SKRIPSI

PENGELOMPOKAN DOKUMEN BERBASIS ALGORITMA GENETIKA

Cornelius David Herianto

NPM: 2015730034

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS
UNIVERSITAS KATOLIK PARAHYANGAN
2019

UNDERGRADUATE THESIS

GA-BASED DOCUMENT CLUSTERING

Cornelius David Herianto

NPM: 2015730034

DEPARTMENT OF INFORMATICS
FACULTY OF INFORMATION TECHNOLOGY AND SCIENCES
PARAHYANGAN CATHOLIC UNIVERSITY
2019

LEMBAR PENGESAHAN

PENGELOMPOKAN DOKUMEN BERBASIS ALGORITMA GENETIKA

Cornelius David Herianto

NPM: 2015730034

Bandung, 22 Mei 2019

Menyetujui,

Pembimbing

Kristopher David Harjono, M.T.

Ketua Tim Penguji

Anggota Tim Penguji

Husnul Hakim, M.T.

Rosa De Lima, M.Kom.

Mengetahui,

Ketua Program Studi

Mariskha Tri Adithia, P.D.Eng

PERNYATAAN

Dengan ini saya yang bertandatangan di bawah ini menyatakan bahwa skripsi dengan judul:

PENGELOMPOKAN DOKUMEN BERBASIS ALGORITMA GENETIKA

adalah benar-benar karya saya sendiri, dan saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan.

Atas pernyataan ini, saya siap menanggung segala risiko dan sanksi yang dijatuhkan kepada saya, apabila di kemudian hari ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya, atau jika ada tuntutan formal atau non-formal dari pihak lain berkaitan dengan keaslian karya saya ini.

Dinyatakan di Bandung,
Tanggal 22 Mei 2019

Meterai
Rp. 6000

Cornelius David Herianto
NPM: 2015730034

ABSTRAK

Pengelompokan (*clustering*) merupakan sebuah metode untuk menggabungkan himpunan objek ke dalam kelompok-kelompok sedemikan rupa sehingga objek dalam kelompok (*cluster*) lebih mirip (karena suatu hal) satu sama lain daripada objek di kelompok lain [1]. *Document clustering* (pengelompokan dokumen) merupakan proses pengelompokan yang dilakukan terhadap suatu koleksi dokumen. Pengelompokan dokumen diterapkan dalam beberapa bidang seperti penambangan web, mesin pencari (*search engine*), dan temu kembali informasi (*information retrieval*) [3]. Hal yang dilakukan dalam pengelompokan dokumen adalah mengukur kemiripan (*similarity*) antar dokumen dan mengelompokan dokumen yang serupa. Salah satu algoritma pengelompokan yang paling sering digunakan adalah *K-means*. Namun, algoritma *K-means* memiliki kekurangan yaitu dapat terjebak dalam *local optimum*. *Local optimum* adalah suatu solusi yang optimal (baik maksimal maupun minimal) diantara kandidat solusi yang berdekatan dalam masalah optimasi. Dikatakan lokal karena solusi ini hanya optimal apabila dibandingkan dengan kandidat solusi yang berdekatan, tidak optimal secara keseluruhan (*global optimum*).

Algoritma genetika atau biasa disebut *Genetic Algorithm* (GA) adalah suatu algoritma pencarian yang terinspirasi dari proses seleksi alam yang terjadi secara alami dalam proses evolusi. GA merupakan metode penyelesaian masalah yang menggunakan genetika sebagai pemodelannya. Dalam penelitian ini, GA akan digunakan sebagai solusi dari masalah *local optimum*. *Local optimum* dapat diatasi oleh GA yang sudah terbukti efektif dalam masalah pencarian dan optimasi. GA dapat digunakan untuk mengelompokkan dokumen dengan beberapa adaptasi terhadap representasi kromosom, fungsi *fitness*, seleksi, persilangan, dan mutasi.

Algoritma genetika dan algoritma *K-means* diuji menggunakan suatu *dataset* berlabel untuk membandingkan waktu dan hasil pengelompokan dari kedua algoritma tersebut. 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 membuktikan bahwa algoritma genetika sudah dapat mengelompokan dokumen dengan hasil yang memuaskan. 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*.

