Peringkasan Dokumen dengan Metode Non-Negative Matrix Factorization

Herastia Maharani[1], Monica Sanjaya[2]

Departemen Teknik Informatika, Institut Teknologi Harapan Bangsa

Jalan Dipatiukur no.80-84,Bandung - Indonesia

[1]herastia@ithb.ac.id

[2]monica.sanjaya@ymail.com

Abstrak — **Peringkasan dokumen teks merupakan pengambilan informasi dari sebuah dokumen teks dengan mengambil sebagian teks dalam dokumen yang dianggap mengandung informasi paling penting. Penelitian ini mengimplementasikan metode *Non-negative Matrix Factorization* (NMF) untuk melakukan peringkasan pada dokumen dalam bahasa Indonesia dan bahasa Inggris. Pengujian dilakukan dengan membandingkan hasil ringkasan metode NMF ini dengan hasil ringkasan yang diperoleh dari sebuah situs peringkasan.**

Kata kunci* — peringkasan, dokumen teks, *non-negative matrix factorization

Abstract — ***Text document summarization is the process of extracting information from text document by selecting particular parts from the documents which are considered to have the most important information. This research used Non-negative Matrix Factorization (NMF) method to summarize documents written in Bahasa Indonesia and also in English. Evaluation is performed by comparing the summary extracted from NMF with summary obtained from a summarization website available on the internet.***

Keywords*— summarization, text document, non-negative matrix factorization

I. PENDAHULUAN

Perkembangan teknologi seperti dengan adanya internet, lebih memudahkan manusia dalam mencari dan menemukan informasi yang dibutuhkannya. Karena perkembangan ini pula, jumlah dokumen terutama dokumen teks secara eksplosif meningkat dari hari ke hari, dan menjadi sangat sulit untuk mengelola informasi yang ada dengan membaca semua teks.

Peringkasan sebuah artikel merupakan sebuah cara pengambilan informasi dari sebuah dokumen teks dengan mengambil sebagian teks dalam dokumen yang dianggap mengandung informasi paling penting. Selain itu, dengan adanya ringkasan, manusia dapat dengan mudah dan lebih cepat mengerti dan memahami isi sebuah dokumen tanpa harus membaca keseluruhan dokumen yang pasti memerlukan waktu yang lebih lama.

Maka dari itu diperlukan sebuah sistem yang mampu meringkas sebuah dokumen teks secara otomatis. Peringkasn dokumen adalah sebuah proses mereduksi sebuah dokumen ke dalam versi yang lebih pendek tanpa kehilangan informasi yang terkandung di dalamnya. Jika dilihat dari bentuk ringkasan yang dihasilkan, peringkasan dibagi menjadi dua jenis, yaitu ekstraksi dan abstraksi. Hasil peringkasan ekstraktif berisi bagian-bagian dari dokumen asli yang dianggap mewakili isi dokumen, contohnya paragraf atau kalimat yang dianggap penting. Sedangkan hasil peringkasan abstraktif tidak mengambil bagian teks asli dari dokumen, melainkan berupa hasil pembentukan struktur baru sesuai dengan struktur tata bahasa yang digunakan.

Sudah ada banyak metode yang diimplementasikan untuk peringkasan sebuah dokumen teks, seperti *word cluster*, *learning algorithm*, *vectorial approach*, *fuzzy approach*, *genetic algorithm*, dan lain-lain. Salah satu metode yang tidak tergantung pada karakteristik bahasa yang digunakan adalah *Non-negative Matrix Factorization* (NMF). Penelitian ini menggunakan metode NMF untuk dokumen bahasa Indonesia dan bahasa Inggris untuk melihat apakah ada pengaruh dari perbedaan bahasa terhadap presisi ringkasan yang dihasilkan.

