Improving K Nearest Neighbor into String Vector Version for Text Categorization

Taeho Jo

School of Game, Hongik University, 2639 Sejongro Sejong South Korea 30016

tjo018@hongik.ac.kr

***(Pt9)Abstract*— This research is concerned with the string vector based version of the KNN which is the approach to the text categorization. Traditionally, texts have been 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, this research propose that texts should be encoded into string vectors the similarity measure between string vectors is defined, and the KNN is modified into the version where string vector is given its input. The proposed KNN version is validated empirically by comparing it with the traditional KNN version on the three collections: NewsPage.com, Opiniopsis, and 20NewsGroups. The goal of this research is to improve the text categorization performance by solving them.**

***Keyword*— String Vector, K Nearest Neighbor, Text Categorization**

Taeho Jo works currently as a faculty member in Hongik University, 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. Recently, he is awarded in the world wide biography dictionary, Marquis Who's Who in the World, two times in 2016 and 2018.