PERINGKASAN TEKS OTOMATIS MENGGUNAKAN HARMONY SEARCH ALGORITHM-BASED SENTENCE EXTRACTION

Ardan Budi Pramana[1], Z.k. Abdurahman Baizal[2], Kemas Rahmat Saleh Wiharja[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Saat ini jumlah terbesar dari informasi tekstual ada di repositori seperti WEB. Untuk proses seperti informasi yang berjumlah besar Peringkasan teks otomatis (automatic text summarization) menjadi hal yang begitu penting. Peringkasan teks otomatis (automatic text summarization) adalah proses menyaring informasi paling penting dari sebuah sumber atau beberapa sumber untuk membuat sebuah versi ringkas dari teks dengan memanfaatkan aplikasi yang dijalankan pada komputer.

Tugas akhir ini menggunakan sebuah metode untuk membuat ekstraksi berdasarkan tiga faktor readability , cohesion, dan topic relation dengan menggunakan harmony search berdasarkan seleksi kalimat untuk membuat ringkasan. Setelah dibuat ringkasan , dievaluasi dengan fungsi fitness berdasarkan ketiga faktor. Selain itu dicari parameter dari harmony search yang dapat menghasilkan ringkasan yang optimal. Pengujian juga dilakukan menggunakan ROUGE evaluation toolkit untuk melihat hasil recall, precision dan f-measure dengan membandingkan hasil ringkasan manusia atau ringkasan referensi. Parameter harmony search yang dapat menemukan ringkasan optimal global mempunyai nilai HMCR sebesar 0.9. Paraemter HMCR diatas 0.9 yang mendekati 1 tidak lebih baik dikarenakan proses diversifikasi menjadi berkurang. Hasil dari pengujian menggunakan ROUGE-2 membuktikan bahwa rata-rata hasil ringkasan menggunakan algoritma harmony search 50% mendekati ringkasan referensi.

Kata Kunci : automatic text summarization, readability , cohesion, and topic relation., harmony search

Abstract

Currently the largest amount of textual information in a repository such as WEB. To process such large amounts of information text Summarization automatic (automatic text summarization) becomes so important. Automatic Text Summarization (automatic text summarization) is the most important filter information from a source or multiple sources to create a compact version of the text by using applications that run on the computer.

This final project uses a method to create three-factor extraction based on readability, cohesion, and topic relations by using harmony search based on the selection of the sentence to make a summary. Once created summary, evaluated by a fitness function based on three factors. Also look for the harmony search parameters that can generate an optimal summary. Testing is also done using the Rouge evaluation toolkit to see the results of recall, precision and f-measure by comparing the results of human summary or summary reference. Harmony search parameters that can find a summary of the global optimum has a value of 0.9 HMCR. Paraemter HMCR above 0.9 is close to 1 is not better because the process of diversification is reduced. Results from tests using Rouge-2 proves that the average summary results using harmony search algorithm, 50% close to the reference summary.

Keywords : automatic text summarization, readability , cohesion, and topic relation., harmony search

1. Pendahuluan

1.1 Latar belakang

Berkembangnya dunia teknologi dan informasi yang semakin pesat menyebabkan kebutuhan mendapatkan informasi juga harus cepat. Informasi merupakan hal yang penting dan sangat dibutuhkan oleh orang banyak. Salah satu bentuk informasi adalah berupa artikel dan berita *online*. Proses untuk mendapat informasi dari sebuah dokumen membutuhkan waktu yang lama karena manusia harus melakukan ekstraksi/abstraksi manual terhadap dokumen. Hal ini mengurangi efektivitas dan efisiensi perolehan informasi di tengah data yang melimpah.Untuk menanggapi pertumbuhan yang pesat dari informasi maka peringkasan teks dapat menolong manusia untuk mengekstrak inti dari informasi tersebut.

