Penerapan Algoritma Genetika pada Peringkasan Teks Dokumen Bahasa Indonesia

Aristoteles

Jurusan Ilmu Komputer FMIPA Universitas Lampung
aristoteles@unila.ac.id

Abstrak. Tujuan penelitian ini adalah meringkas dokumen bahasa Indonesia yang berjenis file teks dengan menggunakan algoritma genetika. Terdapat sebelas fitur teks yang diterapkan pada penelitian ini, yaitu posisi kalimat, *positive keyword*, *negative keyword*, kemiripan antar kalimat, kalimat menyerupai judul, kalimat yang mengandung nama entiti, kalimat yang mengandung data numerik, koneksi antar-kalimat, penjumlahan bobot antar-kalimat, dan kalimat semantik. Penelitian ini terbagi atas tiga tahap yaitu : tahap pengumpulan dokumen, tahap pelatihan, dan tahap pengujian. Hasil pengujian menunjukkan bahwa akurasi dengan pemampatan 30%, 20%, 10% sebesar 47.46%, 41.29% dan 35.01%.

Keywords : peringkasan teks, algoritma genetika.

PENDAHULUAN

Pada saat ini, perkembangan teknologi informasi sangat cepat, salah satunya adalah penggunaan internet. Tujuannya adalah untuk mendapatkan informasi dengan cepat dan akurat. Seiring bertambahnya informasi, maka berbanding lurus dengan dokumen yang ada di dunia internet, salah satu contohnya adalah dokumen berita, Dokumen berita merupakan kumpulan informasi tentang banyak peristiwa penting terjadi dan terbaru secara berkala. Memahami isi dokumen berita melalui ringkasan teks memerlukan waktu yang lebih singkat dibandingkan membaca seluruh isi dokumen, sehingga ringkasan teks menjadi sangat penting. Namun demikian, membuat ringkasan manual dengan dokumen yang banyak akan memerlukan waktu dan biaya yang besar. Sehingga diperlukan suatu sistem peringkasan secara otomatis untuk mengatasi masalah waktu baca dan biaya [1].Peringkasan teks adalah suatu proses yang menghasilkan dokumen yang lebih kecil 50% dari ukuran dokumen [2] dengan tujuan memperoleh informasi dalam waktu singkat [3]. Menurut [4] peringkasan teks adalah proses pencarian informasi yang penting dari sumber (atau beberapa sumber) untuk menghasilkan dokumen yang ringkas bagi pengguna.

Pada penelitian [1], melakukan penentuan tingkat kepentingan atau pembobotan dari sebelas fitur teks untuk meringkas dokumen. Penelitian ini merupakan kelanjutan dari penelitian [1] yaitu meringkas dokumen teks. Hasil ringkasan diuji dengan menggunakan *F-measure, Precision, Recall* [5].

METODE PENELITIAN

Penelitian ini dilakukan dengan tiga tahap yaitu : tahap pengumpulan dokumen, tahap pelatihan, dan tahap pengujian. Ketiga tahap tersebut dapat dilihat pada Gambar 1.

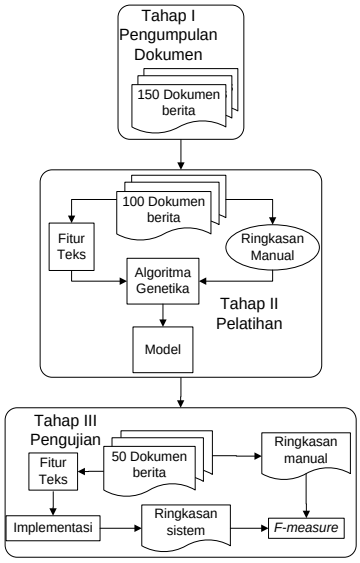
Tahap I Pengumpulan Dokumen

Penelitian ini menggunakan 150 dokumen berita yang beasal dari penelitian [6]. Pada tahap pelatihan digunakan 100 dokumen sedangkan 50 dokumen digunakan untuk pengujian sistem.

