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 tersebut bisa dengan mudah dikelompokkan walaupun *class label* belum diketahui. Ada banyak metode klasterisasi 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 lain dari proses klasterisasi adalah penentuan kata kunci yang mewakili dokumen. Salah satu cara yang dilakukan dalam proses klasterisasi adalah dengan mencari kata yang menjadi inti dari dokumen. Sebagian besar algoritma klasterisasi menggunakan cara *selected globally significant patterns*, yaitu menemukan kombinasi kata yang mewakili sebagian besar dokumen yang diklasterisasi. Terkadang dalam proses ini muncul suatu permasalahan dimana ada *n* buah dokumen yang tidak mengandung kata-kata tersebut. Hasil klasterisasi bisa menjadi terlalu banyak karena *n* buah dokumen tersebut membuat klaster sendiri atau bahkan bisa menjadi terlalu sedikit karena *pattern* yang dianggap mewakili seluruh dokumen, tidak terdapat dalam *n* buah dokumen secara lokal, sehingga dokumen tersebut bisa masuk ke dalam klaster yang tidak tepat. Akibatnya, hasil klasterisasi menjadi tidak bagus.

Permasalahan-permasalahan di atas bisa ditangani dengan menggunakan *Instance Driven Hierarchical Clustering (IDHC)*. Data berdimensi tinggi dan deskripsi klaster yang sulit dimengerti dapat diatasi dengan mereduksi term-term yang tidak frequent. Sedangkan kondisi *overlap* dapat diatasi melalui *duplicate pruning* dan *refinement cluster*. Serta dengan adanya cara *selected locally significant cluster* , algoritma ini menjadi lebih bagus dalam memilih *term* yang bisa mewakili seluruh dokumen dalam suatu *dataset*. Sehingga, dalam kondisi yang buruk, performansi algoritma IDHC bisa lebih unggul dibanding algoritma FIHC dan HFTC. Dan berdasarkan pengujian, nilai *F-Measure* yang didapatkan jauh lebih stabil.

Kata kunci: *klasterisasi, overlap, IDHC, F-Measure*