Abstract

Capturing the semantic relationships between words in a document representation is a difficult problem. Latent Semantic Indexing (LSI) algorithm is one of the best-known dimension reduction. In LSI, documents are indexed by using latent semantic concept. LSI showed a large performance improvements over the TF-IDF representation on small document collections but often do not perform well in large heterogeneous document collections. LSI maps all words to the dimension of the matrix. The greater the greater the number of documents that formed the matrix dimension. In addition, numerical information and documents that may be abbreviations excellent indicator of the topic is no longer available after using LSI. This is due to the LSI, which includes the vocabulary of all terms other than noun or noun is processed in the same way.

In this final project will analyze the performance of an information retrieval system usingHybrid Document Indexing. This approach was used in the indexing of documents to solve the problems of LSI. Hybrid Document Indexing continue using latent semanticconcept and also try to keep specific documents from the collection. Hybrid DocumentIndexing using a combination of LSI for weighting words that contain a noun and the other noun in the document will be TF-IDF weighting.

Test results from this thesis show that the Hybrid Document Indexing using stemming preprocessing proved to be able to find relevant documents even if it does not contain terms of the input query will still be drawn. In addition, the accuracy of search results using this method produces values precision, recall and F-Measure are above 0.50. In the experiment a few number of datasets, dataset, the greater the amount of time the process of indexing and searching will stay longer. The increase is due to the long process of increasing number of documents it will be even greater dimension in processing LSI plus tf-IDF so that processing time becomes longer.

Keywords: *Information Retrieval, Latent Semantic Indexing, Hybrid Document Indexing*