II. PERINGKASAN DOKUMEN TEKS

Salah satu faktor yang berperan penting dalam peringkasan dokumen adalah aturan tata bahasa yang digunakan. Perbedaan tata bahasa membuat banyak metode dikembangkan dengan berdasarkan pada aturan bahasa tertentu sehingga belum tentu bisa digunakan untuk bahasa lainnya. Akan tetapi, ada juga metode peringkasan yang tidak bergantung kepada jenis bahasa yang dipakai, salah satunya adalah *Non-negative Matrix Factorization* (NMF) [4],[3]. Keunggulan metode ini adalah kemampuannya untuk meringkas dokumen dalam berbagai bahasa, sehingga menjadikan peringkasan secara otomatis tidak lagi terbatas bahasa-bahasa tertentu saja. Selain itu metode NMF termasuk ke dalam pembelajaran tidak terawasi (*unsupervised learning*) sehingga proses peringkasan tidak bergantung pada kualitas data pelatihan seperti pada pembelajaran terawasi (*supervised learning*).

Sebelum proses peringkasan dapat dimulai, dilakukan tahap *pre-processing* yang meliputi *stopword removal* dan *stemming*. Dalam tahap *stopword removal*, kata-kata yang dianggap terlalu umum dan tidak memiliki makna yang signifikan terhadap informasi yang dikandung dalam dokumen akan dihilangkan. Contoh *stopword* dalam Bahasa Indonesia antara lain “adalah”, “di”, “ke”, dsb. Sedangkan dalam Bahasa Inggris contohnya adalah “of”, “in”, “thus”, dsb. Adapun

tahap *stemming* merupakan tahap konversi sebuah kata ke dalam bentuk dasarnya dengan cara menghilangkan imbuhan dalam kata tersebut. Dalam penelitian ini digunakan algoritma Porter [5] untuk melakukan *stemming* terhadap dokumen Bahasa Inggris dan algoritma CS Stemmer [1] untuk melakukan *stemming* terhadap dokumen Bahasa Indonesia.

Non-negative Matrix Factorization (NMF) merupakan metode untuk mendekomposisi matriks *term-by-sentence* non-negatif A yang berukuran $m \times n$ menjadi 2 buah matriks, yaitu *Non-negative Semantic Feature Matrix* (NSFM), W, dengan ukuran $m \times r$, dan *Non-negative Semantic Variable Matrix* (NSVM), H, dengan ukuran $r \times n$ [4]. Matriks A merupakan matriks yang berisi bobot *term* dalam kalimat dan berukuran jumlah *term* (m) × jumlah kalimat (n). Metode dekomposisi dengan NMF dapat dinyatakan dalam persamaan yang diberikan di [4].

$$A \approx WH \tag{1}$$

Proses menuju kondisi $A \approx WH$ dilakukan berdasarkan aturan Frobenius Norm [4] yaitu dengan menghitung jarak antara matriks A dengan hasil perkalian W dan H menggunakan persamaan berikut [4].

$$\Theta_E(W,H) \equiv \|A - WH\|_F^2 \equiv \sum_{j=1}^{m}\sum_{i=1}^{n}\left(A_{ji} - \sum_{l=1}^{r} W_{jl}H_{li}\right)^2 \tag{2}$$

Nilai elemen matriks W dan H akan terus-menerus di-*update* untuk mencapai kondisi $A \approx WH$, dengan menggunakan aturan *multiplicative update* yang diberikan di [3].

$$H_{\alpha\mu} \leftarrow H_{\alpha\mu}\frac{(W^T A)_{\alpha\mu}}{(W^T WH)_{\alpha\mu}} \tag{3}$$

$$W_{i\alpha} \leftarrow W_{i\alpha}\frac{(AH^T)_{i\alpha}}{(WHH^T)_{i\alpha}} \tag{4}$$

Setelah kondisi $A \approx WH$ tercapai, proses ekstraksi kalimat ringkasan dilakukan dengan mengambil n kalimat dengan skor kalimat tertinggi. Skor kalimat diperoleh dari perhitungan *Generic Relevance of Sentence* (GRS) setiap kalimat pada elemen-elemen matriks H dengan menggunakan persamaan yang diberikan di [4].

$$GRS_j = \sum_{i=1}^{r}\left(H_{ij}.weight(H_{i*})\right) \tag{5}$$

Keterangan:
r = jumlah baris pada matriks H
i = indeks baris, dengan $1 < i < r$
H_{ij} = elemen matriks H pada posisi (i,j)

