

Clustering Berita Berbahasa Indonesia

Yudi Wibisono, Masayu Leylia Khodra

FPMIPA Universitas Pendidikan Indonesia, Jln Setiabudi 229 Bandung
yudi@upi.edu

KK Informatika Sekolah Teknik Elektro dan Informatika I T B, jln. Ganesa 10 Bandung
masayu@informatika.org

Abstrak

Volume berita elektronik berbahasa Indonesia yang semakin besar merupakan sumber informasi yang berharga. *Clustering* dokumen teks adalah salah satu operasi pada text mining untuk mengelompokkan dokumen yang memiliki kesamaan isi. *Clustering* dapat diaplikasikan untuk menemukan keterkaitan antar berita. Eksperimen dalam paper ini menggunakan 4718 berita dari situs www.kompas.com yang diambil pada bulan Juni 2005 sampai dengan November 2005. Hasil dari eksperimen menunjukkan *clustering* dapat mengungkapkan keterkaitan antar berita yang sebelumnya tidak terlihat. *Clustering* pada eksperimen ini menggunakan algoritma K-means. Penggunaan *stemming*, pembobotan term log-tf dan log-tf.idf juga diujicobakan untuk melihat pengaruhnya terhadap kualitas *cluster* yang diukur dengan *purity*.

Kata kunci: *clustering, K-Means, berita berbahasa Indonesia*

1. Pendahuluan

Jumlah sumber berita berbahasa Indonesia yang tersedia di internet semakin besar. Menurut situs <http://dmoz.org/World/Indonesia/Berita/> (Agustus 2005) terdapat 128 situs berita berbahasa Indonesia. Karakter situs berita yang sering diperbaharui akan menimbulkan aliran informasi berita dalam jumlah besar setiap harinya. *Clustering* dapat digunakan untuk membantu menganalisis berita dengan mengelompokkan secara otomatis berita yang memiliki kesamaan.

Clustering dokumen adalah aktifitas klasifikasi otomatis dokumen tanpa supervisi ke dalam *clusters/grup-grup* (1). Masalah mendasar dalam *clustering* dokumen adalah mengelompokkan dokumen yang mirip satu sama lain.

Walaupun memiliki potensi yang besar, penelitian mengenai *clustering* terhadap *corpus* berbahasa Indonesia masih belum (atau jarang) dilakukan.

Makalah ini berisi hasil eksperimen awal mengenai *clustering* untuk corpus berita berbahasa Indonesia. Diharapkan eksperimen ini dapat menjadi landasan untuk penelitian berikutnya. Oleh karena itu dipilih teknik preprocessing dan algoritma *clustering* yang sesederhana mungkin (K-means).

Pada makalah ini akan dibahas efek penggunaan pembobotan log-tf dibandingkan log-tf.idf dan penggunaan *stemming* pada tahap *preprocessing*.

2. Representasi Dokumen

Koleksi berita berbahasa Indonesia yang digunakan bersumber dari www.kompas.com. Koleksi ini terdiri atas 8237 dokumen yang diambil dari bulan Juli 2005 sampai dengan November 2005.

Berdasarkan URL-nya, Kompas telah memberikan label kategori pada sebagian berita (4718 berita). Kategori yang diberikan Kompas adalah: Metro, Otomotif, Kesehatan, Olahraga, Teknologi dan Gaya Hidup.

Eksperimen dalam paper ini menggunakan berita yang sudah dikategorikan secara manual oleh Kompas.

Setiap dokumen dalam koleksi berita direpresentasikan sebagai sebuah vektor terms. Vektor term untuk suatu dokumen d adalah tuple bobot semua term pada d .

Ada dua metode pembobotan term yang digunakan yaitu log-tf (*logarithmic term frequency*) dan log-tf.idf (*logarithmic term frequency – inverse document frequency*). Kedua metode pembobotan ini didefinisikan sebagai berikut:

$$\text{Logtf}(d,t) = \log(1 + \text{rawtf}(d,t)) \quad (2.1)$$

$$\text{Logtf.idf}(d,t) = \text{logtf}(d,t) * \log\left(\frac{|D|}{n}\right) \quad (2.2)$$

$\text{rawtf}(d,t)$ adalah frekuensi kemunculan term t pada dokumen d ; $|D|$ adalah jumlah semua dokumen pada koleksi; n adalah jumlah dokumen yang mengandung term t .

