BAB 6

KESIMPULAN DAN SARAN

6.1 Kesimpulan

Kesimpulan yang dapat diambil dari penelitian ini adalah sebagai berikut:

1. Berdasarkan *dataset* yang telah digunakan, algoritma genetika dapat digunakan dalam pengelompokan dokumen. Namun, diperlukan beberapa adaptasi terhadap komponen-komponen dalam algoritma genetika sebelum dapat digunakan untuk mengelompokkan dokumen. Adaptasi yang perlu dilakukan di antaranya adalah:

   - Merepresentasikan dokumen ke dalam suatu model ruang vektor.
   - Kromosom tersusun atas K *centroid* dalam bentuk vektor.
   - Fungsi *fitness* yang digunakan adalah menggunakan *intracluster similarity*.

2. Perangkat lunak yang menggunakan algoritma genetika untuk mengelompokkan dokumen telah berhasil dibuat. Berdasarkan hasil eksperimen menggunakan *dataset* dalam penelitian ini, rata-rata nilai *purity* dari hasil pengelompokan menggunakan algoritma genetika adalah sebesar 0.799, lebih baik 56% dibandingkan dengan menggunakan algoritma K-means. Hal ini terjadi karena algoritma genetika dapat dengan lebih baik mengatasi *local optimum* dibandingkan dengan algoritma *K-means*. Namun dari segi waktu, algoritma genetika membutuhkan waktu 4365% lebih lama dibandingkan dengan algoritma *K-means*. Hal ini disebabkan oleh proses komputasi yang dilakukan pada algoritma genetika jauh lebih banyak dan kompleks dibandingkan dengan algoritma *K-means*.

3. Representasi kromosom yang kurang tepat juga menjadi alasan algoritma genetika berjalan dengan lambat. Mulai generasi kedua, *centroid* yang menyusun kromosom bersifat tidak *sparse* (memiliki sedikit elemen bernilai nol). *Centroid* yang tidak *sparse* memperlambat proses perhitungan menggunakan *cosine similarity* karena seharusnya perhitungan menggunakan *cosine similarity* dapat mengabaikan elemen berbobot nol. Namun karena banyak elemen yang tidak berbobot nol, maka hanya sedikit elemen yang dapat diabaikan dalam proses perhitungan menggunakan *cosine similarity*.

4. Metrik yang digunakan dalam penelitian ini yaitu *intracluster similarity* kurang merepresentasikan seberapa baik suatu hasil pengelompokan. Berdasarkan hasil eksperimen, nilai *intracluster* similarity tidak berbanding lurus dengan nilai *purity*. Salah satu kemungkinannya adalah karena semakin jauh jarak suatu anggota *cluster* dari *centroid*, maka nilai *intracluster similarity* semakin kecil. Nilai *purity* merupakan suatu nilai biner sehingga sejauh apapun suatu anggota *cluster* dari *centroid*, objek tersebut tetap merupakan anggota dari *cluster*. Kemungkinan yang terjadi adalah banyak objek yang jaraknya cukup jauh dari *centroid* namun tetap merupakan bagian dari *cluster* tersebut karena jarak ke *centroid* lain lebih jauh.

6.2 Saran

Saran dari penulis untuk peneliti selanjutnya agar dapat mengembangkan penelitian ini adalah sebagai berikut:

1. Mencari suatu metrik yang lebih mendekati nilai *purity*. Beberapa alternatif metrik yang bisa dicoba untuk mengembangkan penelitian ini adalah dengan menggunakan metrik *intercluster* atau menggunakan metrik *silhouette*.

2. Menggunakan representasi kromosom yang lain sehingga dapat menjaga agar vektor dari *centroid* tetap *sparse*. Hal ini dapat dilakukan dengan mengubah representasi kromosom menjadi dokumen dan keanggotaannya dalam *cluster*.

3. Memproses dokumen dengan tipe selain TXT seperti file dengan ekstensi DOC, DOCX, PDF, dan lain-lain. Selain itu, pengembangan dari penelitian ini adalah dengan memperhitungkan atribut lain dari suatu dokumen selain teks seperti gambar, metadata (penulis dokumen, waktu dibuat) dan lain-lain.

DAFTAR REFERENSI

[1] Gan, G., Ma, C., dan Wu, J. (2007) *Data clustering: theory, algorithms, and applications*. Siam.

[2] Raposo, C., Antunes, C. H., dan Barreto, J. P. (2014) Automatic clustering using a genetic algorithm with new solution encoding and operators. *International Conference on Computational Science and Its Applications*, pp. 92–103. Springer.

[3] Shah, N. dan Mahajan, S. (2012) Document clustering: a detailed review. *International Journal of Applied Information Systems*, **4**, 30–38.

[4] Maulik, U. dan Bandyopadhyay, S. (2000) Genetic algorithm-based clustering technique. *Pattern recognition*, **33**, 1455–1465.

[5] Holland, J. H. (1992) Genetic algorithms. *Scientific american*, **267**, 66–73.

[6] Sivanandam, S. dan Deepa, S. (2007) *Introduction to Genetic Algorithms*. Springer Science & Business Media.

[7] Zhai, C. dan Massung, S. (2016) *Text data management and analysis: a practical introduction to information retrieval and text mining*. Morgan & Claypool.

[8] Mecca, G., Raunich, S., dan Pappalardo, A. (2007) A new algorithm for clustering search results. *Data & Knowledge Engineering*, **62**, 504–522.

[9] Russell, S. J. dan Norvig, P. (2016) *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.

[10] Srinivas, M. dan Patnaik, L. M. (1994) Genetic algorithms: A survey. *computer*, **27**, 17–26.

[11] Schütze, H., Manning, C. D., dan Raghavan, P. (2008) *Introduction to information retrieval*. Cambridge University Press.

[12] Aizawa, A. (2003) An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, **39**, 45–65.

[13] Ahn, C. W. dan Ramakrishna, R. S. (2003) Elitism-based compact genetic algorithms. *IEEE Transactions on Evolutionary Computation*, **7**, 367–385.