Nilai GRS_j menunjukkan skor untuk setiap vektor kolom ke-j pada matriks H. Sedangkan nilai *weight* merupakan bobot untuk elemen (i,j) pada matriks H.

$$weight(H_{i*}) = \frac{\sum_{q=1}^{n} H_{iq}}{\sum_{p=1}^{r}\sum_{q=1}^{n} H_{pq}} \tag{6}$$

Keterangan:
n = jumlah kalimat
q = indeks kolom kalimat, dengan $1 < q < n$
H_{iq} = elemen-elemen matriks H pada posisi baris i tertentu
p = indeks baris, dengan $1 < p < r$
H_{pq} = elemen-elemen matriks H pada posisi (p,q), yang merupakan keseluruhan elemen pada matriks H

III. Rancangan Sistem

Alur kerja sistem yang dibangun secara umum diberikan di Gambar 1. Sebagaimana telah dijelaskan di Bagian II, *pre-processing* meliputi penanganan *stopwords* dan *stemming*. Proses peringkasan dokumen sendiri dimulai dengan pembentukan matriks A hingga ekstraksi kalimat ringkasan berdasarkan nilai GRS tertinggi.

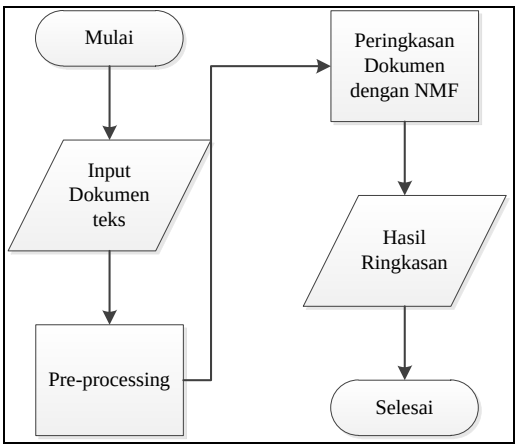
IV. Hasil Pengujian dan Evaluasi

Pengujian dilakukan dengan membandingkan hasil ringkasan sistem dengan ringkasan acuan. Yang dimaksud dengan ringkasan acuan disini ringkasan yang diperoleh dari situs peringkas dokumen sebagai pembanding dalam menentukan tingkat presisi dari aplikasi yang dibangun. Hasil ringkasan sistem untuk dokumen Bahasa Inggris akan dibandingkan dengan hasil ringkasan dari situs http://www.tools4noobs.com/. Sedangkan hasil ringkasan sistem untuk Bahasa Indonesia dibandingkan hasil ringkasan manual berdasarkan situs http://www.bacain.com/.

Dalam pengujian ini, digunakan beberapa nilai inisialisasi awal yang berbeda untuk matriks W (*Non-negative Semantic Feature Matrix*) dan matriks H (*Non-negative Semantic Variable Matrix*). Penggunaan nilai inisialisasi yang berbeda ini bertujuan untuk melhat pengaruh dari nilai awal kedua matriks ini terhadap kalimat ringkasan yang dihasilkan. Nilai inisisalisasi yang dipilih berada dalam interval 0.1–0.25 karena interval tersebut merupakan interval bilangan acak terbaik yang menghasilkan nilai *Frobenius Norm* peling kecil dengan waktu eksekusi yang tersingkat [2].

Evaluasi terhadap hasil ringkasan sendiri dilakukan dengan melihat presisi dari ringkasan yang dihasilkan. Nilai presisi didefinisikan sebagai persentase kalimat ringkasan yang sesuai dengan ringkasan acuan yang digunakan. Berdasarkan pengujian terhadap 20 dokumen uji (10 dokumen berbahasa Inggris dan 10 dokumen berbahasa Indonesia) diperoleh hasil presisi rata-rata sebesar 51.87%.

