PENGGUNAAN METODE RELEVANCE MEASURE DAN LATENT SEMANTIC ANALYSIS (LSA) DALAM MEMBUAT IKHTISAR DOKUMEN BERITA

Agung Triwibowo[1], Adiwijawa[2], Moch Arif Bijaksana[3]

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

Abstrak

Pada zaman teknologi informasi saat ini, ketersediaan berita elektronik sangat dibutuhkan oleh masyarakat. Akan tetapi, berita yang terlalu panjang menyebabkan pembaca membutuhkan waktu yang lama dalam membaca teks berita elektronik. Untuk itu, kita memerlukan suatu sistem yang dapat membuat ikhtisar dari artikel-artikel atau dokumen-dokumen teks secara otomatis.
Tugas Akhir ini menerapkan metode relevance measure (TF - IDF) yang menggunakan konsep pembobotan dan LSA yang dapat mengidentifikasi kalimatkalimat yang penting secara semantik yang berguna untuk mengasilkan output sebuah ringkasan ekstrakif. Pada penerapan sistem ini, sebelumnya dokumen harus di preprosesing, kemudian diekstrak menjadi token dan dilanjutkan dengan proses stopword removal dalam menambahkan hasil keluaran ringkasan.
Pengujian dari sistem ini menggunakan evaluasi ROGUE. Pada hasil pengujian menunjukkan bahwa ringkasan menggunakan metode TF-IDF modifikasi lebih baik dibandingkan TF-IDF biasa. Ringkasan menggunakan metode LSA menghasilkan akurasi terbaik dibandingkan dengan metode TF-IDF dan modified TF-IDF.

Kata Kunci : LSA, peringkasan teks, preprocess, relevance measure(TF-IDF)

Abstract

In this information technology era recently, the availibility of electronic news is very needed by the poeple. But sometimes a very long writing can cost the reader a very long time too to read it. To handle this problem, we need a system that can create a summary from those articles or documents of text automatically
This thesis implements a relevance measure (TF-IDF) that use weight concept and Latent Semantic Analysis to identify important sentences semanticly that is useful to result output an extrative summarise. In an application of this system , first documents or articles must be preproccesed, then extracted to be token and continued by stopword removal process in advance result of summary output.
Testing of this system use ROUGE evalution. In testing result, it shows that modified TF-IDF better that usual TF-IDF. Then, LSA method also show best accuration if we compare with TF-IDF or modified TF-IDF method.

Keywords : LSA ,text summarization, preprocess, relevance measure(TF-IDF),

1. Pendahuluan

1.1 Latar Belakang

Dengan semakin berkembangnya ilmu pengetahuan dan teknologi, membuat jumlah informasi menjadi begitu banyak, salah satunya informasi berupa teks terutama pada media elektronik. Oleh karena itu, dibutuhkan informasi yang singkat dan padat yang merepresentasikan isi dokumen. Sehingga dikembangkan sebuah sistem yang mampu meringkas dokumen yaitu, peringkas teks otomatis (*automatic text summarization*).

Ikhtisar atau yang biasa disebut dengan ringkasan adalah suatu informasi penting dari suatu sumber atau sumber informasi ganda menurut kebutuhan-kebutuhan tertentu. Dengan ringkasan, kita dapat membuat keputusan-keputusan efektif dan mendapat informasi bermanfaat dalam waktu lebih singkat. Peringkasan sudah dimulai sejak tahun 1950. Edmundson menyajikan suatu survei dari metode-metode yang ada untuk ikhtisar yang otomatis *(automatic summarization)*.

Banyak metode maupun pendekatan yang digunakan untuk melakukan peringkasan teks otomatis. Terdapat dua buah pendekatan dilihat dari teknik pengambilan ringkasan, yaitu ekstraksi dan abstraksi. Di mana ekstraksi merupakan teknik menyeleksi materi dari sumber yang berupa teks sedangkan abstraksi merupakan teknik meringkas teks dengan cara mereformulasikan kembali versi aslinya.

Tugas Akhir ini menggunakan pendekatan ekstraksi dalam melakukan peringkasan. Ringkasan dari suatu teks diekstraksi dengan melakukan pemilihan dari bagian dokumen yang penting untuk menghasilkan hasil yang lebih singkat. Ringkasan manusia sering kali dilakukan dengan cara meng-*cut* dan *paste* dari suatu dokumen untuk menghasilkan ringkasan. Kita dapat belajar jenis operasi yang biasanya dilaksanakan secara manual yang dilakukan manusia untuk mengekstraksi dan meng-edit kalimat-kalimat kemudian mengembangkan program yang dapat bekerja secara otomatis *(automatic programs)* untuk menirukan operasi tersebut. Granularitas-granularitas dari ekstraksi berupa frase-frase (2 atau 3 kata-kata) seperti pada kalimat. Pendekatan ekstraksi mungkin punya permasalahan pada koherensinya. Salah satu pendekatan ektraksi yaitu dengan menggunakan metode pembobotan TF-IDF dan metode LSA.

