PERINGKAS TEKS OTOMATIS PADA BERITA SINGLE-DOKUMEN DENGAN MENGGUNAKAN ALGORITMA LINTASAN TERPENDEK (AUTOMATIC TEXT SUMMARIZATION ON SINGLE-DOCUMENT NEWS USING SHORTEST PATH ALGORITHM)

Fathonah Arin Firmawati[1], Adiwijawa[2], Imelda Ataina[3]

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

---

Abstrak

Peringkas Teks Otomatis adalah proses meringkas teks menjadi versi lebih singkat. Terdapat dua tipe teks sebagai masukan sistem peringkas, yaitu singledocument dan multi-document. Sedangkan metode peringkasan yang dapat digunakan adalah abstraksi dan ekstraksi.

Pada Tugas Akhir ini diimplementasikan peringkasan ekstraksi dengan menggunakan algoritma lintasan terpendek. Dan metode ini berbasis graf, dimana kalimat adalah titik (simpul) dan relasi kalimat adalah sisi (edge). Ringkasan akan diperoleh dengan mencari lintasan terpendek pada graf yang telah dihitung cost masing-masing sisi. Kemudian kalimat-kalimat pada teks yang masuk dalam lintasan terpendek, akan diekstrak sebagai ringkasan.

Pengujian pada sistem ini menggunakan evaluasi ROUGE. Pada hasil pengujian menunjukkan bahwa early(j) pada teks berita berbahasa Indonesia bernilai 2 untuk indeks kalimat (j) kurang dari sama dengan 5 dan bernilai 1 untuk indeks yang lainnya. Ringkasan pada teks asli dan teks modified menghasilkan akurasi yang hampir sama untuk setiap skenario.

Kata Kunci : shortest path, peringkasan teks, stopword, graf berarah dan berbobot.

Abstract

Automated Text Summarization is the process of summarizing the text into shorter versions. There are two types of text that can be an input the system, there are single-document and multi-document. While summarizing method that can be used are abstraction and extraction.

In this final project is implemented summarizing extraction using the shortest path algorithm. This method is based on a graph, where the sentence is the point (node) and the relationship between sentences is side (edge). A summary will be obtained by finding the shortest path in a graph which has calculated the cost of each side. Then the sentences in the text that included in a shortest path, will be extracted as a summary.

The test on this system is using evaluation ROUGE. In the test results indicate that early (j) in Indonesian language news text has value 2 for index sentence (j) is less than equal to 5 and the value 1 for the others. The summary of the original text and modified text produce nearly same accuracy for each scenario.

Keywords : shortest path, text summarization, stopword, weighted directed graph.

---

1. Pendahuluan

Pada bab ini akan dijelaskan latar belakang, perumusan masalah, tujuan, beserta metedologi penyelesaian masalah pada Tugas Akhir.

1.1 Latar Belakang

Pada saat ini perkembangan teknologi informasi semakin meningkat seiring berkembangnya media elektronik yang menyediakan berbagai informasi secara *on-line*. Salah satu bentuk informasi tersebut adalah dokumen maupun artikel berita. Kebutuhan *user* akan dokumen yang berupa berita menyebabkan *user* membutuhkan waktu yang lebih lama untuk membaca keseluruhan dokumen. Oleh karena itu, dibutuhkan informasi yang singkat dan padat yang merepresentasikan isi dokumen. Sehingga dikembangkan sebuah sistem yang dapat meringkas dokumen yaitu, peringkas teks otomatis *(automatic text summarization)*.

Banyak metode maupun pendekatan dalam meringkas teks diantaranya berdasarkan cara pengambilan ringkasan adalah metode ekstraksi dan abstraksi. Ekstraksi adalah metode peringkasan kalimat dengan langsung mengekstrak kalimat yang dianggap penting dari teks aslinya, sedangkan abstraksi mengandung *reformulation* kalimat dari teks asli [3]. Metode ekstraksi lebih mudah diterapkan dalam peringkasan teks karena tidak melibatkan *natural language processing*. Kemudian dilihat dari dokumen sebagai masukan, terdapat dua macam dokumen, yaitu *single-document* dan *multi-document*. Pada *single-document* sumber ringkasan hanya terdiri dari sebuah dokumen, sedangkan *multi-document* sumber ringkasan terdiri dari minimal dua dokumen atau lebih yang mempunyai topik yang sama [3].