Tahap II Pelatihan

Tujuan dari tahap pelatihan ini adalah untuk menentukan bobot atau tingkat

kepentingan dari tiap-tiap fitur teks. Penentuan bobot dilakukan dengan menggunakan algoritma genetika. Bobot yang optimal dapat dijadikan model untuk peringkasan teks. Menurut [1], terdapat sebelas fitur teks tiap kalimat dalam dokumen. Berikut ini sebelas fitur teks yaitu :



Gambar 1. Metode peringkasan teks

Posisi Kalimat (f1)

Posisi kalimat adalah letak kalimat dalam sebuah paragraf. Pada penelitian ini diasumsikan bahwa kalimat pertama pada tiap paragraf adalah kalimat yang paling penting. Oleh karena itu, penelitian ini mengurutkan kalimat tersebut berdasarkan posisinya.

$$\text{Score}_{f_1}(s) = \frac{X}{N} \qquad (2.1)$$

Positive keyword (f2)

Positive keyword adalah kata yang paling banyak muncul pada sebuah kalimat.

$$\text{Score}_{f_2}(s) = \frac{1}{\text{length}(s)} \sum_{i=1}^{n} \text{tf}_i * P(s \in S|\text{keyword}_i) \qquad (2.2)$$

Asumsikan s adalah kalimat dalam ringkasan dokumen, S adalah kalimat dalam dokumen, f_2 adalah fitur teks *positive keyword* (fitur teks kedua), n adalah jumlah *keyword* dalam kalimat, tf_iadalah banyaknya *keyword* ke-i yang muncul dalam kalimat.

***Negative keyword* (f3)**

Negative keyword merupakan kebalikan dari fitur teks positive keyword. Negative keyword adalah kata yang sedikit muncul dalam kalimat.

$$\text{Score}_{f_3}(s) = \frac{1}{\text{length}(s)} \sum_{i=1}^{n} \text{tf}_i * P(s \notin S|\text{keyword}_i) \qquad (2.3)$$

Kemiripan Antar-Kalimat (f4)

Kemiripan antar-kalimat merupakan kata yang muncul dalam kalimat sama dengan kata yang muncul dalam kalimat lain.

$$\text{Score}_{f_4}(s) = \left|\frac{\text{Keyworddalams} \cap \text{Keyworddalamantarkalimat}}{\text{Keyworddalams} \cup \text{Keyworddalamantarkalimat}}\right| \qquad (2.4)$$

Kalimat yang Menyerupai Judul Dokumen (f5)

Kalimat yang menyerupai judul dokumen adalah kata yang muncul dalam kalimat sama dengan kata yang ada dalam judul dokumen.

$$\text{Score}_{f_5}(s) = \left|\frac{\text{Keyword dalam s} \cap \text{Keyword dalam judul}}{\text{Keyword dalam s} \cup \text{Keyword dalam judul}}\right| \qquad (2.5)$$

Kalimat yang Mengandung Nama Entiti (f6)

Nama entiti adalah sebuah kumpulan kata yang memiliki makna atau membentuk nama sebuah institusi, nama orang, nama pulau. Misalnya Institut Pertanian Bogor merupakan kumpulan kata yang memiliki makna sebuah institusi perguruan tinggi.

$$Score_{f_6}(s) = \frac{nama\ entiti\ dalam\ (s)}{Panjang\ kalimat\ (s)} \qquad (2.6)$$

Kalimat yang Mengandung Nama Numerik (f7)

Pada peringkasan teks mempertimbangkan data numerik, karena dalam kalimat yang berisi data numerik terdapat kalimat yang penting.

