Abstract No: 311 ***ICT, Mathematics and Statistics***

CLUSTERING ENGLISH NEWS ARTICLES BASED ON RELEVANT DOMAINS: COMPARATIVE STUDY USING THREE CLUSTERING ALGORITHMS

N. Disayiram[1] and R.A.H.M. Rupasingha[2*]

[1]Department of Computing and Information Systems, Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka
[2]Department of Economics and Statistics, Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka

The news tells us about what happens around us. Nowadays, people use news sites to read exciting news. News has many categories. The preferable choice of the news category differs for each newsreader. In the end, every news category is important. Every day lots of news is published on news websites. Typically, news sites categorize the news, but all the categories are not included on that site. Most news sites prioritise some categories, and other categories get lower media coverage. It is, therefore, difficult to find the relevant types of news. These problems give complexity to the newsreaders and relevant content seekers to find the relevant section on the news sites. The clustering of English news based on the relative category gives solutions to overcome those problems. This study aims to cluster news articles based on the relevant domain using machine-learning algorithms. We consider five domains: politics, sports, health, technology, and business. The online collected data was converted into vector format by using the term frequency-inverse document frequency vectorization. Then, the three clustering algorithms: Expectation Maximization, Simple Kmeans, and Hierarchical Clustering based on agglomerative technique, were separately applied to the body of the news and the news headline. The accuracy is calculated through the classes to clusters evaluation model in the WEKA tool. The results show that the Expectation Maximization algorithm achieved the highest accuracy of 87.9%, while it was 83.8% for the Simple Kmeans algorithm. Further, the Hierarchical Clustering method achieved the minimum accuracy results. The comparison results between the heading of news and the body of news show that the body of news performed better than the heading of news to cluster the news articles.

Keywords: Clustering, Domain, Machine learning, News article