Pengujian pada Tabel I memperlihatkan rata-rata presisi ketiga nilai inisialisasi yang digunakan lebih baik dibandingkan dengan menggunakan nilai *random*. Walaupun rata-rata presisi untuk nilai awal 0.1, 0.25, dan 0.175 ternyata sama (51.87%), namun distribusi presisi di setiap dokumen untuk ketiga nilai tersebut berbeda. Pengaruh lain dari nilai inisialisasi yang digunakan adalah pada rata-rata jumlah iterasi yang diperlukan untuk mencapai kondisi berhenti. Dari ketiga nilai yang diuji, inisialisasi dengan dengan nilai 0.175 memberikan rata-rata jumlah iterasi yang paling sedikit.



Gambar 1. Alur kerja sistem peringkasan dokumen

TABEL I

HASIL PENGUJIAN 20 DOKUMEN

No	Pengujian	Rata-rata Presisi	Rata-rata Iterasi
1	Artikel Bahasa Inggris dengan Inisialisasi Nilai Awal *random*	53.17	312
2	Artikel Bahasa Indonesia dengan Inisialisasi Nilai Awal *random*	30.99	402
3	Artikel Bahasa Inggris dengan Inisialisasi Nilai Awal 0.1	56.56	399
4	Pengujian Artikel Bahasa Indonesia dengan Inisialisasi Nilai Awal 0.1	47.18	422
5	Artikel Bahasa Inggris dengan Inisialisasi Nilai Awal 0.25	56.56	438
6	Artikel Bahasa Indonesia dengan Inisialisasi Nilai Awal 0.25	47.18	425
7	Artikel Bahasa Inggris dengan Inisialisasi Nilai Awal 0.175	56.56	384
8	Artikel Bahasa Indonesia dengan Inisialisasi Nilai Awal 0.175	47.18	344

Contoh dokumen Bahasa Indonesia dan ringkasan yang dihasilkan diberikan pada Gambar 2 dan Gambar 3. Dari contoh yang diberikan di Gambar 3, dapat dilihat bahwa ringkasan yang dihasilkan berupa kalimat-kalimat asli dari dokumen sumber. Kalimat yang dipilih sebagai ringkasan adalah kalimat yang memiliki bobot GRS tertinggi dan dianggap mewakili isi dari dokumen sumber. Walaupun isi kalimat hasil ringkasan cukup mewakili isi dokumen sumber namun belum terdapat kesinambungan antar kalimat. Hasil ringkasan akan lebih mudah dipahami jika dilakukan pemrosesan tambahan untuk menggabungkan kalimat-kalimat tersebut, antara lain penambahan atau pengurangan kata-kata sambung.

Apple membuat kejutan dengan meluncurkan iPad generasi keempat, yang diluncurkan bersamaan dengan iPad mini. Munculnya iPad 4 ini hanya berjarak tujuh bulan sejak iPad generasi ketiga atau The New iPad diluncurkan. Tentu saja kejutan Apple ini ditanggapi dengan kecewa oleh para pemilik iPad. Menurut studi yang dilakukan QuickSurveys Toluna, dari 2000 pemilik iPad yang diwawancarai, sekitar 45 persen pemilik iPad tidak senang dengan kehadiran iPad 4. Mereka mengungkapkan kemarahannya kepada Apple yang dinilai terlalu cepat meluncurkan generasi terbaru iPad. Tentu ini beralasan, mereka beranggapan kehadiran iPad 4 yang terlalu cepat menjadikan iPad 3 yang baru mereka beli menjadi produk usang. Ketidakpuasan peluncuran iPad 4 tersebut hampir merata pada semua pemilik iPad. Posisi puncak tentu ditempati pemilik iPad 3 yang tercatat sebanyak 50 persen, diikuti pemilik iPad 2 sebanyak 45 persen, dan pemilik iPad generasi pertama sebanyak 40 persen. Selain itu, satu dari empat responden meyakini reputasi Apple telah ternoda setelah menghapus Google Maps dari sistem operasi iOS 6 dan menggantinya dengan Apple Maps. Seperti diketahui, layanan Apple Maps menuai kecaman karena tidak akurat dan beberapa wilayah hilang dari pemetaan.

Gambar 2. Contoh Dokumen Bahasa Indonesia

Munculnya iPad 4 ini hanya berjarak tujuh bulan sejak iPad generasi ketiga atau The New iPad diluncurkan.

