Abstrak

Jumlah dokumen yang sangat banyak dan harus ditangani memerlukan sitem pengorganisasian secara otomatis. Pendekatan yang sangat popular untuk mengelompokkan dokumen adalah dengan pemodelan berdasarkan ruang lingkup vektor yang merepresentasikan teks, didapatkan dari sejumlah *term* yang terletak dalam dokumen.

Teknik clustering berdasarkan matriks frekuensi term-dokumen menderita akan adanya *noise* yang diakibatkan penggunaan kata-kata yang berbeda tetapi memiliki arti yang sama. Relasi semantik (sinonim) seperti ini harus diatasi.

Metode yang digunakan dalam tugas akhir ini menggunakan *Latent Semantic Indexing(LSI)* yang dikombinasikan dengan *double clustering* untuk mengurangi dimensi dari ruang vektor. Dengan cara ini, teknik clustering diimplementasikan dalam ruang vektor yang lebih kecil dan *noise* yang berkurang

K-means adalah salah satu analisa klaster yang sederhana. *K-means* tidak dapat mendeteksi *noise*. Penggunaan *K-means* yang ditambahkan dengan LSI dan teknik *double clustering, noise* dapat ditangkap dan dikurangi. Ketika *noise* berkurang, performansi seperti *purity, recall, precision* dan *F-measure* dari K-*means* dapat ditingkatkan.

Kata kunci: *clustering, Latent Semantic Indexing(LSI), K-means*