UC SANTA BARBARA

THE Current

Science + Technology

Computer scientist William Wang receives Karen Spärck Jones Award

Sonia Fernandez
January 23, 2023

Share this article

When UC Santa Barbara computer science professor William Wang received a call from an unfamiliar number at 7 a.m. on a Monday, he did what most people trying to avoid a robocall would do: He ignored it.

The caller was persistent though, and eventually Wang got curious and figured out where the mysterious number was coming from: the British Computer Society (BCS). The next time they called, he picked up — only to receive a robust congratulations for being chosen as the recipient of the BCS's Karen Spärck Jones Award.

"It was actually the chair of the award committee," Wang recalled, noting the BCS awards panel informs prize recipients by phone à la the Nobel."

We take great pleasure in sending our warmest congratulations to William Wang upon receiving the Karen Spärck Jones Award," said Tresa Pollock, interim dean of the UCSB College of Engineering. "We at UCSB are well aware of Professor Wang's innovative research in the field of natural language processing, and this well-deserved award will extend his recognition further. This is wonderful news."

“I’m humbled and honored to receive this prestigious award,” Wang said, “because I first learned about Karen Spärck Jones’s algorithms when I was a student, and she’s truly an inspiring pioneer in information retrieval and natural language processing.”

An eminent honor in the field of natural language processing (NLP) and information retrieval (IR), the Karen Spärck Jones Award is named for a computer scientist who — in addition to being a woman in computer science, a rarity at the time — merged linguistics with statistics, developing the principles that underlie today’s web searches, predating Google by two decades.

“She basically invented the modern information retrieval algorithm,” Wang said. If you want to search anything online, he explained, the task for Google and other search engines becomes one of relevance, as they work to retrieve and rank links to information relevant to the searcher in an ocean of other possible links. User diversity, location and robustness to user input are all challenges that search engines face as they scour the internet for the right information.

“Karen Spärck Jones invented an algorithm called TF-IDF,” Wang said. Short for “term frequency-inverse document frequency,” the algorithm goes beyond just counting how often a search term appears in a document in the engine’s calculations of relevance. Rather, it weighs how important a search term or phrase is in a document (such as a webpage) among a collection of other documents, known as a corpus.

“The TF-IDF algorithm actually balances the frequency of the terms and how likely they are to appear in other documents,” Wang explained. This has the effect of not only canceling out very common words such as “the” or “is” that don’t carry much useful information, it also circumvents search engine optimization (SEO) practices where documents are stuffed with keywords to increase search rankings, though the documents themselves have little to no relevance to the user.

The algorithm is more than 50 years old, but it still holds up, Wang said. “Nowadays, even though people have come up with neural network or deep learning-based approaches, without training, they still struggle to compare with Karen Spärck Jones’s system from 40–50 years ago,” he said. “It was one of the earliest algorithms that I implemented as a computer science student.”

TF-IDF continues to be a mainstay of Wang’s current work, in which he and his students are developing an open-domain answering system — an engine that can

give long responses to questions using natural language, deriving information from a large corpus. Using newer machine learning training techniques and the TF-IDF algorithm, his aim is to add specificity to relevance.

“How can you make use of heterogenous data? There’s tabular data, there’s text data, there’s speech, image, video,” he said. “My research is about how we can generate knowledge from these heterogenous sources.”

Wang, the Duncan and Suzanne Mellichamp Chair in Artificial Intelligence Design at UC Santa Barbara, is also the co-director of the Natural Language Processing Group and of the Center for Responsible Machine Learning. He will receive the BCS Karen Spärck Jones Award and deliver a lecture in Dublin in April 2023.

Tags

Artificial Intelligence

Media Contact

Daniel Smith

Media & Office Coordinator

(805) 893-2191

danielsmith@ucsb.edu

Share this article

About UC Santa Barbara

The University of California, Santa Barbara is a leading research institution that also provides a comprehensive liberal arts learning experience. Our academic community of faculty, students, and staff is characterized by a culture of interdisciplinary collaboration that is responsive to the needs of our multicultural and global society. All of this takes place within a living and learning environment like no other, as we draw inspiration from the beauty and resources of our extraordinary location at the

edge of the Pacific Ocean.