Abstract

Today, document categorization technique was done using *machine learning* approach which taken example documents as an example for learning process, then the result from learning process will be used as references for categorizing other documents. There exist a symptom that different people uses different words to express the same concept or idea. Because of that, an approach which comparing documents no just from the similarity of the words, but also from the conceptual similarity of the words is needed.

Prototype software using *probabilistic latent semantic indexing* (PLSI) for categorizing documents was built on this final assignment. PLSI is a novel approach from *latent semantic indexing* (LSI). PLSI performance such as effectiveness and efficiency was measured from the software produced.

It was found that resulted performance was influenced by dimension used, the bigger dimension used, then *recall*, *precision,* and decomposition time will also increased and *error* will decreased. Number of data used also influenced resulted performance, but *terms* which used on the data also influenced resulted performance on number of data. From the analysis, it was found that *TFDIF* weighting will make the resulted performance worse than *TF* weighting.

Keywords: document categorization, *probabilistic latent semantic indexing*