Jurnal Ilmiah
ILMU KOMPUTER
Universitas Udayana

Vol. 6, No. 2, September 2013 **ISSN 1979-5661**

PENDEKATAN POSITIONAL TEXT GRAPH UNTUK PEMILIHAN KALIMAT REPRESENTATIF CLUSTER PADA PERINGKASAN MULTI-DOKUMEN

I Putu Gede Hendra Suputra, Agus Zainal Arifin, Anny Yuniarti

Jurusan Teknik Informatika, Fakultas Teknologi Informasi, Institut Teknologi Sepuluh Nopember (ITS)

Kampus ITS, *Sukolilo,* Surabaya 60111, Indonesia

Email: hendra.suputra@gmail.com

ABSTRAK

Coverage and saliency are major problems in Automatic Text Summarization. Sentence clustering approaches are methods able to provide good coverage on all topics, but the point to be considered is the selection of important sentence that can represent the cluster's topic. The salient sentences selected as constituent to the final summary should have information density so that can convey important information contained in the cluster. Information density from the sentence can be mined by extracting the sentence information density (SID) feature that built from positional text graph approach of every sentence in the cluster. This paper proposed a cluster representative sentence selection strategy that used the positional text graph approach in multi-document summarization. There are three concepts that used in this paper: (1) sentence clustering based on similarity based histogram clustering, (2) cluster ordering based on cluster importance and (3) representative sentence selection based on sentence information density feature score. The candidate summary sentence is a sentence that has greatest sentence information density feature score of a cluster. Trials conducted on task 2 DUC 2004 dataset. ROUGE-1 measurement was used as performance metric to compare the use of SID feature with other method namely Local Importance and Global Importance (LIGI). Test result showed that the use of SID feature was successfully outperform LIGI method based on ROUGE-1 values where the greatest average value of ROUGE-1 that achieved by SID features is 0.3915.

***Kata kunci**: multi-document summarization, sentence clustering, similarity based histogram clustering, sentence information density, positional text graph.*

ABSTRAK

Coverage dan saliency adalah masalah utama dalam peringkasan teks otomatis. Pendekatan clustering kalimat merupakan metode yang mampu memberikan cakupan yang baik (good coverage) terhadap semua topik, namun yang perlu diperhatikan adalah pemilihan kalimat penting (salient setence) yang mampu merepresentasikan topik cluster. Kalimat-kalimat penting yang terpilih menjadi penyusun ringkasan akhir harus memiliki kepadatan informasi sehingga mampu menyampaikan informasi penting yang terkandung di dalam cluster. Kepadatan informasi dari kalimat dapat digali dengan mengekstraksi fitur sentence information density (SID) yang dibangun dari pendekatan positional text graph setiap kalimat di dalam cluster. Pada paper ini diajukan sebuah strategi pemilihan kalimat representatif cluster menggunakan pendekatan positional text graph pada peringkasan multi-dokumen. Terdapat tiga konsep yang digunakan pada paper ini yaitu: (1) clustering kalimat dengan similarity based histogram clustering, (2) pengurutan cluster berdasarkan cluster importance dan (3) pemilihan kalimat representatif berdasarkan skor fitur sentence information density. Kalimat yang menjadi kalimat kandidat ringkasan adalah kalimat yang memiliki skor fitur sentence information density terbesar dari suatu cluster. Uji coba dilakukan terhadap dataset task 2 DUC 2004. Pengukuran ROUGE-1 digunakan sebagai pengukuran performa untuk membandingkan penggunaan fitur SID dengan metode lain yaitu local importance dan global importance (LIGI). Dari hasil ujicoba didapatkan bahwa penggunaan fitur SID berhasil mengungguli metode LIGI berdasarkan nilai ROUGE-1 dimana nilai rata-rata ROUGE-1 terbesar yang mampu diraih fitur SID 0.3915.

