NATURE | NEWS

Artificial-intelligence institute launches free science search engine

Semantic Scholar comes from centre backed by Microsoft co-founder Paul Allen.

Nicola Jones

02 November 2015

Brett Hartman/TED/CC BY-NC 2.0

Oren Etzioni, chief executive officer of the Allen Institute for Artificial Intelligence. *(CC BY-NC 2.0).*

With Google Scholar, PubMed, and other free academic databases at their fingertips, scientists may feel they have plenty of resources to trawl through the ever-growing science literature.

But a search engine unveiled on 2 November by the non-profit Allen Institute for Artificial Intelligence (AI2) in Seattle, Washington, is working towards providing something different for its users: an understanding of a paper's content. "We're trying to get deep into the papers and be fast and clean and usable," says Oren Etzioni, chief executive officer of AI2.

The free product, called Semantic Scholar, is currently limited to searching about 3 million open-access papers in computer science. But the AI2 team aims to broaden that to other fields within a year, Etzioni says. His team is well financed: AI2 was founded and is backed by Microsoft co-founder Paul Allen, who has given the institute more than US$20 million since 2013.

Semantic Scholar offers a few innovative features, including picking out the most important keywords and phrases from the text without relying on an author or publisher to key them in. "It's surprisingly difficult for a system to do this," says Etzioni. The search engine uses similar 'machine reading' techniques to determine which papers are overviews of a topic.

The system can also identify which of a paper's cited references were truly influential, rather than being included incidentally for background or as a comparison. "That's a really good feature," says Jose Manuel Gomez-Perez, who works on search engines and is director of research and development in Madrid for the software company Expert System. Semantic Scholar also extracts figures from the papers to present in the search result.

Scant competition

There are only a few widely used free academic search engines. Google Scholar is by far the largest, encompassing an estimated 100 million or more documents. But it has its problems. "A significant proportion of the documents are not scholarly by anyone's measure," says Péter Jacsó, an information scientist who studies search engines at the University of Hawaii at Manoa. When Jacsó analysed

Google Scholar in 2009, he found some comical results, including papers that were apparently cited a decade before they were published; tags such as 'table of contents' being misidentified as the author; and page numbers being mistaken for the year of publication. Although some of those issues have been fixed, says Jacsó, “there are still millions and millions of errors.”

“Google has access to a lot of data. But there’s still a step forward that needs to be taken in understanding the content of the paper,” says Gomez-Perez.

Another free service, called Microsoft Academic Search, which includes more than 30 million documents, has effectively stopped adding new papers. Microsoft says that it is incorporating the data into its Bing search engine. An alternative search engine created by academics, called CiteSeer, contains 5.3 million records; its director, Lee Giles of Pennsylvania State University in University Park, is a collaborator of Semantic Scholar.

Early days

Despite the narrow field of competition, says Jacsó, a new effort is not really important unless it is likely to become a “Google Scholar beater” — and it is too early to say whether that is the case for Semantic Scholar, he notes. So far, the new service has big shortcomings, says Jacsó, including failing to pick up publication titles for many search results.

Google Scholar pioneer on search engine’s future

Etzioni says that the service finds about 80% of available free-access papers — including self-archived material — in journals, conference websites or in the records of academic institutions. However, unlike Google Scholar, the search engine cannot see behind paywalls. "We don’t have a way past the paywall — it's a limitation for us," Etzioni says. “But we feel the tide is turning. More and more stuff is available somewhere.”

The AI2 team started with the field of computer science so that they could analyse results in topics that were familiar to them. But they aim to expand to other areas in 2016. Medicine is a particular priority, says Etzioni. “I’ve talked to people who say that doctors are in the emergency room looking up things on Google Scholar on their phones,” he says. “They have what I’d consider a fairly blunt instrument.”

Semantic Scholar is one step towards AI2’s ambition, which is to create a computerized service that can read through the scientific literature to identify useful hypotheses and experiments. (Others, including IBM, are also working on similar tools.)

“Our goal is to enable researchers to find answers to some of science’s thorniest problems,” says Etzioni.

Nature | doi:10.1038/nature.2015.18703