BAB 1

PENDAHULUAN

1.1 Latar Belakang

Clustering merupakan suatu metode untuk mencari dan mengelompokkan data yang memiliki kemiripan karakteristik (*similarity*) antara satu data dengan data yang lain. *Clustering* merupakan salah satu metode data mining yang bersifat tanpa arahan (*unsupervised*). Yang dimaksud metode *unsupervised* yaitu metode ini diterapkan tanpa adanya latihan (*training*) dan guru (*teacher*) serta tidak memerlukan target *output*. Dalam data mining ada dua jenis metode *clustering* yang digunakan dalam pengelompokan data, yaitu *hierarchical clustering* dan *non-hierarchical clustering* (Santosa, 2009).

Teknik *clustering* saat ini juga telah banyak digunakan untuk mengatasi permasalahan yang terkait dengan segementasi data. Implementasi *clustering* ini dapat diterapkan pada berbagai bidang sebagai contoh dalam hal *text mining*. Teknik *clustering* dapat digunakan sebagai metode dalam mengelompokkan dokumen teks yang memiliki kesamaan konten/isi dan tema dari teks tersebut. Tujuan utama dari metode *cluster* adalah pengelompokan sejumlah data atau obyek ke dalam *cluster* (kelompok) sehingga dalam setiap *cluster* akan berisi data yang memiliki kesamaan karakteristik dari data tersebut. Pada pengelompokan dokumen, sekumpulan dokumen yang belum diberi label kelasnya akan dikelompokkan sesuai dengan karakteristik-karakteristik kata yang dimiliki setiap dokumen tersebut. Hal tersebut dilakukan untuk memudahkan pengorganisiran dokumen pada kebutuhan lebih lanjut.

Untuk melakukan proses *clustering* terdapat sebuah algoritma yang sering digunakan karena sifatnya yang relatif cepat dan mudah beradaptasi yaitu algoritma K-Means. Algoritma K-Means merupakan algoritma pengelompokan iteratif yang melakukan partisi set data ke dalam sejumlah K *cluster* yang sudah ditetapkan di awal. Pemilihan K titik data sebagai pusat *cluster* awal juga mempengaruhi hasil *clustering*. Sifat tersebut menjadi karakteristik alami K-Means yang dapat mengakibatkan hasil *cluster* yang didapat pada percobaan berbeda dengan hasil setelah proses *clustering*. Kondisi tersebut dikenal sebagai solusi *local optimum,* yang

artinya algoritma K-Means sangat sensitif terhadap lokasi awal pusat *cluster*. (Prasetyo, 2012).

Perkembangan teknik *clustering* mengakibatkan berbagai penelitian dilakukan untuk menghasilkan *cluster* dengan tingkat akurasi yang semakin baik. Salah satu teknik yang digunakan oleh Kim (2008) yaitu mengkombinasikan algoritma genetika dengan K-Means untuk pengelompokan pelanggan dalam membuat sistem pengambilan keputusan pada jual beli *online*. Dari hasil penelitian tersebut, bisa disimpulkan bahwa algoritma genetika dengan *K-Means* mampu menghasilkan pengelompokan (*clustering*) yang lebih baik dibandingkan dengan metode *Self Organising Map* (SOM) yang berbasis *neural network*.

(Feng dan Wang, 2011) melakukan penelitian untuk mengoptimalkan algoritma K-Means dalam menentukan pusat awal *cluster*, dimana pada hasil penelitian sebelumnya menunjukkan bahwa algoritma K-Means memiliki kelemahan tidak hanya dalam menentukan data awal, tetapi tingkat akurasi *clustering* pun berkurang dan metode menjadi kurang efektif. Sehingga untuk memperoleh *cluster* yang efektif dan akurat maka Feng dan Wang mengoptimalkan Algoritma K-Means dengan Algoritma Genetika menjadi sebuah Algoritma Hibrid. Hasil penelitian menunjukkan bahwa algoritma itu memiliki kualitas *cluster* dan *performance* yang baik.

(Chittu dan Sumathi, 2011) dalam penelitiannya menyebutkan ada sebuah algoritma *Clustering* baru yang diusulkan yaitu *Modified Genetic Algorithm Initializing K-Means* (MGAIK). MGAIK diinspirasi oleh sebuah metode *initialization* algoritma genetik untuk *K-Means clustering* tapi beberapa fitur perbaikan dari GAIK. Akhirnya, ketika perbandingan yang dilakukan mendapatkan hasil bahwa MGAIK lebih baik dari yang sederhana algoritma genetika.

Berdasarkan dari fakta-fakta yang diperoleh dari penelitian sebelumnya maka peneliti akan mengembangkan algoritma genetika untuk menentukan pusat *cluster* awal dari k-means *clustering* pada pengelompokan dokumen teks. Percobaan dilakukan untuk mengetahui bagaimana perbedaan hasil *clustering* menggunakan algoritma genetika dan secara random dalam penentuan pusat *cluster* awal, sehingga hasil pengelompokan dokumen diharapkan dapat mendekati solusi *global optimum*.

1.2 Rumusan Masalah

Adapun rumusan masalah dalam penelitian ini adalah bagaimana menentukan pusat *cluster* awal pada k-means *clustering* dengan menggunakan algoritma genetika pada masalah pengelompokan dokumen.

1.3 Batasan Masalah

Agar penelitian ini lebih fokus maka objek kajian akan dibatasi sebagai berikut :

1. Objek yang diteliti pada penelitian ini adalah dokumen teks berupa konten berita dari beberapa media berita *online*.
2. Dokumen teks yang akan di kelompokan berekstensi *.txt* dengan konten berita yaitu bertema teknologi, kesehatan dan ekonomi.

1.4 Tujuan dan Manfaat

Tujuan dari penelitian ini yaitu untuk menentukan pusat cluster awal k-means dengan memanfaatkan algoritma genetika agar dapat mengoptimasi pusat cluster awal. Sedangkan manfaat dari penelitian ini yaitu pengguna dapat melakukan pengelompokan dokumen secara otomatis dengan hasil pengelompokan yang diharapkan sesuai dengan karakteristik masing-masing dokumen.

1.5 Sistematika Penulisan

Agar penulisan tesis ini sesuai dengan format yang ditentukan maka diberikan sistematika penulisan yang terdiri dari :

Bab 1 : Pendahuluan meliputi latar belakang, rumusan masalah, batasan masalah, tujuan dan manfaat dan sistematika penulisan.

Bab 2 : Tinjauan Pustaka yang berisi mengenai teori-teori yang berkaitan dengan penelitian ini yaitu algoritma k-means *clustering*, algoritma genetika, *similarity* dan *text mining*.

Bab 3 : Metodologi penelitian meliputi kerangka kerja penelitian dan metode yang digunakan.

Bab 4 : Hasil dan Pembahasan meliputi algoritma genetika, pengkodean biner sebagai kromosom, penentuan populasi, evaluasi *fitness*, *crossover*, mutasi, penentuan generasi terbaik, algoritma K-Means *clustering* dan *user interface* aplikasi.

Bab 5 : Kesimpulan dan Saran yang dapat memberikan kesimpulan terhadap penelitian dan saran untuk pengembangan dan penelitian berikutnya.