Abstract

Bangla document categorization is an emergent topic now-a-days. Every document has some keywords that reflect its category. Document categorization refers to an automatic categorization of a document based on the keywords it contains. An expedient keyword selection method is necessary to correctly classify a document. TF-IDF [1], Naive Bayes [2][3][4], KNN [5] are some of the trending methods used in Document Categorization. Some of models are also used in Bangla Document Categorization. In this research, Term Graph concept was mainly focused. TGM [5] is never used before for Bangla document categorization. So, the concentration was Term Graph concept mixing with other existing models for categorizing Bangla documents. Experiments are also performed by changing and tuning feature selection method. Maximum 3-size subsets are used in experiment. Features were selected by changing selecting formula. Sometime all features were selected and sometime less important features were removed for increasing accuracy and reducing space complexity.

References

Index Terms

Computer Science  Information Sciences

Keywords

TF-IDF, TGM, KNN, SVM, NB.