Kata-kata kunci: Algoritma genetika, Pengelompokan dokumen, Algoritma *K-means*, TF-IDF, *Local optimum*

ABSTRACT

Clustering is a method of creating groups of objects, or clusters, in such a way that objects in one cluster are very similar and objects in different clusters are quite distinct [1]. Document clustering is an organization of documents into clusters. It has been studied intensively because of its wide applicability in various areas such as web mining, search engines, and information retrieval [3]. It is measuring similarity between documents and grouping similar documents together. One of the most frequently used algorithm in clustering is K-means. However, K-means can easily stuck in local optimum. Local optimum of an optimization problem is a solution that is optimal (either maximal or minimal) within a neighboring set of candidate solutions. This is in contrast to a global optimum, which is the optimal solution among all possible solutions, not just those in a particular neighborhood of values.

Genetic Algorithm (GA) is a search algorithm inspired by the natural selection process that occurs naturally in the evolutionary process. GA is a problem solving method that used genetics as its model. A solution candidate is modeled as an individual in GA. Set of these individuals are called population. Every individual in a population is represented by a chromosome. Chromosome is a collection of parameters which formed a solution. That parameters is called gene. Every individual in the population is assigned, by means of a fitness function, a measure of its goodness In this study, GA will be used as a solution to local optimum, which have proven to be effective in search and optimization problems. GA can be used to cluster documents by adapting chromosome representation, fitness function, selection, crossover, and mutation.

Genetic algorithm and K-means algorithm will be tested using a labeled datasets to compare the running time and clustering result. Based on the experimental results using the datasets in this study, the average purity value of clustering results using genetic algorithms is 0.799, 56% greater than using the K-means algorithm. This proves that the genetic algorithm is able to cluster documents with satisfactory results. But in terms of running time, genetic algorithms take 4365% more time than the K-means algorithm. This is caused by the computational process carried out on the genetic algorithm is far more complex than the K-means algorithm.

Keywords: Genetic algorithm, Document clustering, K-means algorithm, TF-IDF, Local optimum

Dipersembahkan kepada Tuhan YME, keluarga tercinta,
dan diri sendiri

KATA PENGANTAR

Puji syukur kepada Tuhan Yang Maha Esa atas berkat yang diberikan kepada penulis sehingga dapat menyelesaikan skripsi dengan judul **Pengelompokan Dokumen Berbasis Algoritma Genetika** dengan baik dan tepat waktu. Selama menjalani proses perkuliahan dan penyusunan skripsi, penulis telah mendapat banyak bantuan dan dukungan dalam menghadapi hambatan yang ada. Oleh karena itu, penulis inign menyampaikan rasa terima kasih kepada:

1. Keluarga penulis yaitu Papa dan Mama yang selalu mendukung penulis secara moral dan materiil sehingga penulis dapat menyelesaikan proses perkuliahan dan skripsi ini dengan baik. Serta kepada Michelle dan Vincent sebagai adik penulis yang selalu mendukung penulis dalam menyelesaikan skripsi ini.

2. Bapak Kristopher David Harjono, M.T. sebagai dosen pembimbing yang telah membimbing penulis hingga dapat menyelesaikan skripsi ini.

3. Bapak Husnul Hakim, M.T. dan Ibu Rosa De Lima, M.Kom. sebagai dosen penguji yang telah membantu dalam menguji dan memperbaiki skripsi ini.

4. Ibu Mariskha Tri Adithia, P.D.Eng selaku Ketua Program Studi Teknik Informatika Fakultas Teknologi Informasi dan Sains Universitas Katolik Parahyangan.

5. Arlin Sasqia Puspa Shiffa sebagai sahabat yang selalu mendampingi penulis dalam penyusunan skripsi dari awal hingga akhir bahkan membantu penulis mempersiapkan presentasi sidang skripsi.

6. Vania Stephanie dan Khezia Josephine sebagai sahabat penulis yang senantiasa memberikan dukungan, semangat, masukkan yang berguna untuk penulis, serta menghibur penulis selama proses penyusunan skripsi ini terutama saat sedang mengalami kesulitan.

7. Teman-teman dari Grup MANAYGBILANGWGAGUNA yaitu AK, Otung, Devie, Gilbert, Hoshea, Jeane, Khen, Mark, Matthew, Oyeng, Bebe, Rifo, Rizky, Sean, Vinny, dan WM yang telah menghibur dan mendukung penulis dalam penyusunan skripsi ini.

8. Teman-teman dari Grup Korea yaitu Ario, Dandy, Fakhry, Felis, Hima, Hizkia, Irvan, Acong, Joshua, Kezia, Edrick, Ocin, Momon, Thoby, Victor, Yona, Yudhis, dan Matthew Ariel sebagai teman seperjuangan di Teknik Informatika UNPAR angkatan 2015 yang telah memberi semangat dan dukungan kepada penulis.

