String Vector based KNN for Text Categorization

Taeho Jo

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

tjo018@hongik.ac.kr

***Abstract*—This research proposes the string vector based version of the KNN as the approach to the text categorization. Traditionally, texts should be encoded into numerical vectors for using the traditional version of KNN, and encoding so leads to the three main problems: huge dimensionality, sparse distribution, and poor transparency. In order to solve the problems, in this research, texts are encoded into string vectors, instead of numerical vectors, the similarity measure between string vectors is defined, and the KNN is modified into the version where string vector is given its input. As the benefits from this research, we may expect the better performance, more compact representation of each text, and better transparency. The goal of this research is to improve the text categorization performance by solving them..**

***Keyword*—Text Categorization, String Vector based KNN, 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.