Implementasi Algoritma Genetika Pada Aplikasi Peringkas Dokumen Berita Bahasa Indonesia

Jasman Pardede[1], Dewi Rosmala[1], Putu Joli Artaguna[1]
[1]Jurusan Teknik Informatika Fakultas Teknologi Industri
Institut Teknologi Nasional Bandung
Jasmanpardede78@gmail.com[1], d_rosmala@itenas.ac.id[1], putuguna@hotmail.com[1]

INTISARI

Banyaknya informasi atau data yang tersimpan dalam bentuk dokumen bisa mencapai puluhan bahkan ratusan halaman. Membaca semua halaman satu persatu bisa dikatakan kurang efisien. Membaca banyak halaman dengan waktu yang singkat dan terburu-buru ada kemungkinan data yang didapat kurang relevan. Untuk mengatasi permasalahan tersebut, pada penelitian ini dibuat sebuah aplikasi peringkas dokumen. Pada penelitian ini algoritma yang digunakan adalah Genetic Algorithm, karena Genetic Algorithm mengambil nilai terbaik dari hasil seleksi beberapa kemungkinan secara random. Hasil ringkasan di dapat dari nilai Individu terbaik. Satu individu terdiri dari beberapa kromosom (kalimat). Kalimat kalimat dari individu terbaik akan diambil sebagai kalimat ringkasan. Dari kalimat ringkasan akan dilakukan pemapatan (compression rate) sebesar 50%. Dari hasil penelitian, di dapat bahwa dengan algoritma genetika mampu mencari nilai terbaik dalam individu sebesar ***11326.16****. dari hasil compression rate nilai individu terbaik didapat kalimat kalimat dengan nilai terbaik antara lain :* ***4408.3720, 2131.226, 1612.671****.*

Kata Kunci : Peringkas dokumen, algoritma genetika, berita Bahasa Indonesia

ABSTRACT

The amount of information or data saved as document can be tens or hundreds pages. Reading every page one by one can be called inefficient. Reading many pages in a short time will make the information less relevant. To solve this problem, in this research made a summarize document application. The algorithm used in this study is genetic algorithm because this algorithm works taking the best value from randomize chance. The result of summarization was gotten from the best individual. Individual was created from several chromosomes. The sentences from the best individual will be taken as result. From summarized sentence will be compressed of 50% rate. From the result, the algorithm can search the best value from one individual and the best value is ***11326.16****. From the compression rate of best individual has gotten the best value of sentences, they are:* ***4408.3720, 2131.226, and 1612.671****.*

Keywords: summarize document, genetic algorithm, Indonesian news.

Latar Belakang

Banyaknya informasi atau data yang tersimpan dalam bentuk dokumen bisa mencapai puluhan halaman bahkan ratusan. Untuk memahami satu dokumen dengan banyak halaman membutuhkan waktu yang lama, hal itu bisa dikatakan kurang efisien. Membaca banyak halaman dengan waktu singkat dan terburu - buru dapat mengakibatkan data yang di dapat kurang relevan atau tidak sesuai dengan yang di inginkan. Selain itu juga bagi orang yang membaca dokumen tebal, tidak cukup hanya dengan membaca lama tapi juga perlu membuat ringkasan kecil agar bisa di mengerti isi dari suatu dokumen yang di baca.

Ada beberapa algoritma yang dapat digunakan dalam aplikasi peringkas dokumen yaitu *Latent Semantic Analysis* (LSA), *Hybrid Hidden Markov Model Extraction Method* dan *Genetic Algorithm.*

Genetic algorithm (GA) adalah algoritma pencarian heuristic yang di dasarkan atas mekanisme evolusi biologis. GA pertama kali diperkenalkan oleh John Holand dari Universitas Michigan (1975). Berikut ini dijabarkan struktur secara umum yang dipakai dalam Algoritma Genetika, yaitu :

- *Genetic* (Gen), sebuah nilai yang menyatakan satuan dasar yang membentuk suatu arti tertentu dalam satu kesatuan gen yang dinamakan kromosom. Dalam Algoritma Genetika, gen ini bisa berupa nilai biner, float, integer maupun karakter, atau kombinatorial.
- *Chromosome* (Kromosom), gabungan gen gen yang membentuk nilai tertentu.
- Generasi, menyatakan satu siklus proses evolusi atau satu iterasi di dalam algoritma genetika.
- Individu, menyatakan suatu nilai atau keadaan yang menyatakan salah satu solusi yang mungkin dari permasalahan yang di angkat.

