

ABSTRACT

The biggest problem in the classification process is the amount of web pages available features. The features are a token or a word that appears on the web page. The problem can be addressed using feature selection techniques to obtain optimal feature. Feature selection techniques in this study using genetic algorithms and sequential minimal optimization. Genetic algorithms are used to select features, while the sequential minimal optimization is used to evaluate it. Classifier algorithms used in this study is Naïve Bayes and Decision Tree (J48). The process of feature selection using the proposed algorithm is able to select features of the dataset WebKb 45.20% and 42.73% in the dataset Routers. The average value generated the highest accuracy after implementing the proposed method is 89.33% algorithmically J48. Overall, feature selection techniques can improve the accuracy of the average value of the acquisition resulting in the classification process web pages.

Keywords – Feature selection, genetic algorithm, SMO, web page classification, wrapper.

INTISARI

Permasalahan terbesar dalam proses klasifikasi halaman web adalah besarnya fitur yang tersedia. Fitur yang dimaksud merupakan sebuah token atau kata yang muncul di halaman web. Masalah tersebut dapat ditangani menggunakan teknik seleksi fitur untuk mendapatkan fitur yang optimal. Teknik seleksi fitur pada penelitian ini menggunakan algoritme genetika dan sequential minimal optimization. Algoritme genetika digunakan untuk memilih fitur, sedangkan sequential minimal optimization digunakan untuk mengevaluasinya. Algoritme pengklasifikasi yang digunakan pada penelitian ini adalah NaiveBayes dan Decision Tree (J48). Proses seleksi fitur menggunakan algoritme yang diusulkan mampu menyeleksi fitur sebesar 45,20% pada dataset WebKb dan 42,73% pada dataset Routers. Nilai rerata akurasi tertinggi yang dihasilkan setelah menerapkan metode usulan adalah 89,33% menggunakan algoritme J48. Secara keseluruhan, teknik seleksi fitur dapat meningkatkan perolehan nilai rerata akurasi yang dihasilkan pada proses klasifikasi halaman web.

Kata Kunci : Algoritme genetika, klasifikasi halaman web, seleksi fitur, SMO, wrapper.