Pada tugas akhir ini, dibangun dan diimplementasikan *Shortest path algorithm* yang merupakan algoritma berbasis graf untuk melakukan peringkasan dengan metode ekstraksi pada *single-document*. Metode ini telah berhasil digunakan untuk meringkas teks berbahasa Inggris dan mudah dalam implementasi, serta tidak bergantung pada suatu bahasa tertentu. Karena kelebihannya yang tidak bergantung pada suatu bahasa tertentu, maka metode peringkasan dengan menggunakan algoritma lintasan terpendek akan coba diimplementasikan pada teks berbahasa Indonesia. *Shortest path algorithm* berbasis pada pencarian lintasan terpendek dari kalimat pertama hingga kalimat terakhir pada sebuah graf yang merepresentasikan teks asli. Simpul pada graf mewakili kalimat dan sisi mewakili kemiripan antar kalimat [7]. *Word overlap* atau jumlah kata yang sama antara dua kalimat digunakan untuk menentukan kemiripan. Oleh karena itu, jika ada dua kalimat yang memiliki minimal satu kata yang sama maka akan ada sisi diantara nya [7].

Kemudian setelah graf terbentuk dan *cost* pada setiap sisi telah dihitung, ringkasan dapat diekstrak berdasarkan kalimat-kalimat yang terpilih pada lintasan terpendek. Dokumen sebagai data *input-an* sistem merupakan artikel berita yang didapat dari beberapa situs internet.

1.2 Perumusan Masalah

Pada tugas akhir ini masalah yang diselesaikan yaitu:

1. Bagaimana mengimplementasikan algoritma lintasan terpendek untuk menghasilkan ringkasan pada peringkasan dokumen.
2. Melakukan analisis, apakah metode ini menghasilkan ringkasan yang mempresentasikan isi secara jelas dan juga memberikan informasi-informasi yang penting dari teks asli?

Adapun batasan masalah pada tugas akhir ini yaitu:

1. Data yang digunakan adalah dokumen tunggal berita berbahasa Indonesia.
2. Peringkasan dilakukan secara *off-line*.
3. Hasil ringkasan merupakan ekstraksi dari kalimat-kalimat yang terdapat pada teks asli.
4. Judul setiap teks menjadi kalimat pertama atau kalimat awal teks tersebut.
5. Representasi graf yang digunakan adalah graf berarah dan berbobot.
6. Hasil ringkasan sistem yang diperoleh dibandingkan dengan hasil ringkasan manual yang dibuat oleh manusia.

1.3 Tujuan

Tujuan dari Tugas Akhir ini adalah :

1. Membangun sistem peringkas otomatis dengan menggunakan *shortest path algorithm* yang dapat menghasilkan ringkasan sebagai representasi dari dokumen yang diringkas.
2. Mengevaluasi hasil ringkasan yang dihasilkan oleh sistem peringkas otomatis yang dibangun dari *shortest path algorithm* dengan membandingkan terhadap ringkasan manusia menggunakan *tool* ROUGE.

1.4 Metodologi Penyelesaian Masalah

Metodologi yang dilakukan untuk menyelesaikan Tugas akhir ini adalah :

1. Melakukan studi literatur
   Khususnya mengenai peringkasan teks otomatis, *shortest path algorithm*, dan algoritma yang berbasis graf.
2. Pencarian dan Pengumpulan data
   Mengumpulkan dokumen-dokumen berita yang diperlukan untuk mendukung penyelesaian masalah. Data diperoleh dari situs-situs berita seperti kompas dan detik.
3. Analisis kebutuhan dan perancangan
   a. Mempelajari *shortest path algorithm* untuk peringkasan teks yang akan digunakan dalam implementasi perangkat lunak.
   b. Melakukan perancangan sistem yang akan dibangun.
   c. Mempelajari *tool* ROUGE yang akan digunakan sebagai analisa hasil ringkasan sistem.

4. Implementasi
   a. Membangun sistem yang telah dirancang sebelumnya dengan bahasa pemrograman PHP.
   b. Dokumen sebagai masukan sistem berekstensi file (.txt).
5. Pengujian dan analisis hasil
   Melakukan pengujian sistem dan menganalisa hasil keluaran sistem yang berupa ringkasan teks, apakah dapat merepresentasikan informasi utama pada teks asli.
6. Pengambilan kesimpulan dan penyusunan laporan tugas akhir.