Rumusan Masalah

Berdasarkan latar belakang maka dapat dirumuskan permasalahan sebagai berikut :

1. Bagaimana menganalisa dan mendesain Algoritma Genetika dalam aplikasi.
2. Bagaimana membangkitkan *genetic* (gen), *chromosome* (kromosom) dan individu pada teks dokumen.
3. Bagaimana proses genetika yang terjadi dalam penelitian
4. Bagaimana menentukan nilai dari fitur-fitur teks yang dipakai dalam peringkasan dokumen berita bahasa Indonesia.

Tujuan Penelitian

Tujuan dari penelitian ini untuk Mengimplementasikan Algoritma Genetika terhadap peringkasan dokumen berita bahasa Indonesia.

Batasan Masalah

Batasan masalah yang dipakai dalam pembuatan aplikasi peringkas dokumen berita Bahasa Indonesia adalah sebagai berikut:

1. Menghasilkan ringkasan teks berjenis Ekstraksi bukan berjenis Abstrak.

2. Menggunakan teks berjenis dokumen berita bahasa Indonesia.
3. Format penulisan pada dokumen berita yang akan di ringkas menggunakan format yang benar.

Algoritma Genetika[4]

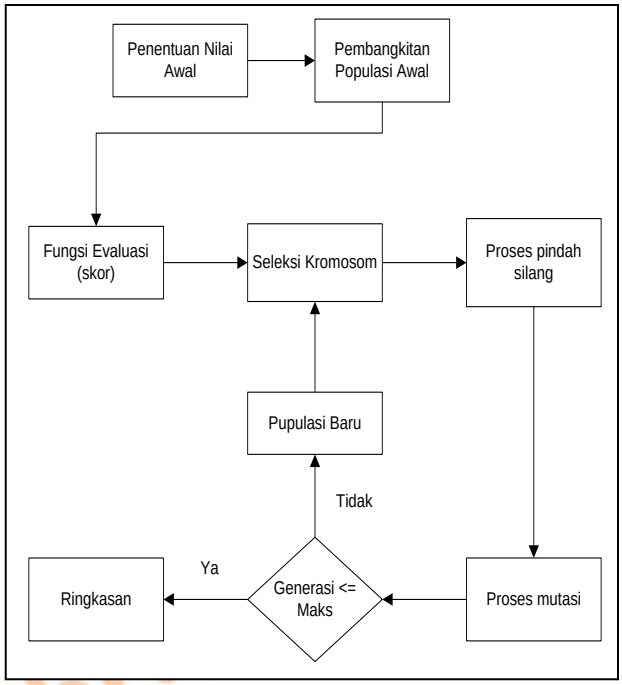
Menurut Goldberg (1998) algoritma genetika atau *genetic algorithm* adalah algoritma pecarian yang didasari pda genetic alamiah dan seleksi alamiah. GA dapat di aplikasikan untuk menyelesaikan permasalahan optimasi kombinasi, yaitu dengan mendapatkan suatu nilai solusi optimal terhadapt suatu permasalahan yang mempunyai banyak kemungkinan (Hermanto 2003).

Dalam penelitian ini, ada beberapa hal yang dilakukan oleh Algoritma Genetika, yaitu :

1. Penentuan gen dari dokumen yang ada.
2. Penentuan kromosom dari dokumen.
3. *Fitness function* yang mengevaluasi kromosom (kalimat).
4. *Fmeasure function* yang mengevaluasi setiap individu (kumpulan beberapa kalimat).
5. Proses genetic (seleksi kromosom, pindah silang, mutas) yang menghasilkan sebuah populasi baru dari populasi yang ada.

