GA-based keyword selection for the design of an intelligent web document search system

周智勳, 李建興, 陳雅惠
Computer Science & Information Engineering
Computer Science and Informatics
chlee@chu.edu.tw

Abstract

The main steps for designing an automatic document classification system include feature extraction
and classification. In this article a method to improve feature extraction is proposed. In this
method, genetic algorithm was applied to determine the threshold values of four criteria for extracting
the representative keywords for each class. The purpose of these four threshold values is to
extract as few representative keywords as possible. This keyword extraction method was combined
with two classification algorithms, vector space model and support vector machine, for examining
the performance of the proposed classification system under various extracting conditions.

Keyword：web document classification; keyword extraction; genetic algorithm; vector space method;