Beberapa metode pembobotan kata secara umum yang biasa digunakan untuk teks yaitu *term frequency* (TF) dan *inverse document frequency* (IDF). Pembobotan TF IDF*(relevance measure)* memperhitungkan kemunculan term tidak hanya pada kalimat yang memiliki term tersebut, tetapi menimbang kemunculan term dalam keseluruhan dokumen.

Latent Semantic Analysis (LSA) yang diilhami dari pengindeks laten semantik dan menerapkan *Singular Value Decomposition* (SVD) ke dokumen sentence matrix [7]. *Singular Value Decomposition* (SVD) adalah suatu alat mathematical sangat tangguh untuk menemukan dimensi-dimensi pokok ortogonal data multidimensional. SVD mempunyai aplikasi-aplikasi di dalam banyak bidang dan dikenal oleh nama-nama yang berbeda, *Latent Semantic Analysis* (LSA) di dalam pengolahan teks. SVD dalam pengolahan teks diberi nama LSA karena SVD

berlaku untuk *document-word matrices*, kelompok-kelompok dokumen yang bersifat secara semantis berhubungan dengan satu sama lain.

Tugas Akhir ini dilakukan dua metode text summarization untuk membuat *generic text summaries* dengan peringkat (*ranking*) dan mengekstraksi kalimat-kalimat dari dokumen asli. Metode yang pertama menggunakan metode *relevance measure* untuk me-rank keterkaitan kalimat dan metode yang kedua gunakan *latent semantic analysis* yang dapat mengidentifikasi kalimat-kalimat penting secara semantis, untuk membuat ikhtisar.

1.2 Perumusan Masalah

Dalam pengerjaan Tugas Akhir ini, ada beberapa rumusan masalah yang coba untuk diselesaikan, yaitu sebagai berikut :

a) Bagaimana mengimplementasikan ringkasan teks yang *generic* menggunakan *relevance measure* and *Latent Semantic Analysis*?
b) Bagaimana tingkat akurasi hasil ringkasan tersebut?

Dengan mempertimbangkan kompleksitas yang mungkin ada pada suatu ikhtisar, maka pada Tugas Akhir ini terdapat beberapa batasan masalah, yaitu :

a) Membuat ikhtisar dengan menggunakan metode *Relevance Measure* dan metode Latent Semantic Analysis (untuk mengidentifikasi *important sentences*)
b) Menggunakan dokumen tunggal teks berita berbahasa Indonesia.
c) Peringkasan teks dilakukan secara *offline*.
d) Ringkasan berupa hasil ekstraksi kalimat-kalimat dari dokumen asli.
e) Tidak melakukan *stemming* terhadap teks masukan.
f) Hasil ringkasan yang diperoleh dibandingkan dengan hasil ringkasan manual.

1.3 Tujuan

Tujuan yang ingin dicapai dari Tugas Akhir ini yaitu :

a) Menganalisis dan mengimplementasikan Relevance Measure dan Latent Semantic Analysis untuk membuat ringkasan otomatis.
b) Menganalisis dan membandingkan performansi hasil ringkasan berdasarkan dua metode tersebut berupa akurasi berdasarkan *precision, recall,* dan *f-measure* dengan menggunakan Rouge.

1.4 Metodologi Penyelesaian Masalah

Dalam menyelesaikan Tugas Akhir ini, ada beberapa tahapan metode akan dilakukan, yaitu :

a) Studi Literatur
   Studi literatur bertujuan untuk meningkatkan pemahaman terhadap metode-metode yang akan digunakan dalam Tugas Akhir, yaitu metode Relevance measure dan LSA, dengan cara mencari referensi yang berkaitan dengan metode-metode tersebut dan kemudian mendalami materinya.
b) Analisis dan Perancangan Perangkat Lunak

Pada tahap ini akan dilakukan analisis terhadap perangkat lunak yang diperlukan dalam membangun implementasi dari *Summarization*, baik berupa bahasa pemrograman yang akan digunakan sampai pada algoritma dan struktur data yang digunakan.

c) Implementasi
Hasil perancangan pada tahap sebelumnya akan menjadi dasar dalam tahap implementasi ini. Pada tahap ini akan dilakukan pengkodean berdasarkan hasil rancangan di atas.

d) Testing
Di tahap ini, akan dilakukan pengujian terhadap hasil implementasi guna menemukan dan menghilangkan *error/bug* yang mungkin masih ada.

