Pengelompokan Artikel Berita Berbahasa Indonesia dengan *Agglomerative Clustering*

Masayu Leylia Khodra¹⁾, Yudi Wibisono²⁾

¹KK Informatika Sekolah Teknik Elektro dan Informatika ITB, Ganesha 10 Bandung

²Ilmu Komputer Universitas Pendidikan Indonesia, Setiabudi 229 Bandung

e-mail: yudi@upi.edu

Abstrak

Makalah ini mengaplikasikan agglomerative clustering untuk pengelompokan artikel berita berbahasa Indonesia untuk sistem aggregator berita. Agglomerative clustering merupakan teknik clustering hirarki dengan keunggulan jumlah cluster tidak perlu ditentukan, dan kualitas cluster tidak bergantung pada inisialisasi awal anggota cluster. Empat linkage diimplementasikan yaitu single linkage, complete linkage, average linkage, dan average-group linkage. Clustering dilakukan dengan menggunakan fitur leksikal, pembobotan term-frequency inverse document-frequency (tf.idf), cosine similarity, dan minimum anggota cluster adalah tiga. Dengan menggunakan 104 artikel berbahasa Indonesia yang telah dilabeli, kualitas cluster terbaik dihasilkan agglomerative clustering dengan menggunakan complete linkage dan kemiripan minimum 0.3 (purity rata-rata 0.888 dan lima cluster) dan 0.4 (purity rata-rata 0.938 dan empat cluster). Hasil eksperimen juga menunjukkan bahwa complete linkage menghasilkan purity rata-rata terbaik dan konsisten dibandingkan jenis linkage lainnya, dan nilai purity akan semakin tinggi jika parameter min_sim diperbesar, tetapi hal tersebut menyebabkan jumlah cluster yang dihasilkan semakin kecil.

Kata kunci: *agglomerative clustering, berita bahasa Indonesia, aggregator berita, tf.idf, purity*

1. Pendahuluan

Sistem agregator berita membantu pembaca berita dengan mengumpulkan informasi berita dari berbagai sumber dan menyajikannya dalam satu tempat. Aplikasi agregator berita dapat mengumpulkan berita terbaru dalam setiap waktu melalui RSS feed. Dengan sistem agregator berita, pembaca tidak perlu lagi mencari berita. Sistem agregator berita berbahasa Indonesia sudah cukup banyak saat ini, mulai dari aplikasi babe (baca berita), kurio, scoop, sampai dengan Google News edisi Indonesia. Sistem agregator berita yang dikembangkan dalam penelitian ini dapat dilihat lebih detil pada [1], termasuk perbedaannya dengan sistem yang ada.

Dalam sistem agregator berita, pengelompokan berita berbahasa Indonesia berperan penting. Setiap kelompok berita menyatakan satu topik berita yang anggotanya merupakan artikel-artikel berita dari berbagai portal berita. Kualitas kelompok berita sangat penting karena dapat membantu pembaca untuk memilih topik berita yang diinginkan sebelum membaca artikel-artikel yang relevan.

Pengelompokan berita berbahasa Indonesia telah banyak dilakukan oleh para peneliti [2-5] dengan berbagai teknik dan tujuan. K-Means merupakan teknik yang paling sederhana dan umum digunakan untuk berita bahasa Indonesia [2-5] karena K-Means mudah diimplementasikan [2,6] dan kompleksitas waktunya $O(NKd)$ dengan N jumlah data, K jumlah cluster, dan d dimensi data [9]. Selain K-Means, teknik lainnya yang pernah diaplikasikan untuk kumpulan berita bahasa Indonesia antara lain fuzzy C-Means [5], Self Organizing Map (SOM) [7], single-pass clustering [8], dan suffix-tree clustering [9]. Beberapa penelitian bertujuan untuk melihat pengaruh preproses [2], seleksi fitur [4], pembobotan sampel [5] terhadap kualitas cluster yang dihasilkan. Ambarwati & Winarko (2014) mencari kategori kecenderungan berita tahun 2007-2009 dengan Self Organizing Map (SOM) [7], tetapi penelitian [7] tidak

melakukan pengujian kualitas *cluster* yang dihasilkan. Google News menggunakan *agglomerative clustering* untuk pengelompokan berita, termasuk untuk edisi Indonesia.