9. Teman-teman penulis lain yang tidak dapat disebutkan satu persatu. Terima kasih untuk segala dukungannya sehingga skripsi ini dapat diselesaikan dengan baik.

Akhir kata, semoga skripsi ini dapat bermanfaat bagi pembaca dan dapat menjadi dasar untuk penelitian yang terkait dengan skripsi ini.

Bandung, Mei 2019

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Pengelompokan (*clustering*) merupakan prosedur untuk mencari struktur alami dari suatu kumpulan data. Proses ini melibatkan pemilihan data atau objek ke dalam kelompok (*cluster*) sehingga objek-objek dalam cluster yang sama akan lebih mirip satu sama lain dibandingkan dengan objek yang berada di *cluster* lain. *Clustering* berguna untuk mereduksi data (mereduksi data dengan volume besar ke dalam kelompok-kelompok dengan karakteristik tertentu), mengembangkan skema klasifikasi (juga dikenal sebagai taksonomi), dan memberikan masukkan atau dukungan terhadap hipotesis mengenai struktur suatu data.

Clustering merupakan salah satu teknik pembelajaran tak terarah (*unsupervised learning*). Pembagian kelompok dalam *clustering* tidak berdasarkan sesuatu yang telah diketahui sebelumnya, melainkan berdasarkan kesamaan tertentu menurut suatu ukuran tertentu [2].

Document clustering (pengelompokan dokumen) merupakan proses pengelompokan yang dilakukan terhadap suatu koleksi dokumen. Pengelompokan dokumen diterapkan dalam beberapa bidang seperti penambangan web, mesin pencari (*search engine*), dan temu kembali informasi (*information retrieval*) [3]. Hal yang dilakukan dalam pengelompokan dokumen adalah mengukur kemiripan (*similarity*) antar dokumen dan mengelompokan dokumen yang serupa. Suatu dokumen dapat terdiri dari beberapa jenis informasi seperti teks, jenis tulisan, ukuran tulisan, warna tulisan, dan gambar.

Salah satu algoritma pengelompokan yang paling sering digunakan adalah *K-means* yang dilakukan dengan cara membagi data ke dalam K kelompok. Kelompok tersebut dibentuk dengan cara meminimalkan jarak antara titik pusat *cluster* (*centroid*) dengan setiap anggota *cluster* tersebut. Titik pusat *cluster* dicari dengan menggunakan rata-rata (*mean*) dari nilai setiap anggota *cluster*. Dalam hal ini, setiap anggota *cluster* dimodelkan sebagai vektor dalam n dimensi (n merupakan banyaknya atribut). *K-means* sudah terbukti efektif dalam melakukan pengelompokan dalam situasi apapun. Namun, cara tersebut tetap saja memiliki kekurangan yaitu dapat terjebak dalam *local optima* tergantung dengan pemilihan *centroid* awal [4].

Masalah *local optima* dapat ditangani menggunakan *Genetic Algorithm* (GA) yang telah terbukti efektif dalam menyelesaikan masalah pencarian dan optimasi. GA merupakan teknik pencarian heuristik tingkat tinggi yang menirukan proses evolusi yang secara alami terjadi [5] berdasarkan prinsip *survival of the fittest*. Algoritma ini dinamakan demikian karena menggunakan konsep-konsep dalam genetika sebagai model pemecahan masalahnya [6].

Dalam GA, parameter dari *search space* dikodekan dalam bentuk deretan objek yang disebut kromosom. Kumpulan kromosom tersebut lalu dikenal sebagai populasi. Pada awalnya, populasi dibangkitkan secara acak. Kemudian, akan dipilih beberapa kromosom menggunakan teknik *roulette wheel selection* berdasarkan fungsi *fitness*. Operasi dasar yang terinspirasi dari Ilmu Biologi seperti persilangan (*crossover*) dan mutasi (*mutation*) digunakan untuk membangkitkan generasi berikutnya. Proses seleksi, persilangan, dan mutasi ini berlangsung dalam jumlah generasi tertentu atau sampai kondisi akhir tercapai.

