Abstract

Text categorization aims to define the categories of documents whose category is unknown. One algorithm in text categorization is multinomial naïve Bayes algorithm which known to have a very simple way of working, effective and has a good performance. In the previous studies, Text categorization reached the stage of categorization only. In this study, the binary file categorization of document will be saved into the database so that data processing can be done easily. One of NoSQL database type is document-oriented and one of the document-oriented DBMS is MongoDB. MongoDB has GridFS and sharding features which can store binary file distribution into multiple machines. By placing data accross multiple machines allows to store more data and handle more loads without the need of powerful machine. From the results of tests performed, text categorization performance value is above 88% in the use of 756 training data and 84 test data. For document oriented database, the best results were obtained with a value of response time 1,713 seconds and throughput of 130,869 transactions per second.

Keywords : Text categorization, NoSQL, document oriented database, GridFS, sharding