Pada makalah ini, pengelompokan berita dilakukan dengan menggunakan *agglomerative clustering*. *Agglomerative clustering* merupakan teknik *clustering* hirarki yang memiliki representasi *cluster* berupa set anggota, bukan *centroid* seperti pada K-Means. Keunggulan dari *agglomerative clustering* adalah jumlah *cluster* tidak perlu ditentukan, dan kualitas *cluster* tidak bergantung pada inisialisasi awal anggota *cluster*. Walaupun *agglomerative clustering* memiliki banyak keunggulan dan telah terbukti aplikasinya pada Google News, belum ada penelitian sebelumnya yang menginvestigasi teknik ini untuk berita bahasa Indonesia. Oleh karena itu, penelitian ini mengimplementasikan *agglomerative clustering* dan melakukan pengelompokan berita dengan *agglomerative clustering*.

Pada bagian selanjutnya, akan dibahas metode *agglomerative clustering*. Kemudian diberikan hasil dan pembahasannya. Pada bagian terakhir, akan dibahas simpulan dan penelitian selanjutnya yang akan dilakukan.

2. Metode *Agglomerative clustering*

Agglomerative clustering merupakan teknik *clustering* hirarki yang memiliki representasi *cluster* berupa set anggota. Proses *clustering* N data dimulai dengan N *cluster* beranggota 1 data. Seperti yang diilustrasikan pada Gambar 1, untuk 7 data A..G, terdapat 7 *cluster* awal. Secara umum, kondisi berhentinya adalah setelah terbentuk 1 *cluster* besar yang berisi semua data. Proses *clustering* diperlihatkan dari bawah ke atas pada Gambar 1.

Gambar 1. Ilustrasi *agglomerative clustering*

Walaupun struktur *cluster* yang dihasilkan berupa hirarki, tetapi penelitian ini belum menggunakan struktur hirarki tersebut. Fokus penelitian ini masih pada kualitas *cluster* terbaik yang bisa dihasilkan. Oleh karena itu, *agglomerative clustering* memiliki parameter *min_sim* yang menyatakan batas ambang kemiripan minimum. Dengan adanya parameter *min_sim*, proses *clustering* N data yang dimulai dengan N *cluster* beranggota 1 data, tidak selalu berakhir dengan 1 *cluster* besar. Proses *clustering* berupa penggabungan *cluster* akan berhenti jika kemiripan maksimum antar *cluster* pada iterasi tersebut lebih kecil dari *min_sim*. Semakin tinggi nilai *min_sim*, jumlah *cluster* yang dihasilkan akan semakin sedikit. Kompleksitas algoritma *agglomerative clustering* adalah $O(N^2)$ [11].

Langkah-langkah pada *agglomerative clustering* adalah sebagai berikut:

1. Setiap data diinisialisasi sebagai satu *cluster* (*singleton cluster*).
2. Hitung linkage antar setiap *cluster*, dan hasilnya disimpan dalam matriks linkage. Pada makalah ini, linkage diimplementasikan dengan kemiripan antar *cluster*.
3. Cari dua *cluster* C_i dan C_j yang kemiripannya paling tinggi (kemiripan maksimum) tetapi bukan merupakan *cluster* yang sama. Seperti yang diilustrasikan pada Gambar 1, pada iterasi pertama, B dan C merupakan pasangan *cluster* dengan kemiripan maksimum. Jika kemiripan maksimum lebih kecil daripada parameter *min_sim*, kondisi berhenti terpenuhi dan langsung ke langkah 6. Jika kondisi berhenti belum terpenuhi, lanjutkan ke langkah berikutnya.
4. Gabung C_i dan C_j menjadi satu *cluster* baru, lalu update matriks linkage. Misalkan pada iterasi pertama, B dan C digabung menjadi satu *cluster* baru, kemudian update nilai linkage cluster B-C dengan A, D, E, F, dan G.
5. Jika jumlah *cluster* masih lebih dari satu dan kondisi berhenti belum dipenuhi, kembali ke langkah 3.
6. Proses selesai.

Perhitungan jarak atau kemiripan antar *cluster* disebut dengan linkage. Terdapat empat linkage yang diimplementasikan dalam penelitian ini yaitu *single linkage*, *complete linkage*, *average linkage*, dan *average-group linkage* [10]. *Single linkage* diimplementasikan sebagai kemiripan maksimum antar anggota dua *cluster*, *complete linkage* merupakan kemiripan minimum antar anggota dua *cluster*, *average linkage* merupakan kemiripan rata-rata antar semua pasangan anggota dua *cluster*, dan *average-group linkage* diimplementasikan sebagai kemiripan *centroid* kedua *cluster*.

