ANALISIS PERBANGAN PEMBOBOTAN KATA PERINGKAS TEKS OTOMATIS

Khairul Ihsan[1], Hetti Hidayati[2], Yanuar Firdaus A.w.[3]

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

Abstrak

Text Summarization adalah sebuah proses untuk menghasilkan ringkasan (summary) dari suatu artikel tapi tetap memiliki gambaran yang akurat dari isi suatu artikel. Tujuannya adalah mengambil sumber informasi dengan mengutip sebagian besar isi yang penting dan menampilkan kepada pembaca dalam bentuk yang ringkas dan sesuai dengan kebutuhan pembaca.
Pada tugas akhir ini mengimplementasikan metode dengan TF-ISF (Term Frequency*Inverse Sentence Frequency) [12] yaitu salah satu metode pada text summarization yang menghasilkan keluaran berupa ringkasan ekstraktif yang terdiri dari kalimat-kalimat beranking tinggi. Sebagai pembanding digunakan metode peringkasan teks dengan TF-IDF (Term Frequency*Inverse Document Frequency) dengan menambahkan konsep ekstraksi frase utama (Keyphrase Extraction) [10] dari suatu teks. Hasil ringkasan yang dihasilkan tetap memiliki/ mengandung bagian-bagian yang penting dari artikel asli sehingga diharapkan dapat membantu pembaca untuk menyerap informasi yang ada dalam artikel melalui ringkasan (summary).
Hasil pengujian menunjukkan bahwa proses summary sangat bergantung pada jenis artikel/dokumen, jumlah kalimat yang dimiliki dan panjang summary yang diinginkan. Performansi dari sistem dinilai menggunakan Precision, Recall dan F-Measure menunjukkan bahwa metode TF-ISF memiliki performansi yang sama dengan TF-IDF yang digunakan pada single document, sedangkan hasil summary pada multidocument menunjukkan bahwa metode TF-IDF dengan Keyphrase Extraction memiliki performansi lebih baik daripada TF-ISF dengan Keyphrase Extraction, karena didukung oleh faktor meningkatnya nilai kemunculan term dan frase utama.

Kata Kunci : Text Summarization, Ekstraksi Frase Utama, TF-IDF, TF-ISF.

Abstract

Text summarization is a process to produce a summary of an article but still have an accurate description of the contents of an article. Objective of this process is to take the source of information by citing most of the important content and show it to the readers in a simple form that appropriate with reader's need.
In this final task is to implement the method of TF-ISF (Term Frequency * Inverse Sentence Frequency) [12] which is one method of text summarization that produces output in the form of extractive summaries of high rank sentences. As a comparison of TF-ISF method, used another text summarization method is TF-IDF (Term Frequency * Inverse Document Frequency) modified using keyphrase extraction[10] concept. Thus, hopefully this system can help readers to get informations from articles through a summary. Text Summarization will produce a text that still has the main points from the original articles
The test results showed that the summary process depends on the type of articles / documents, total sentences and the summary length. System's performance results analyzed using Precision, Recall, and F-Measure showed that TF-ISF method has the same performance with the TF-IDF in processing single document. Otherwise for multidocument summary process, TF-IDF method with Keyphrase Extraction has better performance than TF-ISF method with Keyphrase Extraction, because it is supported by the increasing value of term occurrences and keyphrase.

Keywords : Text Summarization, Keyphrase Extraction, TF-IDF, TF-ISF.

1. Pendahuluan

1.1 Latar Belakang

Automatic text summarization (ATS) atau yang sering disebut peringkas teks otomatis adalah sebuah proses menyaring informasi paling penting dari sebuah sumber atau beberapa sumber untuk membuat sebuah versi ringkas dari teks [15]. Riset peringkasan teks dilakukan pertama kali oleh Luhn pada tahun 1950-an. Sejak itu, mulai berkembang penelitian-penelitian peringkasan teks serta banyak *workshop* dan konferensi peringkasan teks terus dilakukan di seluruh dunia sampai sekarang.

Terdapat dua buah pendekatan dilihat dari teknik pengambilan ringkasan [14], yaitu ekstraksi (*shallower approaches*) dan abstraksi (*deeper approaches*). *Shallower approaches,* kalimat atau kata yang ada dalam ringkasan diambil secara utuh dari dokumen aslinya sedangkan *deeper approaches,* ringkasan yang menambahkan kata baru dan dapat merubah sususan kalimat. Pada umumnya, abstraksi dapat meringkas teks lebih kuat daripada ekstraksi, tetapi sistemnya lebih sulit dikembangkan karena mengaplikasikan teknologi *natural language generation* yang merupakan bahasan yang dikembangkan tersendiri.

