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.
1. Y. h. Lu and Y. Huang, “Document categorization with entropy based tf/idf classifier,” vol. 4, pp. 269–273, May 2009.
2. Y. Wang, J. Hodges, and B. Tang, “Classification of web documents using a naive bayes method,” pp. 560–564, 12 2003.
3. Y. Matsuo and M. Ishizuka, “Keyword extraction from a single document using word co-occurrence statistical information,” vol. 13, 03 2003.
4. M. EL KOURDI, A. BENSAID, and T.-e. Rachidi, “Automatic arabic document categorization based on the naïve bayes algorithm,” 08 2004.
5. V. Bijalwan, P. Kumari, J. Espada, and V. Semwal, “Knn based machine learning approach for text and document mining,” vol. 7, 06 2014.
6. M. Alexandrov, A. Gelbukh, and G. Lozovoi, “Chi-square classifier for document categorization,” pp. 457–459, 02 2001.
7. C.-h. Chan, A. Sun, and E.-P. Lim, “Automated online news classification with personalization,” 03 2002.
8. A. Mesleh, “Support vector machines based arabic language text classification system: Feature selection comparative study.” pp. 11–16, 01 2007.
9. Y. Wang and Z.-O. Wang, “A fast knn algorithm for text categorization,” vol. 6, pp. 3436–3441, 09 2007.
10. M. S. Islam, F. Elahi, and S. Ikhtiar Ahmed, “A support vector machine mixed with tf-idf algorithm to categorize bengali document,” 04 2017.
11. S. Weiss, C. Apte, F. Damerau, and S. Weiss, “Text mining with decision trees and decision rules,” 10 1999.
12. Z. Chen, C. Ni, and Y. L. Murphey, “Neural network approaches for text document categorization,” pp. 1054–1060, 2006.
13. P. Soucy and G. Mineau, “A simple knn algorithm for text categorization,” pp. 647–648, 02 2001.
14. H. Al-Mubaid and S. A. Umair, “A new text categorization technique using distributional clustering and learning logic,” 2006

**Index Terms**

Computer Science Information Sciences

**Keywords**

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