Abstract

The web is a huge repository of information which needs for accurate automated classifiers for Web pages to maintain Web directories and to increase search engines’ performance. In web page classification problem each term in each HTML/XML tag of each Web page can be taken as a feature, an efficient methods to select best features to reduce feature space of the Web page classification problem derived here. Classification of Web page content is essential to many tasks in Web information retrieval such as maintaining, web directories and focused crawling. The uncontrolled nature of Web content presents additional challenges to Web page classification as compared to traditional text classification, but the interconnected nature of hypertext also provides features that can assist the process. As in derived work reviewed in Web page classification, the importance of these Web-specific features and algorithms, describe state-of-the-art practices, and track the underlying assumptions behind the use of information from neighboring pages. This work, our aimed to optimize best features selection for Web page classification problem. Since Firefly Algorithm (FA) is a recent nature inspired optimization algorithm, that simulates the flash pattern and characteristics of fireflies. Clustering
is a popular data analysis technique to identify homogeneous groups of objects based on the values of their attributes. Here FA is used for clustering on benchmark problems which is being found more suitable than Artificial Bee Colony (ABC), Particle Swarm Optimization (PSO), and other nine methods used. The web page optimization using Naïve Bayes classifier (WPCNB) is an improved optimized web page classification using firefly algorithm with NB classifier. this work is tested on research banking data set where firefly algorithm used for web optimization and Naïve Bayes (NB) classifier used for classification of pages in contrast to selected pages with reference to different fireflies. The entitled work is being found better in terms of feature measure (FM), accuracy, precision etc. parameters with respect to existing key concepts. it is also an search optimization approach and can be enhanced by different genetic algorithm (GA) based classifiers use in future.

References

3. Xin-She Yang, Xingshi He, "Firefly Algorithm: Recent Advances and Applications", School of Science, Xi'an Polytechnic University, Vol. 1, issue 1, 2013.
16. Jie Chen, Jian Li, Hao Liao, Qingsheng Yuan, Xiuguo Bao, "Study on Meaningful String Extraction Algorithm for Improving Webpage Classification" IEEE, 2011..

Index Terms

Computer Science
Algorithms

Keywords

Feature, Navie Bayes classifier, j48, f- measure.