Panjang Kalimat (f8)

Panjang kalimat dihitung berdasarkan jumlah kata dalam kalimat dibagi jumlah kata unik dalam dokumen.

$$Score_{f_8}(s) = \frac{jumlah\ kata\ dalam\ (s)}{kata\ unik\ dalam\ dokumen} \quad (2.8)$$

Koneksi Antar- Kalimat (f9)

Koneksi antar-kalimat adalah banyaknya kalimat yang memiliki kata yang sama dengan kalimat lain dalam satu dokumen.

$$Score_{f_9}(s) = \#Jumlah\ koneksi\ antar - kalimat \quad (2.9)$$

Penjumlahan Bobot Koneksi Antar-Kalimat (f10)

Fungsi fitur teks ini adalah menjumlahkan bobot koneksi antar-kalimat. Perhitungan fitur teks penjumlahan bobot koneksi antar-kalimat dilihat pada (2.10) dengan asumsi s adalah kalimat, f_{10} adalah fitur teks penjumlahan bobot koneksi antar-kalimat.

$$Score_{f_{10}}(s) = \sum koneksi\ antar\ kalimat \quad (2.10)$$

Kalimat Semantik (f11)

Kalimat semantik adalah kalimat yang mencirikan hubungan antar kalimat yang didasari semantik. Asumsikan D adalah sebuah dokumen, $t(|t| = M)$ adalah banyaknya kata dalam D, dan $S(|S| = N)$ adalah banyaknya kalimat dalam D.

Matriks kata dapat dilihat pada (2.11), dengan S_j adalah kalimat ke-j dalam dokumen dan t_i adalah *term* ke-i yang muncul didalam dokumen. Pada penelitian ini menggunakan semua *keyword* atau *term* yang ada dalam dokumen kecuali kata-kata *stoplist*.

$$A = \begin{array}{c|cccc} & S_1 & S_2 & \cdots & S_n \\ t_1 & w_{1,1} & w_{1,2} & \dots & w_{1,n} \\ t_2 & w_{2,1} & w_{2,2} & \dots & w_{2,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ t_m & w_{m,1} & w_{m,2} & \dots & w_{m,n} \end{array} \quad (2.11)$$

dengan $w_{i,j}$didefinisikan pada (2.12), dan tf_i adalah banyaknya kemunculan *term* ke-i pada kalimat. SF_i s*entences frequency*$_i$ merupakan banyak kalimat yang mengandung *term* ke-i, sedangkan $ISF_i = \log\left(\frac{N}{SF_i}\right)$ merupakan ukuran diskriminan kemunculan *term* ke-i dalam dokumen, N adalah banyaknya kalimat dalam satu dokumen.

Algoritme Genetika

Menurut [7] algoritme genetika atau genetic algorithm adalah algoritme pencarian yang didasari pada mekanisme genetik alamiah dan seleksi alamiah. Siklus dari algoritme genetika diperkenalkan [7], dapat dilihat pada Gambar 4. Siklus ini terdiri beberapa bagian yaitu: populasi awal, evaluasi fitness, seleksi individu, pindah silang (crossover), mutasi (mutation), dan populasi baru.

Populasi awal adalah sekumpulan kromosom awal yang dibangkitkan secara acak dalam satu generasi. Populasi baru merupakan sekumpulan kromosom baru hasil dari proses seleksi, pindah silang dan mutasi.

Seleksi adalah tahapan dalam algoritme genetika yang berfungsi memilih kromosom yang terbaik untuk proses pindah silang dan mutasi [8] dan mendapatkan calon induk yang baik. Semakin tinggi nilai *fitness* suatu individu semakin besar kemungkinannya untuk dipilih. Jika kromosom memiliki nilai *fitness* kecil, maka tergantikan oleh kromosom baru yang lebih baik.

Pindah silang merupakan komponen yang penting dalam GA [9]. Pindah silang adalah operator dari algoritme genetika yang melibatkan dua induk untuk membentuk kromosom baru. Pindah silang menghasilkan titik baru dalam ruang pencarian yang siap diuji.

Mutasi diperlukan untuk mencari solusi optimum, yaitu 1) mengembalikan gen-gen yang hilang pada generasi berikutnya, 2) memunculkan gen-gen baru yang belum pernah muncul pada generasi

sebelumnya [9].

HASIL DAN PEMBAHASAN

Data Korpus

Penelitian ini menggunakan 150 dokumen berita yang berasal dari harian kompas online [7]. 100 dokumen digunakan untuk data training, sedangkan 50 dokumen digunakan pada tahap pengujian system. Pemampatan 30%, 20%, dan 10% isi dokumen hanya dilakukan pada penelitian ini.