Pemodelan Algoritma Genetika[4]

Pada bagian ini, Algoritma Genetika berfungsi sebagai pencarian pembobotan yang optimal pada tiap ekstraksi fitur teks. Tahap ekstraksi fitur teks digunakan untuk menghitung *fitness function* yang berfungsi untuk mengevaluasi kromosom. Proses Algoritma Genetika ditunjukan pada Gambar 10.

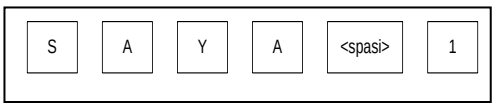


Gambar 1 Proses Algoritma Genetika

(Sumber : Aristoteles, Herdiyeni Y, Ridha A, Julio A. (2012). Pembobotan Fitur pada Peringkasan Teks Bahasa Indonesia Menggunakan Algoritme Genetika. Thesis. Institut Pertanian Bogor)

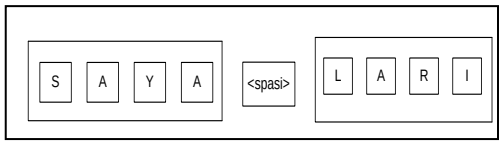
Proses yang terjadi Gambar 1 adalah sebagai berikut:

1. Penentuan nilai awal
   Nilai awal didapat dari dokumen berupa allele, gen, kromosom, individu dan populasi.
   - Allele ditentukan dari setiap karakter yang ada dalam dokumen (angka, huruf, titik, space dll). Gambar 2 memperlihatkan bentuk allele dalam dokumen.



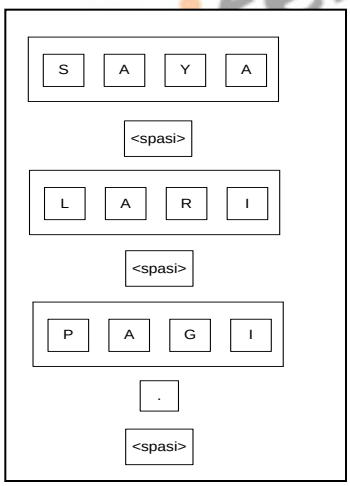
Gambar 2 Menentukan Allele dalam dokumen

- Gen adalah gabungan dari allele. Karakter spasi digunakan sebagai pemisah setiap gen. Gambar 3 memperlihatkan bentuk dari gen.



Gambar 3 menentukan gen dalam dokumen

- Kromosom adalah gabungan dari beberapa gen (kata). Pada penelitian ini, satu kromosom di tandai dengan akhiran titik+spasi (.). Seperti pada Gambar 4 gabungan antar gen sehingga membentuk sebuah kromosom.



Gambar 4 Menentukan kromosom dalam dokumen

- Individu adalah kumpulan dari beberapa kromosom (kalimat), dan
- Populasi adalah kumpulan dari beberapa individu. Dalam penelitian ini yang dimaksud dengan satu populasi adalah hasil ringkasan dari suatu dokumen.

2. Pembangkitan populasi awal

Populasi awal dibangkitkan dari kromosom-kromosom yang telah dikumpulkan dari dokumen, dimana tiap nilai kromosom merepresentasikan nilai ekstraksi fitur teks dan bobot kata masing-masing. Sebuah kromosom direpresentasikan sebagai kombinasi seluruh fitur bobot dalam bentuk $(w_1, w_2, w_3, \cdots, w_n)$.

3. Fungsi Evaluasi

Pemberian nilai pada masing-masing fitur bobot dilihat dari fitur mana yang paling berpengaruh pada saat peringkasan. Nilai yang ada yang ada pada kromosom diterapkan pada (1) yang berfungsi untuk mendapatkan nilai *fitness* tiap kalimat.

$$W_i * (f1 + \ldots + f10) \quad (1)$$

4. Seleksi kromosom

Kromosom yang telah di hitung menggunakan persamaan (2.0) diambil secara acak sebanyak jumlah paragrapf yang ada dalam dokumen. Gambar 5 menjelaskan proses pengambilan kromosom secara acak berdasarkan paragraf.

Gambar 5 Seleksi kromosom secara acak

5. Proses pindah silang

Individu dipilih secara acak untuk dilakukan pindah silang dengan peluang pindah silang 0.65 dan 0.70. Jika nilai individu lebih kecil dari maka terjadi proses pindah silang antara dua individu.

