Abstrak

Saat ini telah banyak dikembangkan teknik klasterisasi, misalnya teknik menggunakan representasi *single-word item*, merepresentasikan dokumen teks sebagai "bag of words" dimana suatu dokumen dipandang sebagai sekumpulan kata-kata. Dalam representasi ini tidak ada urutan antar kata maupun kalimat yang diperhatikan karena setiap kata dianggap berdiri sendiri tanpa ada keterhubungan satu sama lain sehingga tidak tepatnya dalam pelabelan hasil *cluster*.

Permasalahan-permasalahan diatas bisa ditangani dengan menggunakan *Clustering Based On Frequent Word Sequences (CFWS).* Data berdimensi tinggi dapat diatasi dengan mereduksi term-term yang tidak frequent. Pelabelan cluster dilakukan dengan cara menelusuri "word sequences" di tiap dokumen.

Hasil klasterisasi dengan algoritma ini divisualisasikan secara hirarki dalam bentuk *tree.* Berdasarkan pengujian, klaster yang dihasilkan oleh algoritma CFWS ini memiliki kualitas deskripsi klaster mewakili isi berita.

Kata kunci: *clustering, frequent word sequences, CFWS, F-Measure, purity.*