NATURE | NEWS: Q&A

Google Scholar pioneer on search engine's future

As the search engine approaches its 10th birthday, *Nature* speaks to the co-creator of Google Scholar.

Richard Van Noorden

07 November 2014

Amit Basu

Google Scholar, the free search engine for scholarly literature, turns ten years old on 18 November. By 'crawling' over the text of millions of academic papers, including those behind publishers' paywalls, it has transformed the way that researchers consult the literature online. In a *Nature* survey this year, some 60% of scientists said that they use the service regularly. *Nature* spoke with Anurag Acharya, who co-created the service and still runs it, about Google Scholar's history and what he sees for its future.

How do you know what literature to index?

'Scholarly' is what everybody else in the scholarly field considers scholarly. It sounds like a recursive definition but it does settle down. We crawl the whole web, and for a new blog, for example, you see what the connections are to the rest of scholarship that you already know about. If many people cite it, or if it cites many people, it is probably scholarly. There is no one magic formula: you bring evidence to bear from many features.

Where did the idea for Google Scholar come from?

I came to Google in 2000, as a year off from my academic job at the University of California, Santa Barbara. It was pretty clear that I was unlikely to have a larger impact [in academia] than at Google — making it possible for people everywhere to be able to find information. So I gave up on academia and ran Google's web-indexing team for four years. It was a very hectic time, and basically, I burnt out.

Alex Verstak [Acharya's colleague on the web-indexing team] and I decided to take a six-month sabbatical to try to make finding scholarly articles easier and faster. The idea wasn't to produce Google Scholar, it was to improve our ranking of scholarly documents in web search. But the problem with trying to do that is figuring out the intent of the searcher. Do they want scholarly results or are they a layperson? We said, "Suppose you didn't have to solve that hard a problem; suppose you knew the searcher had a scholarly intent." We built an internal prototype, and people said: "Hey, this is good by itself. You don't have to solve another problem — let's go!" Then Scholar clearly seemed to be very useful and very important, so I ended up staying with it.

Was it an instant success?

It was very popular. Once we launched it, usage grew exponentially. One big difference was that we were relevance-ranking [sorting

results by relevance to the user's request], which scholarly search services had not done previously. They were reverse-chronological [providing the newest results first]. And we crawled the full text of research articles, though we did not include the full text from all the publishers when we started.

It took years in some cases to convince publishers to let you crawl their full text. Was that hard?
It depends. You have to think back to a decade ago, when web search was considered lightweight — what people would use to find pictures of Britney Spears, not scholarly articles. But we knew people were sending us purely academic queries. We just had to persuade publishers that our service would be used and would bring them more traffic. We were working with many of them already before Google Scholar launched, of course.

In 2012 Google Scholar was removed from the drop-down menu of search options on Google's home page. Do you worry that Google Scholar might be downgraded or killed?
No. Our team is continually growing, from two people at the start to nine now. People may have treated that menu removal as a demotion, but it wasn't really. Those menu links are to help users get from the home page to another service, so they emphasize the most-used transitions. If users already know to start with Google Scholar, they don't need that transition. That's all it was.

How does Google Scholar make money?
Google Scholar does not currently make money. There are many Google services that do not make a significant amount of money. The primary role of Scholar is to give back to the research community, and we are able to do so because it is not very expensive, from Google's point of view. In terms of volume of queries, Google Scholar is small compared to many Google services, so opportunities for advertisement monetization are relatively small. There's not been pressure to monetize. The benefits that Scholar provides, given the number of people who are working on it, are very significant. People like it internally — we are all, in part, ex-academics.

How many queries does Google Scholar get every day, and how much literature does the service track? (Estimates place it anywhere from 100 million to 160 million scholarly items).
I'm unable to tell you, beyond a very, very large number. The same answer for the literature, except that the number of items indexed has grown about an order of magnitude since we launched. A lot of people wonder about the size. But this kind of discussion is not useful — it's just 'bike-shedding'. Our challenge is to see how often people are able to find the articles they need. The index size might be a concern here if it was too small. But we are clearly large enough.