Telkom University

5. Penutup

Pada bagian ini akan diambil kesimpulan dari hasil analisis pada bab sebelumnya (Bab 4.3) dan juga saran sebagai perbaikan untuk penelitian selanjutnya.

5.1 Kesimpulan

Berdasarkan analisis terhadap hasil pengujian, diperoleh kesimpulan sebagai berikut :

1. Parameter *early(j)* pada sistem peringkasan teks berita berbahasa Indonesia bernilai 2 untuk indeks kalimat (j) yang kurang dari sama dengan 5 dan bernilai 1 untuk indeks yang lainnya.
2. Secara keseluruhan, ROUGE-1, ROUGE-2, dan ROUGE-L menghasilkan akurasi yang lebih baik daripada ROUGE-W dan ROUGE-S. Ini disebabkan karena pada ROUGE-1, ROUGE-2, dan ROUGE-L perhitungannya tidak memperhatikan urutan sehingga menyebabkan pasangan kata yang ada berjumlah banyak. Tidak seperti pada ROUGE-W yang memperhatikan urutan kata dan ROUGE-S yang memperhatikan pasangan kata *(skip-bigram overlap).*
3. Pada setiap parameter ROUGE, nilai akurasi hasil peringkasan teks berbahasa Indonesia pada teks asli dan teks modified memiliki nilai yang hampir sama untuk setiap skenario.
4. Peringkas teks otomatis dengan menggunakan algoritma lintasan terpendek dapat diimplementasikan pada peringkasan teks berbahasa Indonesia.

5.2 Saran

Berdasarkan analisis dan kesimpulan yang didapat, berikut ini beberapa saran untuk perbaikan penelitian selanjutnya:

1. Mencoba melakukan pemilihan kalimat terakhir pada teks modified secara otomatis oleh sistem yang dibangun.
2. Menambah atau memperbaiki kata-kata pada daftar *stopwords* sehingga peran eliminasi *stopwords* dapat memberikan akurasi yang lebih baik.

Referensi

[1] Hovy, Eduard. 2001. *"Automated Text Summarization"*. Handbook of computation linguistics, Oxford University Press. Available on: http://www.isi.edu/natural-language/people/hovy/papers/05Handbook-Summ-hovy.pdf (10 Januari 2009).

[2] Johnsonbaugh, R. 2002. *"Matematika Diskrit (jilid2)"*. Pearson Education Asia Pte. Ld & PT Prenhallindo. Jakarta.

[3] Kaushik, S. & Luthra P. *"Automatic Text Summarization (ATS)"*. Department of Computer Science & Engineering, IIT Delhi. Available on : 3 Desember 2008.

[4] Lin, Chin-Yew. 2004. "ROUGE: A Package for Automatic Evaluation of Summaries". Available on : 1 januari 2009.

[5] Mihalcea, Rada. *"Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization"*. Department of Computer Science University of North Texas. Available on : http://acl.ldc.upenn.edu/P/P04/P04-3020.pdf (31 Desember 2008).

[6] Munir, Rinaldi. 2005. *"Matematika Diskrit (edisi 3)"*. Penerbit : Informatika Bandung. Bandung.

[7] Sjöbergh, J. & Araki, K. *"Extraction based Summarization Using a Shortest Path Algorithm"*. Available on: http://dr-hato.se/research/shortpath.pdf (20 Februari 2010).

[8] ___. Bellman-Ford Algorithm. Available on : http://en.wikipedia.org/wiki/graph/Bellman-Ford_algorithm.htm (23 Juli 2010).

[9] ___. Djikstra Algorithm. Available on : http://en.wikipedia.org/wiki/graph/Dijkstra's_algorithm.htm (23 Juli 2010).

[10] ___. Floyd Warshall Algorithm. Available on : http://en.wikipedia.org/wiki/graph/Algoritma_floyd-warshall.htm (29 Juni 2010).

[11] ___. Stopword List Yudi Wibisono. Available on : http://fpmipa.upi.edu/staff/yudi/stop_words_list.txt (24 Juni 2010).