Metode pembobotan log-tf.idf digunakan karena metode pembobotan ini paling baik dalam information retrieval.

Nilai bobot suatu term menyatakan kepentingan bobot tersebut dalam merepresentasikan dokumen. Pada pembobotan log-tf.idf, bobot akan semakin besar jika frekuensi kemunculan term semakin tinggi, tetapi bobot akan berkurang jika term tersebut semakin sering muncul pada dokumen lainnya.

2. Preprocessing

Dalam preprosesing, langkah-langkah yang dilakukan adalah *case folding*, *parsing*, pembuangan *stopwords*, *stemming*, dan penghitungan bobot setiap term.

Case folding adalah mengubah semua huruf dalam dokumen menjadi huruf

kecil. Hanya huruf 'a' sampai dengan 'z' yang diterima. Karakter selain huruf dihilangkan dan dianggap delimiter.

Stopwords adalah kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan *bag-of-words*. Contoh *stopwords* adalah "yang", "dan", "di", "dari" dan seterusnya. Pada eksperimen ini digunakan 329 *stopword* (7).

Stemming adalah proses pengembalian berbagai bentukan kata ke dalam suatu representasi yang sama. *Stemming* yang digunakan adalah Porter Stemmer yang disesuaikan dengan aturan bahasa Indonesia (2).

4. Algoritma K-Means

Eksperimen ini menggunakan algoritma yang paling umum digunakan dalam *clustering* yaitu algoritma K-Means. Algoritma ini populer karena mudah diimplementasikan dan kompleksitas waktunya linear. Kelemahannya adalah algoritma ini sensitif terhadap inisialisasi *cluster* (1)

Dasar algoritmanya adalah sebagai berikut:

- 1) Inisialisasi *cluster*
- 2) Masukkan setiap dokumen ke *cluster* yang paling cocok berdasarkan ukuran kedekatan dengan centroid. Centroid adalah vektor term yang dianggap sebagai titik tengah *cluster*.
- 3) Setelah semua dokumen masuk ke *cluster*. Hitung ulang centroid *cluster* berdasarkan dokumen yang berada di dalam *cluster* tersebut.
- 4) Jika centroid tidak berubah (dengan treshold tertentu) maka stop. Jika tidak, kembali ke langkah 2.

Ukuran kedekatan antara dua vektor term t_1, t_2 yang digunakan pada paper ini adalah cosinus sudut antara kedua vektor tersebut.

$$\cos(\langle t_1, t_2 \rangle) = \frac{t_1 \cdot t_2}{|t_1| \cdot |t_2|} \quad 4.1$$

5. Ukuran Evaluasi

Ide utama ukuran evaluasi dalam eksperimen ini adalah membandingkan

hasil *clustering* dengan label manual dari setiap berita.

Purity (3) digunakan sebagai ukuran evaluasi dalam eksperimen ini. *Purity* menggunakan ukuran *precision* yang digunakan pada information retrieval. Setiap *cluster* P yang dihasilkan dianggap seperti hasil dari sebuah *query*. Sedangkan setiap kelas L yang telah dilabeli secara manual dianggap sebagai hasil yang diinginkan dari suatu *query*. *Precision* suatu *cluster* P terhadap kelas label L didefinisikan sebagai:

$$Precision(P, L) = \frac{|P \cap L|}{|P|} \quad (5.1)$$

Nilai *purity* secara keseluruhan didefinisikan sebagai:

$$Purity = \sum_{P \in \mathcal{P}} \frac{|P|}{|D|} \max_{L \in \mathcal{L}} Precision(P, L) \quad (5.2)$$

Purity adalah ukuran “kemurnian” sebuah *cluster*. Semakin tinggi nilainya (maksimum 1), semakin “murni” *cluster* sebagai subset dari kategori yang diwakilinya

6. Eksperimen dan Evaluasi Hasil

Eksperimen dilakukan dengan empat perlakuan:

- 1) *stemming* dengan log-tf
- 2) *stemming* dengan log-tf.idf
- 3) tanpa *stemming* dengan log-tf
- 4) tanpa *stemming* dengan log-tf.idf

Jumlah *cluster* yang dipilih (k) adalah 20. Inisialisasi *cluster* dilakukan dengan cara memilih dokumen secara random. Eksekusi untuk setiap perlakuan dilakukan sebanyak lima kali.