Google Scholar has introduced extra services: author profile pages and a recommendations engine, for instance. Is this changing it from a search engine to something closer to a bibliometrics tool?
Yes and no. A significant purpose of profiles is to help you to find the articles you need. Often you don't remember exactly how to find an article, but you might pivot from a paper you do remember to an author and to their other papers. And you can follow other people's work — another crucial way of finding articles. Profiles have other uses, of course. Once we know your papers, we can track how your discipline has evolved over time, the other people in the scholarly world that you are linked to, and can even recommend other topics that people in your field are interested in. This helps the recommendations engine, which is a step beyond [a search engine].

Are you worried about the practice known as gaming — people creating fake papers, getting them indexed by Google, and gaining fake citations?
Not really. Yes, you can add any papers you want. But everything is completely visible — articles in your profiles, articles citing yours, where they are hosted, and so on. Anyone in the world can call you on it, basically killing your career. We don't see spam for that very reason. I have a lot of experience dealing with spam because I used to work on web search. Spam is easier when people are anonymous. If I am trying to build a publication history for my public reputation, I will be relatively cautious.

What features would you like to see in the future?
We are very good at helping people to find the articles they are looking for and can describe. But the next big thing we would like to do is to get you the articles that you need, but that you don't know to search for. Can we make serendipity easier? How can we help everyone to operate at the research frontier without them having to scan over hundreds of papers — a very inefficient way of finding things — and do nothing else all day long?

I don't know how we will make this happen. We have some initial efforts on this (such as the recommendations engine), but it is far from what it needs to be. There is an inherent problem to giving you information that you weren't actively searching for. It has to be relevant — so that we are not wasting your time — but not too relevant, because you already know about those articles. And it has to avoid short-term interests that come and go: you look up something but you don't want to get spammed about it for the rest of your life. I

don’t think getting our users to ‘train’ a recommendations model will work — that is too much effort.

(For more on recommendation services, see 'How to tame the flood of literature', in *Nature*'s Toolbox section.)

What about helping people search directly for scientific data, not papers?
That is an interesting idea. It is feasible to crawl over data buried inside paywalled papers, as we do with full text. But then if we link the user to the paywalled article, they don’t see this data — just the paper’s abstract. For indexing full-text articles, we depend on that abstract to let users estimate the probable utility of the article. For data we don't have anything similar. So as a field of scholarly communication, we haven’t yet developed a model that would allow for a useful data-search service.

Many people would like to have an API (Application Programming Interface) in Google Scholar, so that they could write programs that automatically make searches or retrieve profile information, and build services on top of the tool. Is that possible?
I can’t do that. Our indexing arrangements with publishers preclude it. We are allowed to scan all the articles, but not to distribute this information to others in bulk. It is important to be able to work with publishers so we can continue to build a comprehensive search service that is free to everybody. That is our primary function, and everything else is in addition to this.

Do you see yourself working at Google Scholar for the next decade?
I didn’t expect to work on Google Scholar for ten years in the first place! My wife reminds me it was supposed to be five, then seven years — and now I’m still not leaving. But this is the most important thing I know I can do. We are basically making the smartest people on the planet more effective. That’s a very attractive proposition, and I don’t foresee moving away from Google Scholar any time soon, or any time easily.

Does your desire for a free, effective search engine go back to your time as a student at the Indian Institute of Technology Kharagpur?
It influenced the problems that appealed to me. For example, there is no other service that indexes the full texts of papers even when the user can see only the abstract. The reason I thought this was an important direction to go in was that I realised users needed to know the information was there. If you know the information is in a paywalled paper, and it is important to you, you will find a way in: you can write to the author, for instance. I did that in Kharagpur — it was really ineffective and slow! So my experiences informed the approach I took. But at this point, Google Scholar has a life of its own.

Should people who use Google Scholar have concerns about data privacy?
We use the standard Google data-collection policies — there is nothing different for Scholar. My role at Google is focused on Google Scholar. So I am not going to be able to say more about broader issues.

Nature | doi:10.1038/nature.2014.16269