***Kata kunci**: peringkasan multi-dokumen, clustering kalimat, Similarity based Histogram Clustering, sentence information density, positional text graph.*

1. PENDAHULUAN

Sistem peringkasan multi-dokumen (*Multi-document Summarization*) berdasarkan metode *extractive* tengah menjadi perhatian para peneliti dalam beberapa tahun terakhir (Randev dkk, 2004; He dkk, 2008; Wan dan Yang, 2008; Sarkar, 2009; Kogilavani dan Balasubramani, 2010; Ge dkk, 2011; Ouyang dkk, 2013). Peringkasan berdasarkan metode

extractive terdiri dari proses pemilihan kalimat-kalimat penting dari dokumen sumber dan proses penyusunan kalimat-kalimat tersebut menjadi bentuk yang lebih singkat (Gupta, 2010). Kalimat-kalimat yang diekstraksi dari dokumen sumber dipilih dengan suatu kriteria tertentu. Walaupun cara ekstraksi kalimat pada metode peringkasan *extractive* bukan merupakan cara yang umum digunakan manusia, namun metode peringkasan *extractive* masih banyak digunakan untuk peringkasan multi-dokumen. Oleh karena itu ringkasan yang dihasilkan dari metode peringkasan *extractive* dituntut harus mampu menyampaikan konten-konten penting dari dokumen-dokumen sumber.

Ringkasan yang baik adalah ringkasan yang mampu mencakup (*coverage*) sebanyak mungkin konsep-konsep penting (*saliency*) yang ada pada dokumen sumber (Ouyang dkk, 2013). *Coverage* dan *saliency* adalah masalah utama dalam metode peringkasan dimana strategi pemilihan kalimat pada metode *extractive* menjadi sangat penting karena harus mampu memilih kalimat-kalimat utama (penting) dan terhindar dari redundansi (*redundancy*) sehingga mampu mencakup banyak konsep.

Clustering kalimat adalah suatu metode alternatif yang mampu memberikan good *coverage* pada ringkasan (Boros dkk, 2001; Sarkar, 2009;). Pencapaian *good coverage* pada ringkasan tidak terlepas dari koherensi *cluster* yang baik. Salah satu metode yang dapat menjamin koherensi *cluster* adalah *Similarity Based Histogram Clustering* (SHC) (Hammouda dan Kamel, 2003; Sarkar, 2009). SHC pertama kali diajukan oleh Hammouda dan Kamel pada tahun 2003. Metode SHC terbukti lebih baik jika dibandingkan dengan *Hierarchical Agglomerative Clustering* (HAC), *Single-Pass Clustering* dan *K-Nearest Neighbor Clustering*. Sarkar (2009) menggunakan pendekatan *clustering* kalimat dengan SHC untuk peringkasan multi-dokumen. *Cluster-cluster* kalimat yang terbentuk selanjutnya diurutkan berdasarkan *cluster importance* dan kemudian dilakukan pemilihan sebuah kalimat representatif pada setiap *cluster* berdasarkan bobot kalimat terbesar yang dihasilkan dari fungsi kombinasi antara fitur *local importance* dan fitur *global importance*. Kalimat representatif tersebut adalah kalimat penting yang menjadi kalimat penyusun ringkasan.

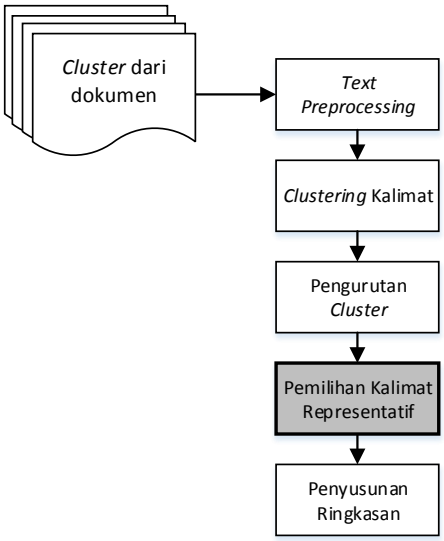
Strategi pemilihan kalimat representatif menjadi sangat penting untuk memecahkan masalah *saliency*. Kalimat yang terpilih harus mampu mewakili topik dari suatu *cluster* tertentu (Sarkar, 2009). Kalimat penting penyusun ringkasan harus memiliki kepadatan informasi yaitu mengandung informasi sebanyak mungkin dari dokumen sumber (He dkk, 2008). Menurut He dkk (2008) fitur kepadatan informasi kalimat (*sentence information density*) dapat digali dengan pendekatan *positional text graph* yang memperhatikan bobot *similarity* suatu kalimat dengan kalimat-kalimat lain. Pendekatan tersebut juga baik untuk menentukan kalimat-kalimat yang informatif karena kalimat tersebut saling berbagi informasi dengan kalimat-kalimat lain (Gupta, 2010). He dkk (2008) mengunakan pendekatan *positional text graph* sebagai salah satu fitur yang digunakan untuk memilih kalimat-kalimat ringkasan. Pada penelitian tersebut dilaporkan bahwa pendekatan *positional text graph* mampu bekerja efektif untuk memilih kalimat ringkasan yang memiliki kepadatan informasi mampu meraih performa terbaik pada sistem peringkasan multi-dokumen.