Menurut studi yang dilakukan QuickSurveys Toluna, dari 2000 pemilik iPad yang diwawancarai, sekitar 45 persen pemilik iPad tidak senang dengan kehadiran iPad 4.

Posisi puncak tentu ditempati pemilik iPad 3 yang tercatat sebanyak 50 persen, diikuti pemilik iPad 2 sebanyak 45 persen, dan pemilik iPad generasi pertama sebanyak 40 persen.

Gambar 3. Contoh Hasil Ringkasan Dokumen Bahasa Indonesia

V. KESIMPULAN

Beberapa kesimpulan yang diperoleh dalam penelitian ini antara lain:

1. Tahap *pre-processing* memiliki pengaruh yang sangat besar terhadap hasil proses peringkasan dokumen yang menggunakan metode NMF. Hasil tahap *pre-processing* akan menentukan isi dari matriks *term-by-sentence* yang akan digunakan dalam metode NMF. Kesalahan dalam proses *stemming* ataupun ketidak-tepatan dalam pemilihan *stopwords* akan berpengaruh terhadap bobot kemunculan kata (*term*) dan juga bobot GRS dari kalimat yang mengandung kata (*term*) tersebut.
2. Pemilihan nilai awal untuk elemen matriks *W* dan *H* sangat berpengaruh pada presisi kalimat ringkasan yang dihasilkan berdasarkan metode NMF, dan juga jumlah iterasi yang diperlukan untuk menghasilkan ringkasan.
3. Kalimat-kalimat hasil ringkasan dari perangkat lunak peringkas dokumen teks ini terkadang tidak berhubungan antar kalimatnya. Hal ini disebabkan pemilihan kalimat ringkasan hanya berdasarkan nilai GRS tertinggi saja tanpa memperhatikan keterhubungannya dengan kalimat sebelumnya atau sesudahnya. Untuk itu diperlukan pengembangan lebih lanjut sehingga hasil ringkasan dokumen dapat berupa kalimat yang memiliki susunan yang harmonis sehingga lebih nyaman untuk dibaca.

4. Diperlukan penanganan khusus untuk menangani *term* yang mengandung karakter numerik seperti tanggal, nominal uang, nomor telepon, dan lain-lain. Tidak adanya penanganan terhadap *term* numerik cenderung mengakibatkan jenis *term* ini memiliki nilai bobot yang kecil dan akhirnya tidak terpilih sebagai bagian dari ringkasan. Padahal pada kenyataannya, kalimat yang mengandung *term* numerik mungkin saja justru memiliki informasi yang penting.

REFERENSI

[1] A. Jelita., *Effective Techniques for Indonesian Text Retrieval*,. Melbourne: RMIT University. 2007.

[2] A. Yuliawati, *Implementasi Peringkas Otomatis pada Dokumen Terstruktur dengan Metode Faktorisasi Matriks Nonnegative*. Surabaya: ITS. 2011.

[3] D. D. Lee and H. S. Seung, *Algorithm for non-negative matrix factorization*. Advance in Neural Information Processing Systems, 13 , 556-562. 2001

[4] J. H. Lee, S. Park, C.-M. Ahn, and D. Kim, *Automatic generic document summarization based on non-negative matrix factorization*. Information Processing and Management, 45, 20-34. 2009.

[5] M. Porter. (2006). The Porter Stemming Algorithm. [Online]. Available : http://tartarus.org/~martin/PorterStemmer/.

Herastia Maharani menyelesaikan pendidikan sarjana di Departemen Teknik Informatika Institut Teknologi Bandung di tahun 2005, dan meraih gelar Magister Informatika di Institut Teknologi Bandung tahun 2010. Saat ini penulis bekerja sebagai staf pengajar di Institut Teknologi Harapan Bangsa, Bandung. Bidang ilmu dan area penelitian yang ditekuni penulis adalah *data mining* dan *information retrieval*. Bidang lain yang saat ini sedang dipelajari oleh penulis adalah *social network analysis*.

Monica Sanjaya, mahasiswa Teknik Informatika Institut Teknologi Harapan Bangsa yang lulus pada tahun 2013.