Tabel 1 berikut memperlihatkan rata-rata *purity* untuk setiap perlakuan. Dari hasil eksekusi dapat dilihat bahwa kualitas *cluster* terbaik diperoleh dengan tidak menggunakan *stemming* dan menggunakan metode pembobotan log-tf.idf. Walaupun demikian tidak ada perbedaan yang signifikan diantara keempat perlakuan.

| Perlakuan | Rata2 Purity | std |
|-------------------------|--------------|-------|
| Stemming, log-tf | 0.559 | 0.029 |
| Stemming, log-tf.idf | 0.500 | 0.074 |
| No Stemming, Log-tf | 0.533 | 0.123 |
| No Stemming, Log-tf.idf | 0.568 | 0.044 |

Tabel 1: Hasil eksperimen dengan k=20 untuk empat perlakuan.

Hal yang tidak diduga sebelumnya adalah penggunaan *stemming* ternyata tidak meningkatkan kualitas *cluster*, kecuali untuk pembobotan dengan log-tf. Hal ini sejalan dengan temuan (4) tentang penggunaan *stemming* pada *information retrieval*. Tetapi perlu diingat bahwa penggunaan *stemming* mengurangi term yang harus diproses dari 61.916 term menjadi 47.362 term. Pengurangan ini akan mempercepat waktu proses dan menghemat media penyimpanan.

Pada pengamatan judul-judul dokumen di dalam *cluster* secara visual, untuk k=20 tidak memperlihatkan hasil yang menarik. Untuk k=100, terlihat beberapa dokumen yang kategorinya berbeda tetapi masuk ke dalam *cluster* yang sama. Contohnya ada berita berkategori otomotif “Toyota Fine-s, Mobil Sport Fuel Cell Hybrid” masuk ke dalam *cluster* yang sama dengan berita “Katalis Rhenium, Mengekstrak Hidrogen dari Air” yang berkategori teknologi.

Pengamatan secara visual pada k=100 juga memperlihatkan terbentuknya beberapa sub kategori. Misalnya untuk kategori “Olahraga” muncul *cluster* untuk cabang olahraga bulutangkis, tinju dan sepakbola. Sedangkan untuk “Teknologi” muncul subkategori: ruang-angkasa dan hewan. Untuk kategori “metro” ada *cluster* yang berisi tentang kriminalitas dan BBM (Bahan Bakar Minyak).

7. Kesimpulan

Clustering pada berita berbahasa Indonesia memiliki potensi yang besar untuk dikembangkan lebih lanjut. *Clustering* berita dapat digunakan untuk mencari berita yang terkait, mencari kategori baru dan memperlihatkan keterkaitan antara berita. Manfaatnya akan lebih besar jika digunakan lebih dari satu sumber berita, misalnya dari koran Kompas dan koran Republika.

Penggunaan log-tf.idf tanpa *stemming* menghasilkan kualitas *cluster* terbaik di dalam eksperimen ini. Tetapi kualitas *cluster* yang ada pada makalah ini masih rendah dan perlu dilakukan penelitian lanjutan untuk meningkatkan kualitasnya.

8. Daftar Pustaka

1. A.K Jain, M.N. Murty and P.J Flynn, *Data Clustering: A Review*, in ACM Computing Surveys, Vol. 32, No. 3 September 1999.
2. J. Asian, H.E.Williams, S.M.M. Tahaghoghi, *Stemming Indonesian*, 28th Australasian Computer Science Conferences in Research and Practice in Information Technology, Vol.38
3. A. Hotho, S. Staab, G. Stumme, *Wordnet improves Text Document Clustering*
4. F.Z.Tala, *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia*, master thesis at Institute for Logic, Language and Computation, Universiteit van Amsterdam
5. A. Hotho, S. Staab, G. Stumme, *Explaining Text Clustering Results using Semantic Structures*.
6. P.S.Bradley, U.M.Fayyad, *Refining Initial Points for K-Means Clustering*
7. Yudi Wibisono, "Klasifikasi Berita Berbahasa Indonesia menggunakan Naïve Bayes Classifier"