Jika diaplikasikan pada pendekatan *clustering* kalimat, dari pernyataan-pernyataan diatas dapat disimpulkan bahwa kalimat representatif yang terpilih sebagai penyusun ringkasan harus memiliki kepadatan informasi dan mampu mencerminkan topik dari suatu *cluster* kalimat.

Oleh karena itu, pada paper ini diusulkan sebuah fitur *sentence information density* yang dibangun dengan pendekatan *positional text graph* sebagai suatu strategi baru untuk memilih kalimat ringkasan pada sistem peringkasan multi-dokumen berdasarkan metode *clustering* kalimat.

2. METODE

Pada bagian ini dijelaskan secara detail berbagai yang dilalui pada sistem peringkasan multi dokumen otomatis yang digunakan pada paper ini. Proses-proses yang dilalui dalam sistem tersebut disusun sesuai *framework* pada Gambar 1.



Gambar 1. *Framework* Peringkasan Multi-Dokumen Berdasarkan Metode *Clustering* Kalimat

2.1 *Text Preprocessing*

Text preprocessing adalah yang pertama dilakukan sebelum input dokumen diolah lebih lanjut menjadi *cluster-cluster* kalimat. Adapun proses-proses yang dilalui dalam *text preprocessing* adalah *segmentation* (segementasi), *stopword removal* dan *stemming*. Pada paper ini *segmentation* dilakukan

terhadap kata dan kalimat. Setiap kata-kata yang diperoleh dari hasil segmentasi input melalui proses *stopword removal*. Selanjutnya menjalani proses *stemming* dengan algoritma *Porter Stemmer*.

2.2 *Clustering* Kalimat

Clustering kalimat adalah bagian yang penting dalam sistem peringkasan otomatis karena setiap topik dalam set dokumen harus diidentifikasi secara tepat untuk menemukan *similarity* dan *dissimilarity* yang ada dalam dokumen sehingga menjamin *good coverage* (Sarkar, 2009). Jika kalimat-kalimat dikelompokkan ke dalam sejumlah *cluster* yang telah ditentukan, *cluster* mungkin tidak koheren karena beberapa kalimat bisa saja terpaksa menjadi salah satu anggota *cluster* meskipun seharusnya tidak. *Cluster-cluster* tidak koheren mungkin mengandung unit-unit teks yang terduplikasi pada *cluster* yang berbeda dan menyebabkan pemilihan kalimat menjadi redundan untuk ringkasan. Sebaliknya, jika *cluster* sangat ketat, sebagian besar *cluster* menjadi *singletons*. Dengan demikian, harus dipilih metode *clustering* yang menjamin koherensi *cluster*. Pada paper ini digunakan algoritma (*Similarity Based Histogram Ratio*) SHC yang diadopsi dari (Hammouda dan Kamel, 2003). SHC dipilih karena SHC menggunakan pendekatan *cluster similarity histogram* yang berguna untuk memonitor dan menjaga koherensi dari *cluster*.

2.2.1 Pengukuran *Similarity*

Metode pengukuran *similarity* kalimat yang digunakan pada paper ini adalah metode *uni-gram matching-based similarity measure* sesuai Persamaan 1. Metode tersebut diajukan oleh Sarkar (2009) karena penggunaan *cosine similarity* memiliki kelemahan jika digunakan untuk mengukur *similarity* kalimat. Metode *Uni-gram matching-based similarity measure* dipilih mengingat kalimat-kalimat adalah unit yang sangat pendek dan *similarity* akan bernilai sangat kecil ketika dihitung dengan pengukuran *cosine similarity*.