Peringkasan teks otomatis (*automatic text summarization* atau *ATS*) adalah pembuatan ringkasan dari sebuah teks secara otomatis dengan memanfaatkan aplikasi yang dijalankan pada komputer. Terdapat dua pendekatan pada peringkasan teks, yaitu ekstraksi dan abstraksi. Pendekatan ekstraksi, pertama kali diperkenalkan oleh Luhn (1958)[7]. Luhn menggunakan teknik statistik sederhana untuk menentukan kalimat yang paling signifikan dalam suatu dokumen. Kalimat ini kemudian diekstrak dari dokumen dan dijadikan ringkasan.

Tugas akhir ini metode peringkasan berbasis graf menggunakan pendekatan ekstraksi. *Graph-based summarization algorithm* atau peringkasan teks berbasis graf merupakan suatu metode peringkasan teks yang *language independent* dan dapat menghasilkan ringkasan ekstraktif. Teks sumber direpresentasikan menjadi sebuah graf sehingga disebut graf tekstual [13]. *Node* pada graf tersebut dapat berupa unit-unit teks seperti kata-kata, kalimat-kalimat, atau paragraf-paragraf dalam teks. *Edge* dalam graf menunjukkan keterhubungan antar *node*. Keterhubungan dapat berupa *similarity* antar kalimat ataupun hubungan leksikal atau gramatikal antar kata/frasa. Konsep *similarity* antar unit teks digunakan dalam proses pembangunan graf tekstual.

Beberapa metode untuk ekstraksi, kalimat yang diekstrak berdasarkan relevansi dengan topik dokumen. Akan tetapi metode-metode tersebut mungkin menghasilkan ringkasan dengan keterbacaan rendah[2]. Ringkasan yang baik adalah setiap kalimat dalam ringkasan terkait dengan topik, mudah dibaca dan koheren (yaitu dapat dilihat sebagai dokumen kecil). Dalam sebuah ringkasan mungkin saja kalimat tersebut sangat tinggi keterhubungannya dengan topik dan keterbacaannya sangat rendah sehingga sulit dibaca. Akan tetapi ada juga ringkasan yang mempunyai keterhubungan dengan topik sangat rendah dan keterbacaanya sangat tinggi. Karena adanya proses *trade offs* antara ketiga faktor tersebut maka pembuatan ringkasan dapat dikatakan sebuah masalah optimasi[2].

Pada kasus peringkasan teks ini merupakan masalah optimasi maka dapat digunakan algoritma optimasi. Salah satu algoritma optimasi adalah algoritma *Harmony Search* (HS). *Harmony Search* adalah algoritma metaheuristik yang berbasis populasi. HS menirukan evolusi yang terjadi pada proses pertunjukan musik, misalnya improvisasi jazz yang berusaha mencari harmoni lebih baik. Dengan analogi tersebut, HS melakukan proses optimasi untuk mendapatkan

keadaan terbaik(optimum global) dengan cara mengevaluasi fungsi objektif. Dengan menggunakan himpunan pola-pola titik nada (*pitches)* yang dikeluarkan oleh alat-alat music, fungsi objektif pada HS dihitung menggunakan himpunan nilai-nilai yang diberikan untuk variable-variabel keputusan (*decision variables*). Jika kualitas suara estetika dapat diperbaiki melalui latihan-latihan demi latihan maka nilai fungsi objektif juga dapat terus ditingkatkan dari iterasi ke iterasi[11]. Algoritma *Harmony Search* ini dapat digunakan untuk menemukan solusi ringkasan yang optimum global.

Dengan menggunakan algoritma HS kalimat-kalimat yang ada pada dokumen akan diekstrak sesuai panjang ringkasan yang diinginkan. Dengan menggunakan himpunan kalimat-kalimat ringkasan yang dikombinasikan, fungsi objektif dihitung untuk menentukan kualitas ringkasan. Kualitas dari ringkasan dapat ditingkatkan dengan mengektrak kalimat-kalimat ringkasan yang baru sehingga nilai fungsi objektif dapat ditingkat dari iteri ke iterasi.

