Abstrak

Data mining merupakan sebuah proses penemuan pola-pola yang menarik dari sekumpulan data berukuran besar. Dalam data mining, banyak fungsi yang dapat dilakukan, seperti : klasifikasi, klastering dan asosiasi. Pada Tugas Akhir ini akan dibahas mengenai *clustering* data kategori dengan menggunakan metode LIMBO (*scaLable InforMation Bottleneck*).

Clustering adalah proses mengelompokkan objek ke dalam suatu kelompok (*cluster*) sehingga objek memiliki kemiripan sangat besar dengan objek lain yang berada pada *cluster* yang sama, tetapi memiliki ketidakmiripan yang besar dengan objek yang berada pada *cluster* berbeda. *Clustering* telah secara luas diimplementasikan diberbagai bidang seperti *market research, pattern recognition, segmentasi pelanggan* dan sebagainya. *Clustering* data bertipe *categorical* mendapat perhatian khusus karena tipe data ini tidak bisa dihitung jarak kedekatan antar objeknya. Selain itu banyak algoritma *clustering* membutuhkan waktu proses yang lama sehingga tidak cocok digunakan untuk data berukuran besar.

Metode *clustering LIMBO* menggunakan struktur pohon untuk mengklaster dataset. *Clustering LIMBO* menggunakan konsep *Distributional Cluster Feature (DCF)* yang menyimpan informasi dari persebaran nilai atribut dan meringkas informasi mengenai sub*cluster*-sub*cluster* dalam *DCF Tree* kemudian membentuk sejumlah perwakilan *cluster* (centroid) yang selanjutnya digunakan dalam proses pemberian label data. Dari hasil analisa didapatkan bahwa nilai *tetha*(ϕ) sebagai salah satu parameter yang diinputkan user dapat mempengaruhi akurasi sistem. Semakin kecil nilai *tetha*(ϕ), jumlah sub*cluster* yang dihasilkan semakin banyak dan akurasi F-measure cenderung semakin naik. Disamping itu, peningkatan jumlah data ikut mempengaruhi waktu untuk pembangunan DCF tree dan proses *clustering*, semakin banyak jumlah data maka semakin lama waktu yang dibutuhkan, karena semakin banyaknya *subcluster* yang terbentuk.

Kata kunci : data mining, *clustering*, LIMBO