6. Mutasi

Sama halnya dengan pindah silang, mutasi dilakukan secara acak dengan menukar salah satu kromosom dalam invidu. Peluang mutasi terjadi antara 0.0001 sampai 0.2.

7. Generasi Maksimal

Generasi maksimal adalah individu yang memiliki nilai paling tinggi setelah perhitungan. Nilai maksimal dapat ditentukan apabila telah iterasi maksimal dari seluruh kromosom yang ada dalam dokumen. Pada penelitian ini iterasi maksimal didapat dari perkalian antara jumlah paragraph dengan jumlah kalimat.

8. Ringkasan

Saat nilai maksimal didapatkan, maka kromosom-kromsom yang ada didalam nilai terbaik akan menjadi kalimat ringkasan.

Ringkasan Teks[1]

Peringkasan teks adalah proses pemampatan teks sumber ke dalam versi yang lebih pendek namun tetap mempertahankan informasi yang terkandung di dalamnya (Barzilay & Elhadad 1997). Ada dua kriteria peringkasan teks yaitu peringkasan teks berdasarkan ekstrasi dan abstraksi (Jezek & Steinberger 2008). Teknik ekstraksi merupakan suatu teknik untuk menyalin unit unit teks yang paling penting atau informative dari teks sumber menjadi ringkasan, sedangkan teknik abstraksi adalah mengambil intisari dari teks sumber kemudian membuat ringkasan dengan menciptakan kalimat – kalimat baru yang merepresentasikan intisari teks sumber dalam bentuk berbeda (Jezek & Steinberger 2008). Ada beberapa fitur yang digunakan dalam peringkasan, antara lain :

Posisi Kalimat (f1)[1]

Posisi kalimat ditentukan dari titik pada setiap kalimat dalam dokumen.

$$\text{Score}_{f1}(s) = \frac{X}{N} \quad (2)$$

Positive Keyword (f2) [1]

Positive keyword adalah kata yang paling banyak muncul dalam dokumen

$$\text{Score}_{f2}(s) = \frac{Banyaknya Keyword kei Dalam(s)}{Banyaknya Keyword Dalam(S)} \quad (3)$$

Negative Keyword (f3) [1]

Negative Keyword adalah kata yang paling sedikit muncul dalam dokumen.

$$Score_{f_3}(s) = \frac{Keyword\ paling sedikit dalam(s)}{Banyaknya Keyword dalam(S)} \quad (4)$$

Kemiripan antar kalimat (f4) [1]

Kemiripan antar kalimat adalah kata yang yang muncul dikalimat sama dengan kata yang muncul dikalimat lain.

$$Score_{f_4}(s) = \left| \frac{\text{Keyword dalam s} \cap \text{Keyword dalam antar kalimat}}{\text{Keyword dalam s} \cup \text{Keyword dalam antar kalimat}} \right| \quad (5)$$

Kalimat yang meyerupai Judul(f5)[1]

Kata dalam kalimat sama dengan kata yang ada dalam judul dokumen.

$$Score_{f_5}(s) = \left| \frac{\text{Keyword dalam s} \cap \text{keyword dalam judul}}{\text{Keyword dalam s} \cup \text{keyword dalam judul}} \right| \quad (6)$$

Kalimat yang mengandung kata entitas (f6) [1]

Entitas adalah sebuah kumpulan kata yang mengandung makna didalamanya seperti makna pulau, nama orang, institusi, tempat dan lain lain.

$$Score_{f_6}(s) = \frac{\text{Nama entiti dalam (s)}}{\text{Panjang kalimat (s)}} \quad (7)$$

Kalimat yang mengandung numerik (f7) [1]

Dalam sebuah kalimat yang mengandung numerik biasanya dianggap penting

$$Score_{f_7}(s) = \frac{\text{Data numeric dalam (s)}}{\text{Panjang kalimat (s)}} \quad (8)$$

Panjang kalimat (f8) [1]

Panjang kalimat dihitung dari jumlah kata dalam kalimat dibagi dengan jumlah semua kata dalam dokumen.