1.2 Perumusan Masalah

Berdasarkan latar belakang tersebut, maka permasalahan yang diangkat pada tugas akhir ini :

1. Bagaimana Algoritma *Harmony Search* dapat menghasilkan peringkasan dokumen dengan menggunakan pendekatan ekstraksi.
2. Bagaimana performansi karakteristik parameter *Harmony Search* yaitu *Harmony Memory Considering Rate* dalam menemukan solusi ringkasan yang optimal.

1.3 Tujuan

Tujuan tugas akhir ini adalah:

1. Mengimplementasikan algoritma *Harmony Search* untuk pembuat peringkasan teks otomatis pada dokumen tunggal berbahasa Indonesia, khususnya artikel berita.
2. Menganalisa performansi karakteristik parameter *Harmony Search* yaitu *Harmony Memory Considering Rate* dalam hal optimasi untuk mendapatkan ringkasan.

1.4 Hipotesa

Pada peringkasan teks untuk parameter *Harmony Memory Considering Rate* (HMCR) semakin mendekati 1 hasil *F-Measure* yang diperoleh tidak terlalu bagus.

1.5 Batasan Masalah

Terdapat beberapa batasan dalam penelitian tugas akhir ini, antara lain:

1. Teks sumber yang akan diringkas adalah artikel berita berbahasa Indonesia yang diperoleh dari *website* berita.
2. Peringkasan dilakukan secara *offline*.
3. Peringkasan yang dilakukan adalah peringkasan dokumen tunggal.
4. Hasil peringkasan berupa ekstraksi dari teks sumber.

5. Tidak melakukan *stemming* terhadap teks masukan.
6. Tanpa mengeliminasi *stopwords*.
7. Evaluasi dilakukan dengan membandingkan *content overlap* antara peringkasan otomatis dan ringkasan referensi dengan menggunakan ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) *evaluation toolkit* [1].

1.6 Metodologi penyelesaian masalah

Metodologi yang dilakukan untuk menyelesaikan permasalahan adalah sebagai berikut:

1. Identifikasi masalah, yakni melakukan identifikasi masalah yang akan di analisa pada tugas akhir ini
2. Studi literatur, tahapan eksplorasi dan studi literatur terhadap *automatic text summarization* dengan algoritma *Harmony Search.* Tahap ini melakukan studi literatur seperti jurnal, website , maupun artikel dan bacaan yang relevan.
3. Analisa dan pengumpulkan data yang digunakan sebagai inputan :
   Data yang digunakan sebagai inputan berupa single dokumen teks. Dokumen tersebut akan direpresentasikan sebagai *Directed Acyclic Graph* (DAG). Setiap kalimat akan direpresentasikan sebagai *node* dan *edge* sebagai penghubung.
4. Analisa dan perancangan pembangunan Algoritma *Harmony Search* berdasarkan langkah-langkah berikut :
   a. Inisialisasi permasalahan dan parameter *Harmony Search* (HS)
      Pada tahap ini dilakukan pendefinisian permasalahan yaitu meminimalisasi nilai *f*(x) dan pendefinisian parameter-parameter yang dibutuhkan algoritma *Harmony Search*, parameter-parameter tersebut antara lain : *Harmony Memory Search* (HMS), *Harmony Memory Considering Rate* (HMCR), *Pitch Adjustment Rate* (PAR), dan *Number of Improvisation* (NI)
   b. Inisialisasi *Harmony Search*.
      Solusi yang dihasilkan *Harmony Search* akan disimpan dalam suatu memori yang disebut dengan *Harmony Memory* (HM).
   c. Improvisasi *Harmony* Baru.
   d. Update *Harmony Memori.*
5. Implementasi dan pembangunan sistem.
6. Pengujian dan analisa hasil
   a. *Testing* sistem, menguji sitem dengan menggunakan data berupa single dokumen teks.
   b. Analisa hasil, performansi ringkasan ditinjau dengan membandingkan antara hasil peringkasan otomatis dan ringkasan referensi dengan menggunakan ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) *evaluation toolkit*.
   c. Analisa parameter HS dapat mengoptimasi ringkasan dengan optimal.

