Clustering suggestion for Chinese news web pages from multi-media sources

邱登裕,潘雅真
Information Management
Computer Science and Informatics
chiuden@chu.edu.tw

Abstract

There exists some news obviously classified into incorrect categories on Chinese web pages portals. This phenomenon is owing mainly the difficulty in automatically classifying Chinese news and the fact that news appearing on web page portals is retrieved from numerous media sources. This study integrates the genetic algorithm and multi-class support vector machine (SVM) classifiers to construct a Chinese news classification method. This study finds that some similar documents are scattered different categories. The main reason for this phenomenon may be that the categories of original media sources differ from those of news web page portals. Similar news thus should be gathered together to form a new category. This study attempts to combine genetic and fuzzy c-means algorithms to propose a new approach to offer clustering suggestions for news web pages that are scattered different categories and derive from multi-media sources.

Keyword： Chinese web pages clustering, multi-class SVM, fuzzy c-means algorithm, genetic algorithm