Getting to Know You

A UCSC computer scientist's research is laying the foundation for the next generation of personalized search engines

By Tim Stephens

In October 2006, the movie rental company Netflix announced a $1 million competition to develop a better movie recommendation system. Within a week, a team led by UC Santa Cruz computer scientist Yi Zhang had already developed a computer algorithm that performed better than the Cinematch system Netflix currently uses to help customers identify movies they'll like.

The Netflix competition offered a well-defined problem right in line with Zhang's interests. An assistant professor of computer science in the Baskin School of Engineering, Zhang specializes in tools for managing large amounts of electronic information. She uses artificial intelligence and machine learning techniques to develop computer systems that can learn about a user's interests and preferences from each interaction and thereby help people find the information they need more efficiently.

The Netflix contest, which will continue through 2011, now involves nearly 20,000 teams from 150 different countries. The company has made its enormous dataset—about 100 million movie ratings by customers—available to contestants, hiding personal information and the most recent rating of each customer. The challenge is to predict customers' most recent ratings based on their previous ratings of other movies. Zhang's team was in first place on the contest's online "leaderboard" for months and remains among the top ten teams.

"It's a very interesting problem, so I play with it when I have the time," she says. "This kind of recommendation system is the future of e-commerce—all the online stores want to make recommendations to their customers."

Lately Zhang has turned her attention back to ongoing research projects. Her Information Retrieval and Knowledge Management Lab is working on projects funded by companies such as Google, Yahoo, and Cisco, and collaborating with researchers in UCSC's Storage Systems Research Center (SSRC).

> "UCSC is taking the initiative to be the UC campus for Silicon Valley. When I taught a course last year at the Silicon Valley Center, I had students from Apple, Intel, and Yahoo." —Yi Zhang

One of Zhang's goals is to develop better search engines for the Internet. She is approaching this problem in various ways, all of which involve computer systems that are able to learn from user feedback to improve the quality of search results over time.

Search engines like Google do an impressive job of finding web sites related to a search query. But web searches can be frustrating when the most useful sites are buried among thousands of less useful "hits." Zhang envisions a search engine that analyzes your choices as you browse through search results, learning what kinds of sites are most useful to you and using that information to guide subsequent searches. The system might also ask you questions to narrow the search.

"When you type in a search query, you typically leave out a lot of information about what you're looking for. A personalized search engine can learn to infer what's hidden in the query," Zhang says.

A related approach, called collaborative information retrieval, optimizes search results based on information from searches performed by other users. Again, it requires a computer system capable of active learning. "In collaborative filtering, the system learns from other people's choices and uses that to guide information retrieval," Zhang says.

Whereas search engines "pull" information from the web in response to a query, a personalized system could also be designed to "push" information without a query. Currently, savvy web surfers can have "RSS feeds" delivered automatically from their favorite web sites. But first they have to find those sites and request the feeds. What if your computer knew what kinds of information you were interested in and found it for you?

"The push-based scenario complements the traditional search engine to meet long-term information needs by delivering things like stock information, news, and entertainment. Of course, different users have different criteria, so the computer has to be able to learn from user feedback," Zhang says.

A computer system with such intimate knowledge of an individual's habits and preferences may be efficient and convenient, but it also raises privacy and security concerns. According to Zhang, however, good system designs can protect privacy and let users control how much information they want to share. "Good designs and good policies can protect privacy to the extent that the user desires," she says.

Zhang divides her time between an office on the UCSC campus and another at UCSC's Silicon Valley Center. She teaches classes in both locations, has students who work at companies in the valley, and often meets with industry representatives to discuss her research.

"UCSC is taking the initiative to be the UC campus for Silicon Valley," Zhang says. "I have many part-time students who are working in industry, as well as full-time students on campus. When I taught a data mining course last year at the Silicon Valley Center, I had students from Apple, Intel, and Yahoo."

Working engineers in Silicon Valley's high-tech industry are a primary target for new degree programs in Technology and Information Management that Zhang is helping to develop at the Baskin School of Engineering. Drawing from traditional fields such as computer science, economics, and business management, Technology and Information Management focuses on the use of information technology for more effective management of businesses.

Zhang's work with SSRC researchers focuses on information retreival from massive databases, such as those maintained by business enterprises. In many respects, Zhang notes, retrieving information from enterprise databases is more challenging than her work on Internet search engines. "Enterprise search problems are pretty difficult. The underlying file system of the database has to be designed to support efficient and effective searching," she says.

The Netflix challenge, however, is relatively straightforward. The contest will award annual prizes to the team whose system delivers the biggest improvement in the accuracy of the recommendations. Winning the grand prize will require at least a 10-percent improvement over the Cinematch system. "It's a nice, well-defined problem," Zhang says. "But the competition is very tough."

Yi Zhang's algorithm has remained among the leaders in Netflix's competition to develop a better system for recommending films to customers.

Jim Mackenzie