$$Score_{f_8}(s) = \frac{\text{Jumlah kata dalam (s)}}{\text{Kata unik dalam dokumen}} \quad (9)$$

Koneksi Antar Kalimat (f9) [1]

Banyaknya kalimat yang memiliki kata yang sama denga kalimat lain.

$$Score_{f_9}(s) = \#\text{Jumlah koneksiantar - kalimat} \quad (10)$$

Penjumlaha bobot koneksi antar kalimat (f10) [1]

Penjumlah bobot koneksi kata tiap kalimat dalam dokumen.

$$Score_{f_{10}}(s) = \sum \text{koneksi antar kalimat} \quad (11)$$

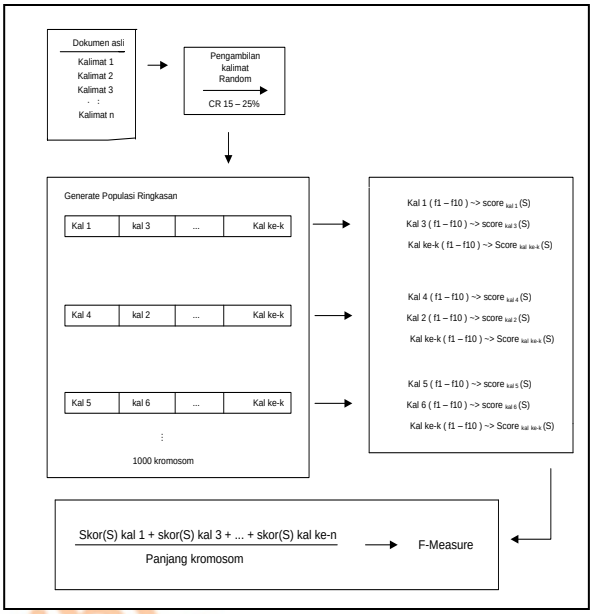
Bobot Kata (w)[2]

Penjumlahan total kemunculan kata tiap kalimat.

$$Log\left(\frac{total\ kalimat}{term\ kemunculan\ kata}\right) \quad (12)$$

Proses Scoring

Perancangan sistem dapat digambarkan seperti yang terlihat pada Gambar 6 yang menjelaskan proses scoring pada apikasi



Gambar 6 proses *scoring* tiakalimat dalam dokumen

Pada Gambar 6 Kalimat dari dokumen asli di ambil secara acak untuk di lakukan *ganerate* populasi ringkasan pada tiap-tiap kromosom. Setelah didapat panjang kromosom (kalimat), dilakukan penghitungan skor pada tiap-tiap kalimat dalam dokumen. Rumus yang digunakan untuk mendapatkan skor(S) yaitu dengan menggunakan rumus (1).

Setelah didapat skor(S) tiap-tiap kalimat dalam kromosom, skor(S) tiap kalimat akan dijumlahkan untuk mendapatkan nilai *F-Measure*. Penjumlahan skor(S) tiap kalimat dapat menggunakan persamaan (13).

$$Skor(S)_{kalimat1} + \ldots + Skor(S)_{kalimatn} \quad (13)$$

Nilai dari (13) akan menjadi nilai *F-Measure*. Nilai *F-Measure* nantinya akan digunakan untuk proses *genetic algorithm*

selanjutnya yaitu mutasi (*mutation)* dan pindah silang (*cross over*).

a. Setelah dihitung nilai tiap kalimat, dilakukan seleksi kromosom. Seleksi kromosom ini berfungsi untuk memilih kromosom-kromosom mana saja yang akan dipilih untuk proses pindah silang, mutasi, dan mendapatkan calon induk yang baik.
b. Peluang pindah silang yang digunakan pada penelitian ini adalah 0.88. Pindah silang terjadi jika peluang yang dihasilkan kromosom yang dijadikan induk lebih kecil dari peluang pindah silang yang telah ditentukan. Teknik yang digunakan pada pindah silang adalah teknik pindah silang satu titik.
c. Peluang mutasi yang digunakan adalah 0.2.
d. Jumlah generasi yang akan ditampilkan disesuaikan dengan jumlah paragraf dan kalimat dalam dokumen.