Dari beberapa tugas akhir yang telah ada [11], secara umum membahas tentang penggunaan metode TF-IDF. Dari pengujian menunujukkan bahwa system tersebut lebih mengidentifikasi dokumen. Oleh karena itu, perlu dilakukan penanganan di mana system mengidentifikasi kalimat dibandingkan dengan dokumen, dengan menggunakan skema pembobotan pada teks yang didasarkan pada jumlah kalimat.

Secara umum terdapat dua jenis metode pemilihan kalimat [12] sebagai hasil ringkasan yaitu metode yang tidak melakukan perhitungan bobot kata dan metode yang melakukan perhitungan bobot kata. Metode yang tidak menghitung bobot kata hanya mengambil beberapa kalimat awal dan akhir, sedangkaan metode yang menghitung bobot kata adalah menggunakan bobot term (kata maupun pasangan kata) dari setiap term yang terdapat dalam kalimat tersebut. Beberapa metode pembobotan diantaranya adalah TF-ISF *(Term Frequency-Inverse Sentence Frequency)* dan TF-IDF *(Term Frequency*Inverse Document Frequency).* Tugas akhir ini mengimplementasikan metode pembobotan kata terhadap dokumen berbahasa Indonesia dengan menggunakan *Keyphrase Extraction.*

Studi kasus yang diambil adalah artikel berita karena merupakan salah satu jenis dokumen teks yang banyak dibutuhkan orang dan dapat dengan mudah diperoleh di internet sebagai salah satu alternatif membaca berita dari koran.

1.2 Perumusan Masalah

Rumusan masalah yang akan dikaji dalam tugas akhir ini adalah :

1. Bagaimana algoritma pembobotan kata dapat menemukan kalimat-kalimat yang mengandung informasi paling penting dari sebuah teks.
2. Membandingkan kinerja metode pembobotan kata pada artikel yang menggunakan Ekstraksi Frase Utama.

Adapun beberapa batasan-batasan dalam tugas akhir ini yaitu :

1. Jenis ringkasan yang dihasilkan berupa *extractive summary*.
2. Koleksi dokumen yang digunakan untuk pengujian berupa artikel berita berbahasa Inggris dengan file berekstensi *.txt* yang diunduh dari DUC *2002 (Document Understanding Conference)*.
3. *Thesaurus* tidak didefinisikan dan ditangani secara khusus, karena merupakan unsur bawaan dari relasi semantik yang dihasilkan oleh WordNet.
4. Pada sistem *term parsing* tidak menangani kesalahan dalam pengetikan kata.
5. Proses *Automated Text Summarization* dilakukan secara *Off-line*.
6. Jenis ringkasan yang dihasilkan berupa *extractive summary* dengan kisaran panjang *summary* antara 10% - 50%.

1.3 Tujuan

Tujuan yang ingin dicapai dalam pelaksanaan Tugas Akhir ini adalah :

1. Mengimplementasikan metode pembobotan kata menjadi sebuah sistem peringkas otomatis untuk dokumen berita.
2. Membandingkan metode pembobotan kata menggunakan ekstraksi frase utama.
3. Membandingkan subjektif tes sebagai pelengkap pengujian validitas hasil *summary*.

1.4 Metodologi penyelesaian masalah

Metodologi yang digunakan dalam pelaksanaan tugas akhir ini, yaitu:

a. Studi Literatur mengenai *text summarization* dengan menggunakan *Keyphrase Extraction* dan Pembobotan Kata serta melakukan studi dari berbagai pustaka seperti *text book*, jurnal ilmiah, dan artikel *web* yang dapat menunjang tugas akhir ini.

b. Analisis mengenai permasalahan yang ada pada *text summarization*, *Keyphrase Extraction* dan Pembobotan Kata.

c. Perancangan solusi berdasarkan analisis permasalahan yang sudah didapatkan seperti perancangan objek dan kelas yang terlibat dalam *summary system*, perancangan basis data, perancangan antarmuka

d. Implementasi, membangun hasil rancangan ke dalam kode program.

e. Melakukan pengujian sistem dan menganalisa hasil keluaran sistem yang berupa ringkasan teks, sejauh mana ringkasan dapat menggambarkan makna utama teks dan mengujinya dengan ROUGE *evaluation toolkit*, dan mempelajari relevansi hasilnya dengan hasil subjektif tes.

f. Penyusunan laporan dan penarikan kesimpulan, melakukan penyusunan laporan dan penarikan kesimpulan terhadap perangkat lunak yang dibuat serta pemberian saran terhadap pengembangan perangkat lunak ini kedepannya dalam bentuk tertulis sebagai laporan penelitian.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Beberapa kesimpulan yang dapat diambil dari Tugas Akhir ini yaitu :