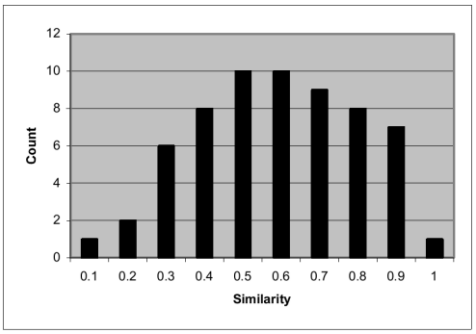
$$sim(s_i, s_j) = \frac{(2^* | s_i | \cap | s_j |)}{| s_i | + | s_j |}, \qquad (1)$$

dimana s_i dan s_j adalah kalimat *s* ke-*i* dan ke-*j*. $| s_i | \cap | s_j |$ merepresentasikan jumlah dari kata-kata yang sesuai (*match*) antara kalimat *s* ke-*i* dan kalimat *s* ke-*j*. $|s_i|$ adalah panjang kalimat *s* ke-*i* yaitu jumlah kata yang menyusun kalimat tersebut.

2.2.2 *Clustering* Kalimat dengan SHC

Konsep utama dari SHC adalah menjaga setiap *cluster* sedapat mungkin berada dalam kondisi koheren pada tingkat yang baik. Hal tersebut dapat dimonitor pada kecenderungan anggota *bin* yang terlihat dari *cluster similarity histogram* pada Gambar 2. Masing-masing *bin* terdiri dari jumlah *similarity* dari masing-masing pasangan anggota *cluster* yang memenuhi interval pada *bin*. *Cluster* yang baik memiliki *similarity histogram* dengan nilai *similarity* dimana untuk setiap pasangan kalimatnya berada pada *bin* maksimum, sedangkan jika cenderung berada pada *bin* minimum maka kualitas *cluster* tersebut cenderung buruk.

Terdapat 10 *bin* pada Gambar 2. Setiap tinggi *bin* merepresentasikan total jumlah dari nilai pasangan *similarity* sesuai dengan interval dari *bin* tersebut. Koherensi antar anggota *cluster* adalah hal yang dioptimalkan pada SHC. Sebuah nilai *threshold* digunakan untuk menentukan tingkat toleransi sebuah *cluster* menerima anggota *cluster* baru. Jika SHC digunakan untuk kasus *clustering* kalimat maka setiap kalimat yang menjadi kandidat anggota *cluster* harus mampu meningkatkan koherensi *cluster* tersebut atau memenuhi nilai *threshold* yang telah ditentukan sebelumnya.



Gambar 2. *Cluster Similarity Histogram* (Sarkar, 2009)

Salah satu cara untuk mendapatkan derajat koherensi yang tinggi dalam *cluster* adalah mempertahankan derajat *similarity* antar anggota tetap tinggi. Dalam *similarity histogram*, ini berarti menjaga distribusi *similarity* agar cenderung ke kanan (ke arah nilai *similarity* 1).

Koherensi yang baik pada setiap *cluster* sangat penting untuk dijaga. Hal tersebut untuk mencegah adanya kalimat-kalimat yang redundan pada pemilihan kalimat penyusun ringkasan. Kualitas dari suatu *similarity histogram* yang merepresentasikan koherensi *cluster* ditentukan dengan menghitung rasio *similarity* yang berada diatas *threshold* dengan total jumlah *similarity* yang ada. Rasio dari *histogram* yang tinggi mencerminkan koherensi yang tinggi pula.

