Extracting Hidden Concepts from Twitter Data for Use in Time Series Prediction

The current research aims to model human behavior by extracting hidden concepts from data from the social media site, Twitter. Twitter allows its 500 million users 140 characters to display their mood, thoughts, activities, likes, and dislikes. By analyzing clusters of highly correlated words and their respective time series, subjects once thought to be strictly qualitative can be analyzed quantitatively. Time series of hourly word usage frequencies are correlated on the daily and weekly scale using both spearman correlation coefficients and cosine distance. These correlations are then analyzed using clustering algorithms such as k-means, singular value decomposition, and latent semantic analysis to uncover patterns in the word frequency data. For example, the words we typically use during the day may differ from the words we typically use at night. Since data obtained from Twitter can be a very accurate indicator of human behavior, there is potential for this data to have strong predictive power. Data from social media sites has already been proven to be predictive in a variety of areas. An area that has yet to be explored using this data is energy consumption. Since individuals often tweet about what they have done, are doing, and will do, there is potential for this data to be used to predict energy usage. The word clustering algorithms are applied to the energy demand curve provided by ISO New England to determine relationships between word frequencies and energy consumption. In future research, this relationship will be developed into a model that can predict energy consumption using word frequencies from Twitter data. This model can then be incorporated into a load forecasting model used by utility companies.