e) Analisis Hasil Implemtasi
Analisis hasil implementasi dilakukan dengan mengukur akurasi atau ketepatan kerja dari program dalam mengikhtisarkan suatu dokumen

f) Pembuatan Laporan Tugas Akhir.
Pada tahap akhir, dilakukan pembuatan dokumentasi yang berupa laporan Tugas Akhir.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Berdasarkan analisis terhadap hasil pengujian, Tugas Akhir ini memiliki kesimpulan sebagai berikut:

1. Ringkasan otomatis menggunakan metode *Modified* TF-IDF menunjukkan peningkatan akurasi lebih baik daripada metode TF-IDF. Peningkatan akurasi tersebut rata-rata sebesar 0,4%.
2. Ringkasan menggunakan metode LSA menghasilkan akurasi yang lebih baik di bandingkan dengan metode TF-IDF dan *Modified* TF-IDF dengan menggunakan evaluasi baik Rouge n-1, Rouge n-2, maupun Rouge w
3. Akurasi hasil peringkasan juga dipengaruhi oleh panjang ringkasan; hasil pengujian secara umum menunjukkan akurasi ringkasan 25% lebih baik daripada akurasi ringkasan 10%, dan ringkasan 50% yang paling baik akurasinya.

5.2 Saran

1. Proses pengambilan data dapat dikembangkan secara *online*, sehingga proses pengambilan data dapat dilakukan secara otomatis.
2. Menambah atau memperbaiki kata-kata pada daftar *stopwords* sehingga peran eliminasi *stopwords* dapat memberikan akurasi yang lebih baik.

Referensi

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

[2] E. Hovy and C. Lin, “Automated text summarization in summarist," *in Proceedings of the TIPSTER Workshop*, (Baltimore, MD), 1998. Available on : http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.130.38 (19 Januari 2009).

[3] Golub, Gene H. and Van Loan, Charles F., 1996, Matrix Computations, Baltimore and London, John Hopkins University Press

[4] I Wayan S.W., ″Membandingkan Pendekatan Latent Semantic terhadap WordNet untuk Semantic Similarity," Universitas Gunadarma, 2006. Available on : http://iwayan.staff.gunadarma.ac.id/Publications/files/708/2006_Kommit_LatentSemantic_IWS.pdf (19 januari 2009).

[5] J. Goldstain, M. kantrowitz, V. Mittal, and J. Carbonell, “Summarizing text documents: Sentence selection and evaluation metrics," *in Proceedings of ACM SIGIR'99, (Berkeley, CA)*, Aug. 1999. Available on : http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.11.9344 (19 Januari 2009).

[6] Kontostathis, A., and Pottenger, Wiliam,M., “A Framework for Understanding LSI Performance,” Lehigh University. Available on : http://citeseerx.ksu.edu.sa/viewdoc/summary?doi=10.1.1.4.87 (20 Juli 2009).

[7] Landauer, T. K., Foltz, P. W., and Laham, D. “Introduction to Latent Semantic Analysis.” *Discourse Processes*,1998, 25, 259-284. Available on : http://lsa.colorado.edu/papers/dp1.LSAintro.pdf (19 Januari 2009).

[8] M. Berry, S. Dumais, and G. O'Brien, “Using linear algebra for intelligent information retrieval," *Tech. Rep. UT-CS-94-270*, University of Tennessess, Computer Science Department, Dec. 1994. Available on : http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.117.3373 (19 Januari 2009).

[9] Regina Barzilay and Michael Elhadad. 1997. “Using lexical chains for text summarization.” *In Proceedings of the ACL’97/EACL’97 Workshop on Intelligent Scalable Text Summarization*, pages 10–17, Madrid, Spain. Available on : http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.29.6428 (19 Januari 2009).

[10] Rosario, Barbara. “Latent Semantic Indexing: An overview”. Infosys 240 spring 2000,Final Paper,2000.

[11] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, pp. 391-407, 1990. Available on : http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.108.8490 (19 Januari 2009).

[12] Y. Gong and X. Liu, "Generic text summarization using relevance measure
and latent semantic analysis," *in Proc. ACM SIGIR*, 2001, pp. 19–25. Available on: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.109.5097 (19 Januari 2009).

[13] ___. "Automatic summarization," *the free encyclopedia,* Available on : http://en.wikipedia.org/wiki/Automatic_summarization (30 Mei 2009).

[14] ___. "Information Retrieval," *the free encyclopedia,* Available on : http://en.wikipedia.org/wiki/Information_retrieval (30 Mei 2009).

[15] ___. "Latent semantic analysis," *the free encyclopedia*, Available on : http://en.wikipedia.org/wiki/Latent_semantic_analysis (19 Januari 2009).