Jika *n* adalah jumlah dari kalimat pada suatu *cluster*, maka jumlah dari pasangan kalimat yang ada pada *cluster* tersebut adalah $n(n+1)/2$. $Sim=\{sim_1, sim_2, sim_3, \ldots, sim_m\}$ adalah kumpulan dari pasangan *similarity* antar kalimat, dimana $m=n(n+1)/2$. *Similarity histogram* dari *cluster* dinotasikan dengan $H=\{h_1, h_2, h_3, ..., h_{nb}\}$. Jumlah dari *bin* yang ada pada suatu *histogram* dinotasikan dengan *nb* sedangkan jumlah *similarity* kalimat yang ada pada *bin* ke-*i* dinotasikan dengan h_i. Fungsi untuk menghitung nilai h_i ditunjukkan pada Persamaan 2.

$$h_i = count(sim_j) \text{ untuk } sim_{li} \le sim_j \le sim_{ui}, \quad (2)$$

dimana sim_{li} ialah batas bawah *similarity* pada *bin* ke-*i* sedangkan sim_{ui} ialah batas atas *similarity* pada *bin* ke-*i*.

Histogram ratio (*HR*) dari suatu *cluster* dapat dihitung dengan Persamaan 3.

$$HR = \frac{\sum_{i=T}^{n_b} h_i}{\sum_{j=1}^{n_b} h_j} \quad (3)$$

$$T = \lfloor S_T * n_b \rfloor, \quad (4)$$

S_T adalah *similarity threshold*. Persamaan 4 menunjukkan jumlah *bin* yang sesuai dengan *similarity threshold* yang dinotasikan dengan *T*.

Penyertaan elemen yang buruk pada suatu *cluster* pada setiap tahap mungkin dapat mempengaruhi kualitas *cluster* dan hal tersebut mungkin menurunkan nilai *HR* menjadi nol. Untuk mencegah masalah tersebut maka ditetapkan sebuah *threshold* minimum untuk *histogram ratio* yaitu HR_{min}. Setiap *cluster* harus berpedoman kepada HR_{min}. Langkah-langkah dari algoritma SHC diperlihatkan pada *pseudocode* pada Gambar 3.

```
1: N ← Empty List {Cluster List}
2: for each sentence s do
3:    for each cluster c in N do
4:       HR_old = HR_C
5:       Simulate adding s to c
6:       HR_new = HR_C
7:       if (HR_new ≥ HR_old) OR ((HR_new > HR_min)
       AND (HR_old−HR_new < ε)) then
8:          Add s to c
9:          exit
10:      end if
11:   end for
12:   if s was not added to any cluster then
13:      Create a new cluster c
14:      ADD s to c
15:      ADD c to L
16:   end if
17: end for
```

Gambar 3. *Pseudocode Clustering* Kalimat Menggunakan SHC

Mengacu pada Gambar 3, metode SHC berjalan secara bertahap dengan menguji setiap kalimat dimana *cluster* silmilarity *histogram* dari setiap *cluster* dihitung sebelum dan sesudah melakukan simulasi penambahan kalimat baru pada suatu *cluster* (baris 4-6). Pada saat membentuk *cluster* baru nilai *histogram ratio* disimpan pada HR_c yang dihitung menggunakan Persamaan 3. HR_{old} adalah nilai *histogram ratio* pada suatu *cluster* sedangkan HR_{new} adalah *histogram ratio* yang diperoleh setelah sebuah kalimat dijadikan anggota dari suatu *cluster* yang diujikan. Nilai minimum dari *histogram ratio* dinotasikan dengan HR_{min}. Parameter ε (*epsilon*) adalah sebuah *threshold* sebagai ambang batas selesih antara HR_{old} dengan HR_{new}. Pada proses pembentukan *cluster*, nilai (HR_{old}) dan nilai (HR_{new}) dibandingkan. Jika nilai HR_{new} lebih atau sama dengan nilai HR_{old}, maka kalimat tersebut ditambahkan ke dalam suatu *cluster*. Atau jika nilai HR_{new} lebih rendah dari nilai HR_{old} namun nilai HR_{new} masih berada diatas HR_{min} dan selisih antara HR_{old} dengan HR_{new} tidak lebih dari ε maka kalimat juga ditambahkan sedangkan selain itu kalimat tidak ditambahkan (baris 7-10). Jika suatu kalimat tidak mendapatkan *cluster* setelah diuji dengan semua *cluster*, maka sebuah *cluster* baru dibentuk dan kalimat tersebut menjadi anggotanya (baris 12-16).