Eksperimen bertujuan untuk mengevaluasi kualitas *cluster* yang dihasilkan oleh *Agglomerative clustering*. Dataset yang digunakan adalah 104 artikel berita berbahasa Indonesia pada 11 April 2016 yang telah dilabeli menjadi 20 *cluster*. Semua artikel ini telah dikategorikan secara otomatis ke dalam berita Pendidikan [1].

Karena data yang akan diproses berupa judul dan teks artikel berita, teks ditransformasi menjadi vektor fitur leksikal dengan pembobotan term-frequency inverse document-frequency (tf.idf) [2]. Anggota *cluster* berupa vektor fitur. Kemiripan antar anggota *cluster* dihitung dengan cosine similarity [2]. Sebagai tambahan batasan spesifik untuk *clustering* berita, ditambahkan minimum anggota dari suatu *cluster* adalah 3 karena *cluster* dengan jumlah artikel hanya 1 atau 2 dapat diabaikan.

Evaluasi kualitas *cluster* dilakukan dengan menggunakan *purity* rata-rata. *Purity* adalah ukuran “kemurnian” sebuah *cluster*. Semakin tinggi nilainya (maksimum 1), semakin “murni” *cluster* sebagai subset dari kategori yang diwakilinya. *Purity* suatu *cluster* dihitung dengan mencari jumlah anggota maksimum yang overlap dengan salah satu label dibagi dengan jumlah anggota *cluster* tersebut [2].

3. Hasil dan Pembahasan

Eksperimen yang dilakukan mengobservasi 2 faktor yaitu jenis linkage (*single*, *complete*, *average*, *average-group*), dan nilai *min_sim* (0.1-1.0). Jumlah *cluster* yang dihasilkan dengan anggota lebih dari minimum anggota (yaitu 3) dicantumkan dalam kurung pada setiap nilai *purity* rata-rata. Karena jumlah *cluster* yang dilabeli sebanyak 20 *cluster*, kombinasi linkage dan *min_sim* tidak hanya ditentukan oleh *purity*, tetapi juga oleh jumlah *cluster* yang dihasilkan.

Tabel 1. *Purity* rata-rata dari kombinasi linkage dan *min_sim*

Linkage	Min_sim						
	0.1	0.2	0.3	0.4	0.5	0.6	≥0.7
Single	0.214 (1)	0.588 (1)	0.630 (2)	0.765 (3)	0.889 (3)	- (0)	- (0)
Complete	0.772 (7)	0.823 (8)	0.880 (5)	0.938 (4)	- (0)	- (0)	- (0)
Average	0.231 (1)	0.747 (3)	0.750 (3)	0.833 (4)	1 (2)	- (0)	- (0)
Average-group	0.212 (1)	0.226 (1)	0.206 (1)	0.250 (1)	0.609 (2)	1 (1)	- (0)

Pada Tabel 1, dapat dilihat bahwa *complete linkage* menghasilkan *purity* rata-rata terbaik dan konsisten dibandingkan jenis linkage lainnya. Selain itu, diperlihatkan juga bahwa nilai *purity* akan semakin tinggi jika *min_sim* diperbesar, tetapi jumlah *cluster* yang dihasilkan semakin kecil. Jumlah *cluster* = 0 artinya tidak ada *cluster* yang berhasil memiliki anggota ≥ 3.

Dalam penelitian ini, kualitas *cluster* terbaik dihasilkan *agglomerative clustering* dengan menggunakan *complete linkage* dan kemiripan minimum 0.3 (5 *cluster*) dan 0.4 (4 *cluster*). *Purity* rata-rata = 1 dari *average linkage* dan *average-group linkage* tidak dipilih karena jumlah *clusternya* hanya 2 dari 20 label yang ada.

