AUTOMATIC INDEXING FOR DOCUMENTS WRITTEN IN BAHASA INDONESIA USING LUCENE FRAMEWORK

Aditya Wahyu Baskoro, Julio Adisantoso, Firman Ardiansyah

Department of Computer Science

Faculty of Mathematics and Natural Sciences

Bogor Agricultural University

julioipb@gmail.com

2010

Abstract

Indexing is a basic step in information retrieval fields that has important roles to provide rapid access during the search process. The purpose of this research is implementing automatic indexing process for documents written in Bahasa Indonesia using Lucene framewok, including the use of stopwords, stemming, and both. All unique terms are weighted using tf-idf function provided by Lucene. Lucene is an open source framework that performs indexing, searching, and partial text analysis. System performance tested with 2000 documents to measure indexing time and 1000 documents to measure performance of retrieval results. In general, performance of the system measured by average precision is good, because 60,35% average retrieval results are relevant to given queries. In general the use of stopwords does not give a significant impact on system performance but stemming does. Stopwords and stemming are useful to reduce the size of index.