Skysong Innovations
Advancing the Arizona State University Knowledge Enterprise

1475 N. Scottsdale Road, Suite 200
Scottsdale, AZ 85287-3538
Phone: 480 884 1996 Fax: 480 884 1984

Novel Clustering Algorithms for Improved Twitter Dataset Creation

Case ID:M16-049P^
Published: 2/26/2020

Inventors

Justin Sampson
Fred Morstatter
Ross Maciejewski
Huan Liu

Contact

Shen Yan
shen.yan@skysonginnovations.com

Social media services like Twitter have become a prominent source of research data for both academic and corporate applications because of the large amount of data available. Twitter allows researchers to obtain low-cost data samples, but past research has shown the data is not a random distribution and thus, is unrepresentative of the population. As a result, the pool of data produced by Twitter has significant bias, leading researchers to form incorrect conclusions. To address the inaccurate representation of the data, scientists are now looking to create more reliable samples by using computer algorithms to recover missed or hidden data filtered by Twitter.

Researchers at ASU have developed methods to increase coverage when gathering data from Twitter by structuring searches using a dynamic-word co-occurrence network. The network comprise clustering algorithms that combine keywords to produce multiple streams of data for comparison. The mechanisms help mitigate inaccuracy of data due to overlap while still producing more data points, ultimately increasing reliability for analysis. Overall, the methods generate more data than the original Twitter sample, giving a set of data with reduced bias and increased quantity all at a lower cost.

Potential Applications

- Social Media Analytics and Monitoring
- Data Mining and Storage
- National Security/Intelligence
- Automated Advertising and Targeted Marketing
- Machine Learning Models

Benefits and Advantages

- Increased Quantity – The clustering methods used produce an increased number of data points than the original Twitter sample
- More Effective – The techniques provide developers with a more representative set of data for a population
- Lower Cost – The methods require very few resources and thus increased savings compared to the explicit cost of purchasing the complete data

For more information about the inventor(s) and their research, please see:

Dr. Huan Liu's directory webpage

Dr. Ross Maciejewski's directory webpage