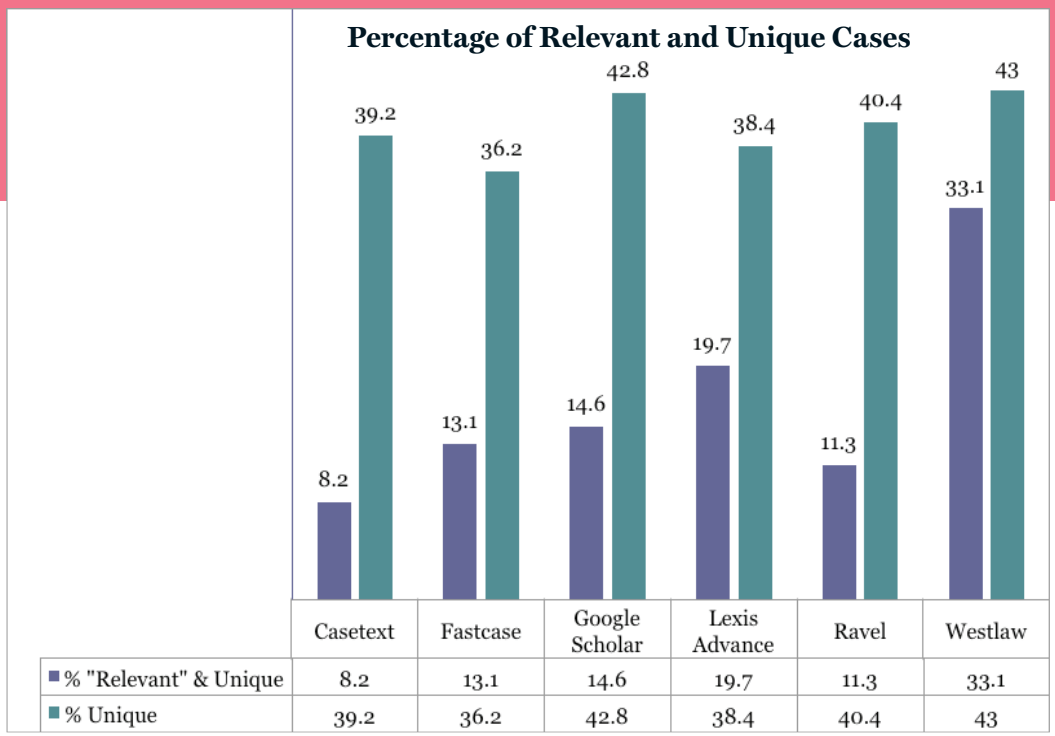
The searches for the study each incorporated known legal concepts. The searches were the kind that a lawyer with any expertise in the area could easily translate into a recognizable legal issue. Here is an example of the kind of search used in the study: *criminal sentence enhancement findings by jury required* (the search was limited to the reported cases in the 6th Circuit).

Lawyers with subject expertise would know that the search is about the constitutionality of increasing the penalty for a crime when the jury did not make a specific finding about the facts that enhanced the penalty. This background statement was given to the RA who ran the search in each of the six legal databases and read the resulting top 10 cases from each database to see whether the cases were relevant or not. This translation—from the human putting in keywords that represent a legal problem to the documents the human-created algorithm determines are responsive—is at the heart of all human/computer legal research interaction. The study tested how the humans creating the algorithms tried to implement that translation. The decision to limit the results to the top 10 was based in part on the assumption that returning relevant results at the top is the goal of every team creating a legal research algorithm, a view that database provider ads and FAQs support. And modern researchers tend to look at the top results and then move on.

The RAs were given a framework for relevance determinations based on the background statement and on explicit instructions for determining relevance: A case was relevant, in our example, if it discussed situations where juries did (or did not) make sufficient factual determinations to support an enhancement of the sentence in a criminal case. If a case was in any way related to determining the contours of the role of the jury, it would be marked as "would definitely be saved for further review" or "would probably be saved for further review." This study does not say that the cases that are "relevant" are necessarily the best cases, just that they are cases playing some "cognitive role in the structuring of a legal argument," as Stuart Sutton put it in *The Role of Attorney Mental Models of Law in Case Relevance Determinations: An Exploratory Analysis*. This is a broad and subjective view of relevance that should resonate with all attorneys who have created mental models of an area of the law.

See the next chart, at the top of this page, "Percentage of Relevant Results in the Top 10," which illustrates relevance in each of our six legal databases.

What is striking about this chart is how many results are *not* relevant. Even within 10 cases, not all of the results relate to the search terms. Westlaw (67 percent relevance) and Lexis Advance (57 percent relevance) performed the best. For Casetext, Fastcase, Google Scholar and Ravel (now owned by Lexis), an average of about 40 percent of the results



were relevant. In terms of each database provider offering a different view of the same corpus of cases, how many of those relevant results were unique?

