String Vector based AHC for Text Clustering

Taeho Jo

Department of Liberal Art, Hongik University, 2639 Sejongro, Sejong, South Korea

tjo018@hongik.ac.kr

***Abstract*—In this research, we propose the string vector based version of AHC algorithm as the approach to the text clustering. Using the traditional version leads to the three main problems: huge dimensionality, sparse distribution, poor transparency, since texts need to be encoded into numerical vectors. In order to solve the problems, in this research, we encode texts into string vectors, define the similarity measure between them, and modify the AHC algorithm into the version where a string vector is given as its input. As the benefits from this research, we expect the better performance, the more compact representation, and the better transparency. Hence, this research is intended to improve the text clustering performance, by solving the problems..**

***Keyword*—Text Clustering, String Vector based AHC, Semantic Operation**

Taeho Jo (M'97–AM'12) This author became a Member (M) of IEEE in 1997, and an Associate Member (AM) in 2012. He was born in 1970, South Korea. He received his Bachelor degree from Korea University in 1994, his Master degree from Pohang University of Science and Technology in 1997, and his PhD degree from University of Ottawa in 2006. His research area spans mainly over text mining, neural networks, machine learning, and information retrieval. He has the four year experience of working for industrial organizations and ten year experience of working for academic ones. Now, he has published more than 100 research papers in peered journals and proceedings, and he is working for Hongik University as a faculty member.