Tabel 2. Judul artikel dari hasil *clustering* dengan *complete linkage* serta *min_sim* 0.3 dan 0.4

min_sim=0.3	Min_sim=0.4
<i>Cluster</i> 1 (<i>purity</i> =1): 1. Robot Karya Mahasiswa Semarang Sabet Juara di Amerika 2. Juara Umum di AS, Tim Robot Unissula Diarak Bak Pahlawan	<i>Cluster</i> 1 (<i>purity</i> =1): 1. Robot Karya Mahasiswa Semarang Sabet Juara di Amerika 2. Juara Umum di AS, Tim Robot Unissula Diarak Bak Pahlawan

<p>3. Juara di AS, Tim Robotik Unissula Diganjar Hadiah ke Mekah</p> <p><i>Cluster 2 (purity=1):</i></p> <ol style="list-style-type: none"> 1. Zaskia Gotik Pimpin Artis Baca Pancasila di DPR 2. Alasan PKB Tunjuk Zaskia Gotik Jadi Duta Pancasila 3. Zaskia Gotik Jadi 'Duta Pancasila,' Kok Bisa? 4. Zaskia Gotik Jadi Duta Pancasila, Denny Cagur Sambut Baik 5. Ini Alasan PKB MPR Mengangkat Eneng Jadi Duta Pancasila 6. Eks Artis Cilik Sindir Zaskia Gotik Jadi Duta Pancasila <p><i>Cluster 3 (purity=0.4):</i></p> <ol style="list-style-type: none"> 1. Native Speaker Permudah Pemahaman Bahasa Inggris 2. Siswa SMA Unjuk Kebolehan Nginggris di Ajang Ini 3. Ketahui Manfaat Kebiasaan Membawa Bekal 4. Lima Kebiasaan Mahasiswa yang Hambat Skripsi 5. Belajar Kelola Tinja, Pemkot Bekasi Harus Jauh-Jauh ke Filipina 6. Pemkot Bekasi Studi Banding Tinja ke Filipina 7. Skripsi Penting untuk Akademis Mahasiswa 8. Marak Joki Skripsi, Ini Tanggapan Menristekdikti 9. Beasiswa 5.000 Doktor Kemenang, Mau? 10. Ini Alasan Tokopedia Dilirik 24 Ribu Pelamar per Bulan <p><i>Cluster 4 (purity=1):</i></p> <ol style="list-style-type: none"> 1. Lima Keuntungan Ikut Try Out Masuk PTN 2. Demi Lulus SBMPTN, Rela Try Out 10 Kali! 3. Ikuti Try Out Online SBMPTN 2016 Okezone! 4. Menabung Demi Ikut Bimbel dan Try Out <p><i>Cluster 5 (purity=1):</i></p> <ol style="list-style-type: none"> 1. Kolaborasi Kampus Indonesia-Eropa untuk Modernisasi Perguruan Tinggi 2. Kemristekdikti Dukung Metode Pembelajaran Inovatif Eropa 3. Proyek Indoped untuk Tingkatkan Mutu Pendidikan Tinggi 4. Uni Eropa Pilih Lima Kampus Indonesia untuk Proyek Indoped 	<p>3. Juara di AS, Tim Robotik Unissula Diganjar Hadiah ke Mekah</p> <p><i>Cluster 2 (purity=1):</i></p> <ol style="list-style-type: none"> 1. Zaskia Gotik Jadi 'Duta Pancasila,' Kok Bisa? 2. Zaskia Gotik Jadi Duta Pancasila, Denny Cagur Sambut Baik 3. Ini Alasan PKB MPR Mengangkat Eneng Jadi Duta Pancasila <p><i>Cluster 3 (purity=1):</i></p> <ol style="list-style-type: none"> 1. Kolaborasi Kampus Indonesia-Eropa untuk Modernisasi Perguruan Tinggi 2. Kemristekdikti Dukung Metode Pembelajaran Inovatif Eropa 3. Proyek Indoped untuk Tingkatkan Mutu Pendidikan Tinggi 4. Uni Eropa Pilih Lima Kampus Indonesia untuk Proyek Indoped <p><i>Cluster 4 (purity=0.75):</i></p> <ol style="list-style-type: none"> 1. Ketahui Manfaat Kebiasaan Membawa Bekal 2. Lima Kebiasaan Mahasiswa yang Hambat Skripsi 3. Belajar Kelola Tinja, Pemkot Bekasi Harus Jauh-Jauh ke Filipina 4. Pemkot Bekasi Studi Banding Tinja ke Filipina
---	---