The final chart, above, "Percentage of Relevant and Unique Cases," reflects how each database provider offers cases that are both unique and relevant in the top 10 results. Westlaw offers the highest percentage of such cases, at just over 33 percent. Lexis Advance has nearly 20 percent unique and relevant cases. Casetext, Fastcase, Google Scholar and Ravel have an average of 12 percent of relevant and unique cases. Of course, you don't have to do the same search in all six databases to find all the relevant cases. All the cases are in all of the databases, and multiple searches may bring those unique and relevant results to the top.

The takeaway is that lazy searching will leave relevant results buried; if an important case is the 57th result from just one search, a researcher is not going to find it. Algorithms are just not going to do the heavy lifting in legal research. At least not yet.

OTHER DATA FROM THE STUDY

The study also looked at the age of cases that were returned in each search. Overall, the oldest cases dominated Google Scholar's results. Almost 20 percent of the results from Google Scholar were from 1921 to 1978. The highest percentage (about 67 percent) of newer cases were returned by Fastcase and Westlaw. Ravel and Lexis Advance had an average of 56 percent newer cases.

Another area of diversity was the number of cases each database returned. The median number of cases returned in response to the same search varied from 1,000 for Lexis Advance to 70 for Fastcase. Casetext, Ravel and Westlaw each returned 180 results at the 50th percentile and Google Scholar returned 180. Each algorithm is set to determine what is responsive to the same search terms in vastly different ways.

For the most part, these algorithms are black boxes—you can see the input and the output. What happens in the middle is unknown, and users have no idea how the results are generated. While legal database providers tend to view their algorithms as trade secrets, they do give some hints in their promotional materials about how the algorithms work. A more detailed discussion of those materials (and other concepts in this article) is available in "The Algorithm as a Human Artifact: Implications for Legal (Re)Search" in the *Law Library Journal*. We need a frank discussion with database providers about what it means for a researcher to search in their databases and how researchers can become better searchers. Knowing that should not violate any trade secrets. Discussing algorithmic accountability with database providers can work, though proactive responses would be better. For example, I asked Lexis Advance

about jurisdictional searching and they released a FAQ on the topic. No trade secrets were revealed, and researchers now have a better understanding of how to effectively search in Lexis Advance. Fastcase has responded to the discussions about algorithmic accountability by releasing an advanced search feature that lets the researcher adjust the relevance ranking for a specific search to privilege the attributes that researcher wants to emphasize. Algorithmic accountability is now open for discussion. Providing the kind of algorithmic accountability that enables researchers to create better searches should be a market imperative for all database providers, so please demand accountability.

As a matter of empirical fact, we now know some things about using legal databases that researchers had suspected but could not prove. We know that Westlaw and Lexis Advance return more relevant and unique results.

These databases have an edge: They've had decades to refine their strategies. Both have a large base of user information. Each has a detailed but different classification system and different sets of secondary sources. Recall that only 28 percent of the cases from Lexis Advance and Westlaw appear in both databases. It may not be so surprising that the results from Lexis and Westlaw are so different, as those results may differ in ways that conform to the respective worldviews encapsulated in their classification systems and the secondary sources their algorithms mine to return results.

This raises questions about two types of viewpoint discrimination that are worth exploring. The first is one familiar to all researchers: Authorial viewpoints are a form of viewpoint discrimination. Attorneys and librarians have always preferred, budgets allowing, to have more than one authorial viewpoint represented in their legal resources. What held true on the treatise level now holds true on the database level, and the differing worldviews of each database provider can be seen as a positive good.

The second kind of viewpoint discrimination results from the 19th-century viewpoint explicitly imported into Westlaw through its Key Number classification system and re-created in Lexis in its own classification system. Scholars have often pointed out that older and more established legal topics (think of contract rescission) fare better in these systems. Newer topics (which have changed over time, from civil rights in the 1960s and '70s to cybersecurity today) are harder to fit into the existing schemes. So it is possible that searches in more established areas of the law will be more successful in these older databases.

If one is searching for solutions to legal problems in emerging areas of the law, it would be worthwhile to try the newer databases and see what their 40 percent of unique cases have to offer. The newer databases also offer new forms of serendipity: "summaries from subsequent cases" and the "black letter law" filters in Casetext, as well as citation visualizations in Ravel and Fastcase, are examples of new ways of adding value to the research process.

A FEW LAST WORDS

Researchers should take away a few key things from the study:

- Every algorithm is different.
- Every database has a point of view.
- The variability in search results requires researchers to go beyond keyword searching.
- Keyword searching is just one way to enter a research universe.
- Redundancy in searching is still of paramount importance.
- Terms and connectors searching is still a necessary research skill.
- Researchers need to demand algorithmic accountability. We are the market, and we can influence the product.

Algorithms are the black boxes that human researchers are navigating. Humans created those black box algorithms. We need better communication between these two sets of humans to facilitate access to the rich information residing in legal databases. ■

Susan Nevelow Mart is an associate professor and the director of the law library at the University of Colorado Law School in Boulder.