Paper No: SC-17 **Smart Computing**

A Comparative Study of Clustering English News Articles Using Clustering Algorithms

N. Disayiram
Department of Computing and Information Systems
Sabaragamuwa University of Sri Lanka, Sri Lanka
dshajiram@gmail.com

R. A. H. M. Rupasingha
Department of Economics and Statistics
Sabaragamuwa University of Sri Lanka, Sri Lanka
hmrupasingha@gmail.com

***Abstract* - The news informs us of what is going on in the world. People nowadays read their interesting news on news websites. There are numerous categories of news. Eachnewsreader has a different preference for news categories. Sportspeople prioritize sports news, whereas technology fans pay attention to the technology segment of the news. At the endof the day, each news category is important. Every day, a large amount of information is released on news websites. News sites usually categorize the news however, not all of the categories arepublished on those sites. Some categories are given higher attention by news outlets, while others receive less coverage. Asa result, finding an appropriate category of news is tough. Theseissues make it difficult for newsreaders and content seekers to find relevant sections on news websites. The clustering of English news articles by relative category provides solutions to these issues. This research aims to use clustering algorithms to cluster news articles depending on the relevant domain/cluster.We consider five news categories: politics, sports, health, technology, and business. The data collected online was converted into a vector format using the term frequency-inversedocument frequency (TF-IDF) vectorization. Then, on the bodyof the news and the news heading, the three clustering algorithms: Expectation-Maximization (EM), Simple K-means,and Hierarchical Clustering based on an agglomerative approach were applied individually. The Waikato Environmentfor Knowledge Analysis (WEKA) tool's classes to clusters evaluation model are used to calculate the accuracy. The EM method had the maximum accuracy of 88.5% with the best results in terms of correctly clustered instances. The comparisonbetween the heading of news and the body of news demonstratesthat the body of news clustered the news items better than the heading of news.**

Keywords - clustering, domain, Machine Learning, news article