7. Pembuatan laporan, melakukan pelaporan hasil pengerjaan tugas akhir berupa analisis sistem yang dibangun beserta dokumentasinya serta kesimpulan akhir.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Berdasarkan analisis terhadap hasil pengujian, diperoleh kesimpulan sebagai berikut:

1. Parameter *Harmony Memory Consideration Size* (HMCR) pada *Harmony Search* yang menghasilkan ringkasan yang optimum global adalah HMCR = 0.9 pada dokumen yang diujikan jika jumlah dokumennya besar. Pada dokumen yang memiliki jumlah kalimat yang kecil dibawah 20 tidak perlu digunakan HMCR yang lebih besar, dikarenakan HMCR dibawah 0.9 sudah dapat menghasilkan ringkasan yang global optimum. Dan jika parameter HMCR mendekati 1 hasilnya tidak lebih bagus dari 0.9 dikarenakan proses diversifikasi (eksplorasi) semakin kecil.
2. Dengan HMCR sebesar 0.9 maka pada saat improvisasi HMCR lebih banyak menggunakan kombinasi vektor solusi yang ada di *Harmony Memory*.

.

5.2 Saran

Berdasarkan hasil analisis dan kesimpulan, terdapat beberapa saran untuk perbaikan pada penelitian peringkasan teks sebagai berikut:

1. Pada inisialisasi *Harmony Memory* diinputkan beberapa kombinasi kalimat awal untuk dokumen artikel bahasa Indonesia sehingga diperoleh hasil ringkasan yang jauh lebih optimal.

Daftar Pustaka

[1] Chin-Yew Lin. 2004. *ROUGE: A Package for Automatic Evaluation of Summaries*. In Proceedings of the ACL-04 Workshop. Barcelona, Spain, pages 74 - 81.

[2] E. Shareghi and L. S. 2008. Hassanabadi.*Text Summarization with Harmony Search Algorithm-Based Sentence Extraction.* In Proceedings of the 5th International Conference on Soft Computing as Transdisciplinary Science and Technology, page 226-331.

[3] E. H. Hovy, 2001, "*Automated Text Summarization Information Sciences Institute Press,* University of Southern California.

[4] Geem, Zong Woo. "*Optimal cost design of water distribution networks using harmony search*". Environmental Planning and Management Program, Johns Hopkins University.

[5] Geem, Zoong Woo. 2009. *Music-Inspired Harmony Search Algorithm.* Springer.

[6] Hongyan Jing, et, al. 1998. *Summarization Evaluation Methods: Experiments and Analysis*. In Proceedings of the AAA1 Intelligent Text Summarization Workshop. page 60-68.

[7] Husni, M, and Zaman, B. 2005. *Perangkat Lunak Peringkas Dokumen Berbahasa Indonesia Dengan Hybrid Stemming*. Petra Christian University, Surabaya.

[8] Inderjeet Mani, Mark T. Maybury, 2001, "Automatic Summarization: Turorial Notes". American/European Conference on Computational Linguistics (ACL/EACL '01) Toulouse, France. http://mitre.org/resources/centers/it/maybury/summarization/summarization.htm

[9] Mahdavi, M. et all. *An improved harmony search algorithm for solving optimization problems*. Applied Mathematics and. Computation, Vol. 188, page 1567–1579, 2007.

[10] Radev D., Allison T., Blair-Goldensohn S., Blitzer J., Celebi A., Drabek E., Lam W., Liu D., Qi H., Saggion H., Teufel S., Topper M. and A. Winkel, *The MEAD Multidocument Summarizer*, MEAD Documentation v3.08, 2003.

[11] Suyanto. 2010. *Algoritma Optimasi Deterministik atau Probabilitic*. Yogyajarta : Graha Ilmu.

[12] Tang, J., Yao, L., and Chen, D. *Multi-topic Based Query-Oriented Summarization*. In Proceedings of SDM. 2009, 1147-1158.

[13] Ziheng ,Lin. 2007. *Graph-Based Methods for Automatic Text Summarization.* Department of Computer Science School of Computing National University of Singapore, Singapore.