ABSTRACT:

In this paper, we propose a hybrid approach of Arabic scripts web page language identification based on decision tree and ARTMAP approaches. We use the decision tree approach to find the general identities of a web document, be it an Arabic script-based or a non-Arabic-based. Then, we use the selected representations of identified pages from the decision tree approach as an input to the ARTMAP neural network for further verification of the diversity of languages detected by the algorithm. From our initial experiments, we found that, although the decision tree approach may achieve a higher accuracy than ARTMAP, the former may not be as reliable as the ARTMAP approach if the language used is extended to other types of Arabic script web documents in different languages (e.g., Urdu, Arabic, Persian, etc.). Therefore, we propose this hybrid decision tree-ARTMAP approach in order to improve the performance of the Arabic script language identification on web documents in a variety of languages. The result shows that the proposed approach has outperformed both decision tree and the default ARTMAP approaches.