Researchers 'text mine' The New York Times, demonstrating ease of new technology

July 26 2006

Performing what a team of dedicated and bleary-eyed newspaper librarians would need months to do, scientists at UC Irvine have used an up-and-coming technology to complete in hours a complex topic analysis of 330,000 stories published primarily by The New York Times.

The demonstration is significant because it is one of the earliest showing that an extremely efficient, yet very complicated, technology called text mining is on the brink of becoming a tool useful to more than highly trained computer programmers and homeland security experts.

"We have shown in a very practical way how a new text mining technique makes understanding huge volumes of text quicker and easier," said David Newman, a computer scientist in the Donald Bren School of Information and Computer Sciences at UCI. "To put it simply, text mining has made an evolutionary jump. In just a few short years, it could become a common and useful tool for everyone from medical doctors to advertisers; publishers to politicians."

Text mining allows a computer to extract useful information from unstructured text. Until recently, text mining required a great deal of preparation before documents could be analyzed in a meaningful way. A new text-mining technique called "topic modeling" -- which UCI scientists used in their New York Times experiment -- looks for patterns of words that tend to occur together in documents, then automatically categorizes those words into topics -- all with minimal human effort.

UCI researchers didn't invent topic modeling, but they developed a technique that allows the technology to be used on huge document collections. They also are among the first to demonstrate its ease and effectiveness by applying it to a newspaper archive. The results reveal few surprises, but the application demonstrates the ability of topic modeling to spot trends and make connections in a way that could be applied to more complicated and cumbersome documents such as those used by medical researchers and lawyers.

Newman and UCI researchers Padhraic Smyth, Mark Steyvers and Chaitanya Chemudugunta presented their research at the recent Intelligence and Security Informatics conference in San Diego.

The topic model, applied to the collection of news articles published from 2000 to 2002, identified patterns of words that occurred together in the stories. From those words, researchers were able to identify topics. Information associated with those topics was charted over time, allowing the scientists to pinpoint what months of the year certain topics were most in the news and how much ink they received from year to year.

For example, the model generated a list of words that included "rider," "bike," "race," "Lance Armstrong" and "Jan Ullrich." From this, researchers were easily able to identify that topic as the Tour de France. By examining the probability of words appearing in stories about the Tour de France, researchers learned that Armstrong was written about seven times as much as Ullrich. Charting information over time, researchers discovered that discussion of Tour de France peaked in the summer months but decreased slightly year to year.

"If I were interested in advertising a product related to the Tour de France, I might want to know whether interest in the Tour de France is increasing or decreasing," Newman said. "This might be very important knowledge."

Including the Tour de France, the model automatically identified a total of 400 topics ranging from renting apartments in Brooklyn and diving in Hawaii to voting irregularities and dinosaur bones. As for newsmakers, topics included Tiger Woods, Elian Gonzalez, Denzel Washington and Barbie.

"Text mining is an incredible tool," Newman said. "It already allows a doctor to identify the common thread in old and new medical research. With topic modeling, connections can be drawn faster and more efficiently in large volumes of text."

About topic modeling: UCI researchers performed their experiment using a statistical topic model based on a text model developed at UC Berkeley in 2003. Thanks to an improved solution technique proposed by Mark Steyvers and a research partner, this model has advanced from academic use to something that is now widely used in the research community. Topic modeling looks for patterns of words that tend to occur together in documents, then automatically categorizes those words into topics. Older text-mining techniques require the user to come up with an appropriate set of topic categories and manually find hundreds to thousands of example documents for each category. This human-intensive process is called supervised learning. In contrast, topic modeling, a type of unsupervised learning, doesn't need suggestions for an appropriate set of topic categories or human-found example documents. This makes retrieving information easier and quicker.

Source: University of California - Irvine

Citation: Researchers 'text mine' The New York Times, demonstrating ease of new technology (2006, July 26) retrieved 26 April 2024 from https://phys.org/news/2006-07-text-york-ease-technology.html