Researchers say readers' identities can reveal much about content of articles

August 12 2013

Articles that people share on social networks can reveal a lot about those readers, research has shown. But a new Carnegie Mellon University study reverses the proposition, asking the question: What can be learned about an article from the attributes of its readers?

To find out, the CMU researchers, along with colleagues at the University of Washington, analyzed almost 3 million news articles and the public profiles of the people who shared those articles on Twitter. This enabled them to generate a few thousand "badges" that characterized the content of the shared news articles and also could be used to analyze any subsequent article, including those that had never been shared or even read.

"Because these badges are based on the readers' self-described interests, rather than only on the words in each article, we found that the badges provide a consistent, reliable means of representing content," said Khalid El-Arini, a recent Ph.D. graduate of the CMU Computer Science Department. "For instance, while what it means to be 'liberal' changes from month to month, there will always be people who describe themselves as 'liberal' on Twitter, allowing us to produce a direct correspondence between 'liberal' topics from different periods of time."

This approach could thus be used in recommendation systems, but also can provide interesting insights about writers and the state of political discourse.

El-Arini will present the findings today at the Conference on Knowledge Discovery and Data Mining (KDD 2013) in Chicago. Other investigators were Min Xu, a Ph.D. student in CMU's Machine Learning Department; Emily Fox, assistant professor of statistics at the University of Washington, and Carlos Guestrin, professor of computer science and engineering at Washington.

In order to train their model, the team began by looking at three months of tweets—from September of 2010, 2011 and 2012—and selecting those that included links to mainstream news articles and came from a user who had filled out a Twitter profile.

Each news article was then downloaded and the most meaningful, unique words were extracted, creating a "bag of words" for each article; similar to a visual word cloud, these bags give greater weight to more important words. Likewise, from each user's Twitter profile, a set of descriptive words, or badges, was extracted.

By comparing the bags of words with badges from the people who shared the articles, the researchers were able to create a dictionary that associated each badge with its characteristic words. For example, people who self-identify with the music badge in their profiles are likely to share articles with words such as "band," "album" and "song." Different dictionaries were created for each year to compensate for interests or topics that change over time. These dictionaries were then used to encode new articles, leading to a document representation based on attributes of potential readers.

"It is important to point out that the badge dictionaries can be used to encode articles never before seen or shared," El-Arini said. "We take the content of the articles, and use our model to predict which types of readers would be likely to share them."

The researchers included a case study that used badges to explore the readership of prominent political columnists. This method showed, not surprisingly, that New York Times columnist Maureen Dowd had readers who tended to be progressive. This association was notable because Dowd never explicitly uses the word "progressive" in the articles analyzed by the researchers. Rather, the algorithm detected that the words Dowd uses in these articles correspond to the type of content self-described progressives tend to share on Twitter.

El-Arini, now a research scientist at Facebook, said the algorithm the team designed also takes into account relationships between badges. The system understands, for instance, that "liberal" is similar to "progressive" and "school" is associated with "student."

"Our methodology leads to thematically coherent topics that are more consistent over time than popular alternative approaches," El-Arini said. "We believe this method will lead to better performance on personalization tasks."

Provided by Carnegie Mellon University