1. Pembobotan kalimat pada metode TF-IDF dengan *Keyphrase Extraction* dan TF-ISF dengan *Keyphrase Extraction* nilai kalimat bergantung dari faktor kemunculan *keyphrase* dan bobot nilai *full weighting*.
2. Berdasarkan dokumen uji yang digunakan, term yang merupakan frase utama pada multi dokumen memiliki nilai TF*IDF yang tinggi, hal ini berbeda dengan TF*ISF, dimana term yang memiliki nilai tertinggi belum tentu merupakan *keyphrase*.
3. Berdasarkan jenis dokumen input, secara keseluruhan peringkasan teks menggunakan metode TF-IDF dengan *Keyphrase Extraction dan* TF-ISF dengan *Keyphrase Extraction* memiliki nilai performansi sama yang digunakan pada *single document* karena jumlah dokumen yang digunakan sama.
4. Untuk proses peringkasan teks *multidocument* metode TF-IDF dengan *Keyphrase Extraction* memiliki performansi lebih baik daripada TF-ISF dengan *Keyphrase Extraction*, karena didukung oleh meningkatnya nilai kemunculan suatu term dan frase utama pada jumlah kalimat yang banyak.

5.2 Saran

Saran-saran yang dapat penulis uraikan untuk keperluan analisis selanjutnya adalah:

1. Sebaiknya proses *keyphrase extraction* diintegrasikan dengan *summary system* yaitu dengan menambah kelengkapan *vocabulary* yang disediakan dengan harapan dapat memperbaiki kinerja sistem dan mengurangi ketergantungan sistem terhadap aplikasi *keyphrase extractor*.
2. Memperbanyak jumlah dokumen uji dengan memperhitungkan optimalitas waktu proses dari sistem.
3. Menganalisis dan menambah jumlah human summary untuk mengetahui performansi dari metode TF-IDF dan TF-ISF.

Daftar Pustaka

[1] Barzilay, Regina., Michael Elhadad. Using Lexical Chains for Text Summarization. Mathematics and Computer Science Dept. Ben Gurion University in the Negev Beer-Sheva. 1997.

[2] Biryukov, Maria. Multidocument Question Answering Text Summarization using Topic Signatures. 23rd August 2004.

[3] Dias,Gael dan Elsa Alves. Topic Segmentation Algorithms for Text Summarization and Passage Retrieval: An Exhaustive Evaluation . Link download :http://www.aaai.org/Papers/AAAI/2007/AAAI07-211.pdf

[4] Foltran, L Alfredo and Ana Ozaki Rivera. Comparison of text sets using Data Mining and Similarity Measure Methods .Link Download : http://unbproteus.googlecode.com/svn/trunk/latex/relatorio.pdf

[5] Erkan, Günes., Dragomir R. Radev. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. School of Information & Department of EECS. University of Michigan, Ann Arbor, MI 48109 USA. 2004.

[6] Manning, Christopher., Prabhakar Raghavan, Hinrich Schütze. Introduction to Information Retrieval. Cambridge University Press. 2008.

[7] Saggion, Horacio. Introduction to Text Summarization and Other Information Access Technologies. Department of Computer Science University of Sheffield England, United Kingdom. 2008.

[8] Turney, Peter D., Learning Algorithms for Keyphrase Extraction. Institute for Information Technology National Research Council of Canada. 1999.

[9] Nurzaitun Purwasih. 2009. Peringkasan Teks Otomatis Dokumen Tunggal Berbahasa Indonesia Menggunakan Graph-based Summarization Algorithm dan Similarity. Fakultas Informatika Institut Teknologi Telkom.

[10] Thiago Alexandre Salgueiro Pardo, Lucia Helena Machado Rino and Maria das Graças Volpe Nunes. 2008. Extractive summarization: how to identify the gist of a text.

[11] Thiago Alexandre Salgueiro Pardo, Lucia Helena Machado Rino and Maria das Graças Volpe Nunes. http://fportfolio.petra.ac.id/user_files/92-008/Rolly-TI-Jurnal-June%202006(new).pdf

[12] Teguh Cahyono. 2004. Metode-metode Text Summarization http://www.hansmichael.com/default.asp?cat=taT

[13] Budhi, Gregorius S. et.al. Indonesian Automated Text Summarization. Link download:http://www.fportfolio.petra.ac.id/user_files/92-008/Paper%20Automated%20Text%20Summarization%20(Greg-Rolly)%20final2.pdf