Keeping pace with the data explosion

March 3 2014, by Robert W. Fisher

A Lehigh group has developed co-factorization machines that employ mathematical analysis to study how social media users interact with tweets.

With a torrent of new content unleashed on the Internet every hour, how do you find the news articles, status updates and videos you want to view? How do websites like Yahoo and Facebook feed you enough interesting content to make you want to click on the ads?

"We process terabytes of data every hour," says Liangjie Hong '13 Ph.D., a scientist at Yahoo! Labs. "You cannot consume it all."

"If you're really engaged, you have too many people to keep up with," agrees Brian Davison, associate professor of computer science and engineering and head of Lehigh's Web Understanding, Modeling and Evaluation (WUME) laboratory. Davison himself follows hundreds of people on sites like Facebook, where he is spending the 2013-14 academic year on sabbatical in the data science group.

Davison and Hong have collaborated on an innovative project that attempts to discern users' behavior from a small sample of online activity and then to predict the types of content users would like to see.

"If we can better understand what you are interested in," says Davison, "we can decide what to filter, rank higher or flag for your attention."

Who retweets what?

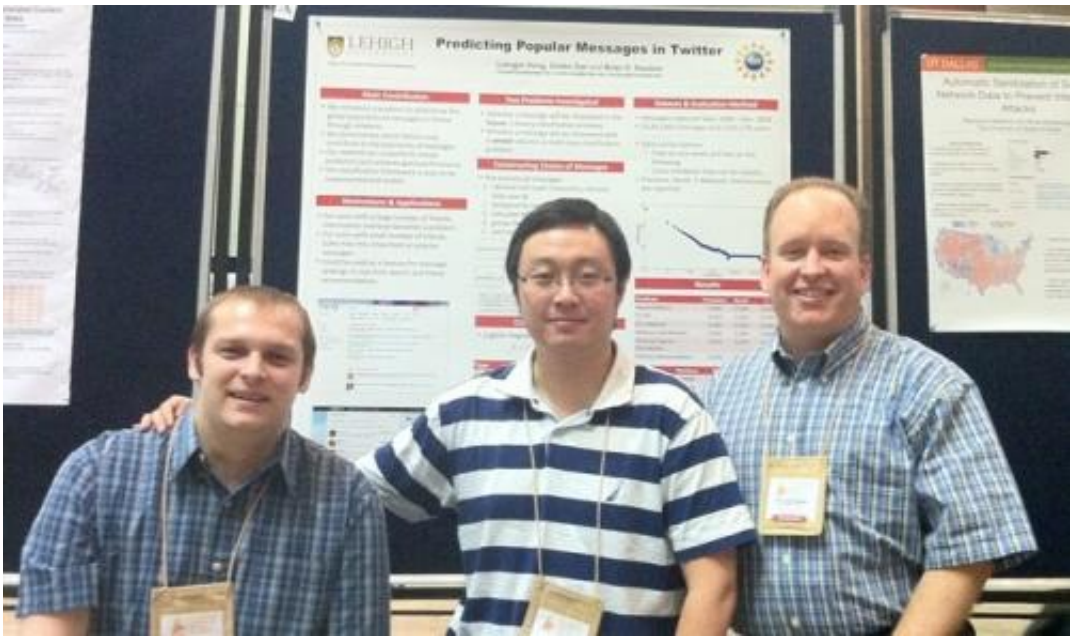
The two researchers analyzed a spurt of Twitter activity, including millions of tweets posted by thousands of individual users. Then they trained an algorithm to predict with high accuracy how often the recipients of tweets would "retweet," or rebroadcast, the messages to their own followers. They received a best poster paper at the 2011 World Wide Web Conference and then decided they could obtain more relevant information by modeling how individual users respond to new content.

"If we could record a user's activities for 24 hours," says Hong, "we would know exactly what they are looking for."

Even power users, however, leave only a handful of clicks to analyze. So Davison and Hong turned their focus to the likelihood that an individual user will pass along a specific message. They developed co-factorization machines, which use a mathematical analysis method to examine how users interact with tweets. Do they reply? Do they retweet? Which tweets do they mark as favorites? The researchers' technique also

determines users' possible interests based, among other things, on the frequency with which specific terms appear in their feeds.

"People have patterns for retweeting," Hong says. Some never retweet. Others primarily pass along tweets from prominent users. Others focus on topics they're passionate about, such as a sports team or an entertainer.



Ph.D. candidate left Ovidiu Dan (left) joined Liangjie Hong (center) and Brian Davison at the 2011 World Wide Web Conference, where the trio received a best poster paper award.

Finding patterns in terabytes of data is a huge challenge, Hong says, but mining data flows for individual patterns can simplify the effort. A particular user is interested in only a tiny fraction of the topics covered on the Internet, he says. This reduces the amount of data that researchers have to process and enables them to make narrow predictions about

users' behavior.

Training machines to learn

The algorithms developed by Davison and Hong improve their predictions through a technique known as machine learning. Rather than explicitly programming rules for how users will respond to tweets, the algorithms "learn" from the results of past interactions to build and refine rules for individual users. This research was a finalist for the best paper award at ACM's (the Association for Computing Machinery) Sixth International Conference on Web Search and Data Mining (WSDM) in Rome in 2013.

The practical approach of the WUME lab prepared Hong well for his role in industry.

"Professor Davison understands that we don't do research in an ivory tower; we have to go outside to see real-world problems," Hong says. Internships at Yahoo Labs and LinkedIn helped him make connections, he says. And the opportunity to use the WUME lab's Hadoop parallel processing environment, at a time when few universities had such facilities, was invaluable.

The quest to model an individual's interests is as old as the card catalog, Davison says, "and it's not going away."

Davison predicts algorithms will soon customize a social media user's experience, ensuring they see every post by a family member while filtering an international colleague's stream so that only English messages are visible, or offering insights on favorite topics from sources the user has yet to hear of.

At Yahoo! Labs, says Hong, scientists run real-time experiments every

day, varying the arrangement and selection of millions of news articles, search results, Tumblr posts, and Flickr photos. A click you made months ago can provide clues that help match today's content with your interests.

"From a business point of view," he says, "if we can provide personalized content, we can make more money from ads."

Davison recognizes that increasing customization can exacerbate the "filter bubble" effect, in which people get information from an echo chamber of like-minded sources. With funding from a Lehigh Faculty Innovation Grant, he is studying how people perceive bias in online news.

"News outlets can tell the same story from very different perspectives," he says. "We want to see if people can recognize different types of bias," such as a political or gender orientation, a lack of objectivity or a negative approach.

Eventually, he plans to tackle the subtle problem of bias by omission, an example of which is when a news outlet never criticizes its corporate parent. Someday you may be able to set a "bot" on your system that will seek out articles that cover your favorite topics with a different slant, he says.

In the meantime, web companies will continue to look for patterns in the data you view and post online, while relying on human nature to fill gaps in the technology.

"How users respond to information and how they decide whether or not to pass it along," says Hong, "is usually based on a mixture of personalization and popularity. We need to offer you the information that is most relevant to you while providing the popular stuff like Kim

Kardashian and Justin Bieber.

"This usually works better than just providing purely personalized solutions."

Provided by Lehigh University