UC San Diego

UC San Diego News Center

September 03, 2014 | By Doug Ramsey

UC San Diego Alumna Wins Best Industry & Government Paper Award

At same conference, UCSD Ph.D. student tracks storefronts behind illicit e-commerce

Over 2,000 people attended the 20th ACM SIG International Conference on Knowledge Discovery and Data Mining (KDD 2014), a premier interdisciplinary conference that brought together researchers and practitioners in late August from data science, data mining, knowledge discovery, large-scale data analytics, and big data. Best paper awards were handed out to academic and industry papers, and this year's Industry & Government award went to University of California, San Diego alumna Diane Hu (M.S. '09, Ph.D. '12), who did her doctoral dissertation under Computer Science and Engineering professor Lawrence Saul.

The UC San Diego alum was honored for the work she did with two co-authors on the paper, "Style in the Long Tail: Discovering Unique Interests with Latent Variable Models in Large Scale Social E-commerce." Hu and her co-authors, Rob Hall and Josh Attenberg, all work at Etsy, Inc., the e-commerce website that bills itself as "the world's most vibrant" marketplace for handmade or vintage items and supplies. Etsy also attracts software developers like Hu with its slogan, "We believe in code as craft."

UC San Diego alumna Diane Hu, data scientist at Etsy, Inc. (Photo by Karen Kristian for Career Contessa)

In the award-winning paper, Brooklyn-based Etsy data scientist Hu and her colleagues tackled the challenge of matching buyers to products "as the size and diversity of the marketplace increases." With over 30 million diverse product listings on Etsy, the company must deal with the problem of capturing shoppers' aesthetic preferences in order to steer them to items that fit their often-eclectic styles. In her talk presenting the award-winning paper, Hu described the methods and experiments underlying two new style-based recommendation systems they developed to serve Etsy buyers. One is called Latent Dirichlet Allocation (LDA), which seeks out 'trending' categories and styles on Etsy. The trends are then factored into the user's "interest" profile. Hu and her colleagues also explored hashing methods to perform fast-nearest-neighbor search on a map-reduce framework, in order to efficiently obtain recommendations. "These techniques have been implemented successfully at very large scale," explained Hu, "substantially improving many key business metrics." In other words, better recommendations translated into stronger sales and happier customers.

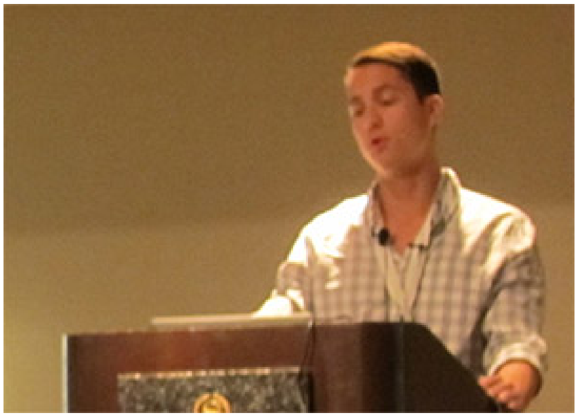
Prior to graduating in 2012, Diane Hu did an internship at Yahoo! Labs. "It was my first time applying my academic training in an industry setting," she recalled in a Q&A for the Career Contessa website recently. "The data we worked with was a lot larger in scale and messier than the toy datasets used in academia, which resulted in having to learn a lot of new concepts and tools in order to apply even the simplest algorithms to the datasets."

"I chose to study machine learning because of its far-reaching impact on many types of Web technologies. It is used in all sorts of applications, ranging from predictive analysis (will housing prices go up?) to classification problems (is this article about politics or sports?) to recommendation systems (what kind of products does this user like that we can recommend to them?), and many more."

"After I finished my Ph.D.," added Hu, who is a photographer in her spare time, "I wanted to put my skills and knowledge to use on real products. The data science role is perfect for this, as I get to apply my knowledge of machine learning to creating features that are used by millions of users that stop by the Etsy website everyday."

Knock It Off

While an alumna of UC San Diego's Computer Science and Engineering (CSE) department took the top industry honor at KDD 2014, CSE faculty and a graduate student were representing UC San Diego at the conference in New York City. 5th-year Ph.D. student Matthew Der (M.S. '13, Ph.D. '15 expected) collaborated on a paper with his three doctoral advisors: CSE professors Lawrence Saul, Stefan Savage and Geoffrey Voelker. Their paper had a catchy title: "Knock It Off: Profiling the Online



Computer science and engineering Ph.D. student Matthew Der delivers his presentation, "Knock It Off", at KDD 2014.

Storefronts of Counterfeit Merchandise." The team's approach was to extract features that reveal when Web pages linked to the same affiliate program share a similar underlying structure. The features were mined initially from a small seed of labeled data, and according to the paper, that data allowed the researchers to "profile the Web sites of 44 distinct affiliate programs that account, collectively, for hundreds of millions of dollars in illicit e-commerce." The researchers also noted "several broad challenges in the large-scale empirical study of malicious activity" on the Internet.

After delivering his presentation, Der had to rush back to San Diego in time to teach CSE 150, Introduction to Artificial Intelligence, which he has been teaching during the second summer quarter, which ends this week on Friday, September 5.

MEDIA CONTACT

Doug Ramsey, , dramsey@ucsd.edu

UC San Diego's Studio Ten 300 offers radio and television connections for media interviews with our faculty, which can be coordinated via studio@ucsd.edu. To connect with a UC San Diego faculty expert on relevant issues and trending news stories, visit https://ucsdnews.ucsd.edu/media-resources/faculty-experts.