Pada Tabel 2, dapat dilihat bahwa daftar judul artikel yang menjadi anggota *cluster*. *Cluster* yang terbentuk pada $\text{min_sim}=0.4$ memiliki kemiripan dengan *cluster* $\text{min_sim}=0.3$. *Cluster*-1 (Tim Robotik Unissula Juara di AS) dan *cluster* 5 (Proyek Indoped) dari $\text{min_sim}=0.3$ sama dengan *cluster*-1 dan *cluster* 3 dari $\text{min_sim}=0.4$. Untuk *cluster*-2 (Zaskia Gotik Jadi Duta Pancasila), walaupun keduanya memiliki *purity*=1, anggota *cluster*-2 dari $\text{min_sim}=0.4$ merupakan subset *cluster*-2 dari $\text{min_sim}=0.3$. *Cluster*-4 dari $\text{min_sim}=0.4$ dan *cluster*-3 dari $\text{min_sim}=0.3$ memiliki topik yang beragam karena *purity*-nya belum 1.

4. Simpulan

Agglomerative clustering telah diimplementasikan dan diaplikasikan untuk 104 artikel berbahasa Indonesia. Kualitas *cluster* terbaik dihasilkan *agglomerative clustering* dengan menggunakan *complete linkage* dan kemiripan minimum 0.3 (*purity* rata-rata 0.888 dan 5 *cluster*) dan 0.4 (*purity* rata-rata 0.938 dan 4 *cluster*). *Complete linkage* menghasilkan *purity* rata-rata terbaik dan konsisten dibandingkan jenis

linkage lainnya. Selain itu, diperlihatkan juga bahwa nilai *purity* akan semakin tinggi jika *min_sim* diperbesar, tetapi jumlah *cluster* yang dihasilkan semakin kecil.

Penelitian berikutnya perlu membandingkan kualitas *cluster* yang dihasilkan oleh teknik *clustering* lainnya dengan menggunakan dataset yang sama. Selain itu, perlu dieksplorasi agar jumlah *cluster* yang dihasilkan semakin mendekati jumlah label seharusnya.

Daftar Pustaka

- [1] M. L. Khodra, Y. Wibisono, A. Fauzan, B. Rahman. Pembangunan Sistem Cerdas Agregator Berita Bahasa Indonesia. Konferensi Nasional Sistem Informasi 2015.
- [2] Wibisono Y., dan Khodra, M. L., *Clustering* Berita Berbahasa Indonesia, *Jurnal FPMIPA UPI*. 2005.
- [3] Husni, Y.D.P. Negara, M. Syarief. *Clusterisasi* Dokumen Web (Berita) Bahasa Indonesia Menggunakan Algoritma K-Means. *Jurnal SimanteC*. Vol. 4, No. 3 Juli 2015.
- [4] D.P.Langgeni, ZK.A.Baizal, Y. A.W. Firdaus. *Clustering Artikel Berita Berbahasa Indonesia Menggunakan Unsupervised Feature Selection*. Seminar Nasional Informatika 2010 (semnasIF 2010). Yogyakarta, 22 Mei 2010.
- [5] A. Hamza. Peningkatan Kinerja *Clustering* Dokumen Teks Menggunakan Pembobotan Sampel. Seminar Nasional Informatika 2011.
- [6] T. Alfina, B. Santosa, dan A. R. Barakbah. Analisa Perbandingan Metode Hierarchical *Clustering*, K-means dan Gabungan Keduanya dalam *Cluster* Data (Studi kasus : Problem Kerja Praktek Jurusan Teknik Industri ITS). *Jurnal Teknik ITS* Vol. 1 (Sept, 2012)
- [7] Ambarwati, E. Winarko. Pengelompokan Berita Indonesia Berdasarkan Histogram Kata Menggunakan Self-Organizing Map. *IJCCS*, Vol.8, No.1, January 2014, pp. 101~110
- [8] A. Z. Arifin dan A. N. Setiono. Klasifikasi Dokumen Berita Kejadian Berbahasa Indonesia dengan Algoritma Single Pass *Clustering*. Prosiding SITIA ITS, 2002.
- [9] A. Z. Arifin, R. Darwanto, D. A. Navastara, dan H. T. Ciptaningtyas. Klasifikasi Online Dokumen Berita dengan Menggunakan Algoritma Suffix Tree *Clustering*. Prosiding Seminar Sistem Informasi Indonesia (SESINDO2008) ITS, 2008
- [10] Pengyu Hong. Introduction to Hierarchical *Clustering* Analysis. 2005
- [11] R.Xu, D.Wunsch II. Survey of *Clustering* Algorithms. *IEEE Transactions On Neural Networks*, Vol. 16, No. 3, May 2005