Fungsi *fitness* tidak hanya berfungsi untuk menentukan seberapa baik solusi yang dihasilkan

namun juga menentukan seberapa dekat solusi tersebut dengan hasil yang optimal [6]. Oleh karena itu, diperlukan fungsi *fitness* yang cocok sehingga GA dapat menghasilkan keluaran yang optimal. Pada masalah *clustering* menggunakan GA, maka fungsi *fitness* yang digunakan harus bisa menggambarkan bahwa seluruh elemen sudah berada dalam *cluster* yang terbaik dan sudah sesuai.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan, rumusan masalah dari penelitian ini adalah sebagai berikut:

1. Bagaimana algoritma genetik dapat digunakan untuk mengelompokkan dokumen?
2. Bagaimana membangun perangkat lunak yang menggunakan algoritma genetik untuk dapat mengelompokkan dokumen?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah disebutkan, tujuan dari penelitian ini adalah sebagai berikut:

1. Mempelajari algoritma genetik dan hubungannya dengan pengelompokan dokumen.
2. Membangun perangkat lunak yang mengimplementasikan algoritma genetik untuk dapat mengelompokkan dokumen.

1.4 Batasan Masalah

Rumusan masalah yang telah disebutkan memiliki ruang lingkup yang cukup luas. Dengan menyadari terbatasnya waktu serta kemampuan, penelitian ini difokuskan dengan memperlihatkan batasan masalah sebagai berikut:

1. Jenis dokumen yang dapat diproses dengan perangkat lunak yang dibuat hanyalah *Text Document* dengan ekstensi *TXT*.
2. Informasi dari dokumen yang diproses dalam pengelompokan hanya berasal dari teks yang menjadi isi dari dokumen tersebut. Gambar dan *metadata* (pemilik, tanggal modifikasi) tidak diperhitungkan.

1.5 Metodologi

Langkah-langkah yang dilakukan dalam penelitian ini adalah:

1. Melakukan studi literatur mengenai model ruang vektor, *Document Clustering* (pengelompokan dokumen), *Genetic Algorithm* (algoritma genetik), dan penggunaan algoritma genetik dalam pengelompokan dokumen.
2. Mencari dokumen yang dijadikan *datasets*.
3. Membuat rancangan perangkat lunak yang menggunakan algoritma genetik sebagai algoritma pengelompokan dokumen.
4. Mengimplementasikan hasil rancangan menjadi perangkat lunak dalam bahasa pemrograman Java.

5. Melatih dan menguji perangkat lunak dengan dokumen yang telah tersedia.
6. Mengevaluasi hasil pengujian lalu melakukan implementasi dan pengujian kembali sampai didapatkan hasil yang sudah sesuai dengan harapan.

1.6 Sistematika Pembahasan

Dokumentasi dari penelitian ini disajikan dalam enam bab dengan sistematika pembahasan sebagai berikut:

1. Bab 1 Pendahuluan
   Bab 1 berisi latar belakang pemilihan "Pengelompokan Dokumen berbasis Algoritma Genetika" sebagai judul dari penelitian ini. Selain itu, dibahas juga rumusan masalah, tujuan penelitian, batasan masalah, serta metodologi penelitian yang menjadi acuan dari penelitian ini.

2. Bab 2 Landasan Teori
   Bab 2 memuat landasan teori yang digunakan dalam penelitian ini. Konsep-konsep yang dibahas yaitu pengelompokan, *local optimum*, K-means, algoritma genetika beserta seluruh operasinya, model ruang vektor, pembobotan term yang terdiri dari bobot frekuensi dan bobot TF-IDF, metrik *Intracluster* untuk mengukur minerja metode *clustering*, dan rvaluasi hasil pengelompokan menggunakan *purity*.

3. Bab 3 Analisis
   Bab 3 memuat hasil analisis berdasarkan landasan teori. Hasil analisis yang ditulis pada bab 3 antara lain analisis *dataset*, representasi dokumen, modifikasi terhadap model ruang vektor beserta metode pembobotannya, representasi kromosom, fungsi *fitness*, dan operasi genetik lainnya.

4. Bab 4 Perancangan
   Bab 4 memuat hasil perancangan berdasarkan hasil analisis pada bab 3. Terdapat tiga bagian dalam bab perancangan yaitu Kebutuhan masukan dan keluaran, perancangan kelas, dan perancangan antarmuka pengguna.

5. Bab 5 Pengujian dan Eksperimen
   Bab 5 memuat hasil pengujian dan eksperimen yang telah dilakukan. Pada bab ini dibahas mengenai skenario pengujian, eksperimen pada algoritma genetika, dan eksperimen pada algoritma *K-means*.

6. Bab 6 Kesimpulan dan Saran
   Bab 6 memuat kesimpulan dari penulis berdasarkan hasil penelitian yang telah dilakukan dan saran untuk peneliti berikutnya agar dapat mengembangkan penelitian ini menjadi lebih baik lagi.