Format Data

Format data yang digunakan pada penelitian ini adalah format XML. Dimana terdapat tag-tag yang digunakan sebagai penanda pembacaan isi dokumen. Contoh format data terlihat pada Gambar 2.

<TITLE>Pemerintah Pede Bisa Bayar Utang Rp 2.000 Triliun</TITLE>
<TEXT>
JAKARTA, KOMPAS.com - Pemerintah mengaku optimis bisa membayar utang negara sekitar Rp 2.000 triliun. Jumlah tersebut sekitar 23 persen dari total produk domestik bruto (PDB) Indonesia sebesar Rp 9.000 triliun. Wakil Menteri Keuangan Mahendra Siregar mengatakan pemerintah saat ini fokus untuk menjaga fiskal negara. Sebab, selama ini keuangan negara mengalami defisit karena sebagian besar keuangan negara digunakan untuk anggaran subsidi bahan bakar minyak (BBM).
</TEXT>

Gambar 2. Format dokumen

Aplikasi Sistem

Aplikasi system terdiri atas pembacaan dokumen dengan format xml, pemotongan kalimat, pemotongan kata, dan pemisahan kata-kata yang tidak penting. Berikut ini pemotongan kalimat yang digunakan pada penelitian ini :

1. Batas kalimat setelah tanda baca . ? !
2. Batas kalimat sesudah tanda pentik, bukan setelah tanda titik
3. Dapat mengenali singkatan, Aris, M.Si

Hasil Pengujian

Tabel 2. Akurasi dari hasil pengujian

	30%	20%	10%
Akurasi	47.46%,	41.29%	35.01%.

Berdasarkan hasil pengujian yang telah dilakukan pada penelitian ini dapat dilihat pada Tabel 1 yang menunjukan pemampatan sangat berpengaruh pada hasil akurasi pengujian.

KESIMPULAN

Hasil penelitian ini dapat disimpulkan bahwa algoritme genetika dapat digunakan untuk mencari tingkat kepentingan yang optimal dari tiap fitur teks. Nilai akurasi 47.46% pada pemampatan 30%. Sedangkan hasil tidak optimal pada pemampatan 10%.

Tidak perlu dibuat sub bab di bagian Kesimpulan.Kesimpulan merupakan simpulan dari analisis yang telah dilakukan serta menjawab tujuan dari penelitian sebagaimana tersirat dalam bagian Pendahuluan. Saran hendaknya singkat saja terkait dengan peluang perbaikan yang mungkin dapat dilakukan untuk kesempurnaan penelitian terkait berikutnya.

DAFTAR PUSTAKA

Aristoteles, Herdiyeni Y, Ridha A, Julio A. (2012). Text Feature Weighting for Summarization of Documents in Bahasa Indonesia Using Genetic Algoritm. International Journal of Science Issues. ISSN 1694-0814.

Radev D, Hovy E, McKeown K. (2002). Introduction to the special issue on text summarization. Computer linguist.

Blake C, Pratt W, Rules B, Fiturs F. (2001). A semantic approach to selecting fiturs from text. ICDM. 59–66.

Manning CD, Raghavan P, Schutze H. (2008). Introduction to Information Retrieval. Cambridge: Cambridge University Press.

Baeza-Yates R, Ribeiro-Neto B. (1999). Modern Information Retrieval. ACM Press New York. Addison-Weslye.

Ridha A. (2002). Pengindeksan otomatis dengan istilah tunggal untuk dokumen berbahasa indonesia [skripsi]. Bogor. Ilmu Komputer, Matematika dan Ilmu

Pengetahuan Alam, Institut Pertanian Bogor.

Goldberg DE. 1989. Genetic algorithms in search,optimization, and machine learning. Addison Wesley Longman, Inc.

Cox E. 2005. Fuzzy modeling and genetic algorithm for data mining and exploration. USA: Morgan Publisher.

Gen M, Cheng R. 1997. Genetic algorithm and engineering design. John Wiley & Sons, Inc. Canada.