Abstrak

Menumpuknya data khususnya dokumen berita berbahasa Indonesia merupakan salah satu hal yang menyebabkan makin terkenalnya teknik klastering. Dengan teknik ini, dokumen berita berbahasa Indonesia ini akan dengan mudah dikelompokkan walaupun *class label* belum diketahui. Ada banyak metode klastering yang bisa digunakan, akan tetapi umumnya metode-metode tersebut belum bisa menangani data berdimensi tinggi, deskripsi klaster yang sulit dimengerti serta masih diizinkannya kondisi *overlap* (kondisi dimana satu dokumen bisa masuk ke dalam beberapa klaster).

Permasalahan-permasalahan di atas bisa ditangani dengan menggunakan *Frequent Itemset-Based Hierarchical Clustering* (FIHC). Data berdimensi tinggi dan deskripsi klaster yang sulit dimengerti dapat diatasi dengan mereduksi term-term yang tidak frequent. Sedangkan kondisi *overlap* dapat diatasi melalui *disjoint cluster*.

Hasil klasterisasi dengan algoritma ini divisualisasikan secara hirarki dalam bentuk *tree.* Berdasarkan pengujian, klaster yang dihasilkan oleh algoritma FIHC ini memiliki kualitas yang bagus, terutama bila dibandingkan dengan algoritma lain yakni *Hierarchical Frequent Term-Based Clustering* (HFTC). Deskripsi klaster yang dihasilkan sudah cukup *meaningful* dan kondisi *overlap* juga dipastikan sudah tidak ada. Semakin besar dataset yang digunakan dalam pengujian, maka *minimun support* yang dibutuhkan menjadi semakin kecil. Dan untuk nilai *minimum suppor*t yang sama, semakin kecil nilai *cluster support* akan mengakibatkan nilai *F-Measure* semakin menurun.

Kata kunci: *klastering, frequent term-based text clustering, FIHC, F-Measure*