Skenario Uji

Hasil ringkasan didapat dari berbagai perhitungan yang dijelaskan sebelumnya. Ada beberapa hal yang dilakukan dalam pengujian sebelum mendapatkan hasil ringkasan dengan nilai maksimal, antara lain:

1. Menghitung nilai setiap fitur dan bobot kata.

   Nilai setiap fitur dihitung dengan rumus yang telah dijelaskan sebelumnya (fitur f1 – f10) dan bobot kata.

2. Menghitung nilai setiap kalimat dalam dokumen

   Nilai dari setiap kalimat didapat dari hasil perkalian antara nilai bobot kata dengan total penjumlahan nilai fitur (f1 – f10). hasil dari perkalian tersebut dinamakan nilai fitness : Fitness = w_i * (f1 + f2 + f3 + f4 + f5 + f6 + f7 + f8 + f9 + f10).

3. Menghitung nilai setiap individu

   Nilai individu didapat dari penjumlahan nilai fitness yang ada dalam satu interasi. Hasil dari penjumlahan nilai fitness dinamakan nilai fmeasure: Fmeasure = $Skor_{kalimat\_ke\_1}$ + $Skor_{kalimat\_ke\_2}$ + … + $Skor_{Kalimat\_ke\_i}$

4. Mencari Fmeasure terbaik

   Nilai Fmeasure terbaik didapat dari perbandingan setiap nilai Fmeasure yang telah dihitung disetiap individu.

5. *Compression rate* 50% dari Fmeasure terbaik.

   Jumlah kalimat yang terdapat dalam nilai Fmeasure terbaik akan dikalikan sebesar 50%. Hasil dari perkalian tersebut akan diambil sebagai kalimat ringkasan. Fungsi dari compression rate adalah untuk mengetahui seberapa berpengaruh nilai compression rate terhadap ringkasan yang didapatkan.

Hasil Pengujian

Dari hasil pengujian aplikasi sesuai dengan skenario pengujian, didapat nilai terbaik dalam individu sebesar 11326.16. Dari hasil *compression rate* nilai individu terbaik didapat kalimat-kalimat dengan nilai terbaik antara lain : 4408.3720, 2131.226, 1612.671

Kesimpulan

Beberapa kesimpulan yang diapat dari hasil penelitian adalah :

1. Algoritma genetika dapat di implementasikan dalam meringkas dokumen berita Bahasa Indonesia dengan menggunakan beberapa parameter peringkasan, yaitu : posisi kalimat, kata yang paling banyak

muncul, kata yang paling sedikit muncul, kemiripan antar kalimat, kalimat yang menyerupai judul dokumen. kalimat yang mengandung nama entity, kalimat yang mengandung numerik, panjang kalimat, koneksi kalimat, dan penjumlahan bobot koneksi antar kalimat‖ sesuai dengan pengujian.

2. Peringkasan dokumen bertipe ekstraksi sangat ditentukan oleh besarnya nilai *compression rate.* Semakin besar nilai *compression rate* (25%, 30%, 50%) nya, semakin terwakili hasil ringkasan dokumen dari dokumen aslinya.

Daftar Pustaka

1. Aristoteles, Herdiyeni Y, Ridha A, Julio A. (2012). Pembobotan Fitur pada Peringkasan Teks Bahasa Indonesia Menggunakan Algoritme Genetika. Thesis. Institut Pertanian Bogor.
2. Tala, F. Z. (2003). A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia. M.S. thesis. M.Sc. Thesis. Master of Logic Project. Institute for Logic, Language and Computation. Universiteti van Amsterdam The Netherlands.
3. Yeh J, Ke H, Yang W, Meng I. 2005. *Text summarization using a trainable summarizer and latent semantic analysis*. Information Processing & Management. 41(1). 75-95.
4. Golberg, David E. 1989. Genetic Algorithm in Search, Optimizaton & Machine Learning. The university of Alabama.
5. Hirao, T., Okumura, M., Yasuda, N., & Isozaki, H. (2007). Supervised automatic evaluation for summarization with voted regression model. Information Processing & Management, 43(6), 1521-1535.
6. The Modelling Language user Guide, Grady Booch, James Rumbough, and Ivar Jacobson. Addison Wesley. 1999.