2.3 Pengurutan *Clustering Ordering*

Pengurutan *cluster* dilakukan karena pada proses *clustering* menggunakan algoritma SHC tidak pernah ada pengetahuan khusus berapa jumlah *cluster* yang akan terbentuk. Sehingga sangat penting untuk mengetahui *cluster-cluster* mana saja yang terpilih menjadi kandidat ringkasan akhir (Sarkar, 2009).

Cluster importance adalah sebuah metode yang melakukan pengurutan *cluster* berdasarkan nilai penjumlahan bobot dari kata-kata yang merupakan kata *frequent* (sering muncul) yang terkandung dalam *cluster*. Sebuah *threshold* (θ) ditetapkan untuk menentukan apakah suatu kata tersebut termasuk kata *frequent* atau tidak terhadap seluruh dokumen input. Jika frekuensi suatu kata memenuhi *threshold* θ maka kata tersebut dianggap sebagai kata yang memiliki bobot. Pendekatan *cluster importance* bertujuan mengukur pentingnya suatu *cluster* berdasarkan jumlah kata-kata *frequent* yang ada pada suatu *cluster*.

Misal dari hasil *clustering* kalimat terbentuk N buah *cluster* sehingga $C=\{c_1, c_2, c_3, \ldots, c_N\}$. Pengurutan dilakukan dengan mencari bobot dari *cluster* ke-1 sampai ke-*N* yang diurutkan berdasarkan bobot *cluster importance*. Pengurutan *cluster* berdasarkan bobot *cluster importance* dihitung dengan Persamaan 5.

$$Weight(c_j) = \sum_{w \in c_j} \log(1 + count(w)), \quad (5)$$

dimana bobot dari *cluster c* ke-*j* dinotasikan dengan *Weight*(*cj*), *count (w)* adalah jumlah dari kata *w* pada koleksi *input* dan *count (w)* lebih dari *threshold* θ. Bobot dari *cluster* merepresentasikan kekayaan informasi yang dikandung dalam *cluster* tersebut. Sebelum menghitung bobot suatu *cluster* terlebih dahulu dilakukan *stopwords removal* yaitu menghilangkan kata-kata yang tidak penting yang terdapat pada kumpulan dokumen input.

2.4 Pemilihan Kalimat Representatif

Pemilihan kalimat yang menjadi kalimat ringkasan dalam paper ini didasarkan pada tingginya skor suatu kalimat didalam suatu *cluster* tertentu. Penentuan skor kalimat dihitung berdasarkan skor fitur *sentence information density*. Fitur *sentence information density* adalah fitur yang mampu mencerminkan banyaknya kandungan informasi dari suatu kalimat sehingga dapat mewakili kalimat-kalimat lainya. Proses pemilihan kalimat representatif *cluster* berdasarkan skor fitur *sentence information density* selanjutnya disebut dengan metode SID.

Metode SID menggunakan pendekatan *positional text graph* yang diadopsi dari He dkk (2008). Dalam paper ini digunakan pendekatan *clustering* kalimat sehingga *sentence information density* dihitung mengacu pada pasangan-pasangan *similarity* kalimat yang ada dalam suatu *cluster* kalimat. Nilai *similarity* untuk setiap kalimat yang ada pada *cluster* dihitung dengan kalimat lainnya kemudian dibentuk *similatiry matrix* yang digunakan untuk membentuk *graph* kalimat yang merepresentasikan *position information*. *Graph* digambarkan sebagai $P = (V, E)$, dimana P merepresentasikan *graph*, $V = \{s_1, s_2, \ldots, s_n\}$ adalah *vertex* pada *graph* yang merepresentasikan kalimat-kalimat dalam suatu *cluster*, dan $E=\{(s_i, s_j)\}$ adalah kumpulan dari suatu *edge* pada *graph* dimana bobot *edge* tersebut dihitung berdasarkan *similarity* antar dua kalimat dalam *cluster*.

Graph P dibangun berdasarkan kalimat-kalimat dalam *cluster*. Saat pertama *graph P* kosong, setelah itu semua kalimat dalam suatu *cluster* dimasukan sebagai *vertex*. Langkah kedua hitung nilai *similarity* untuk setiap pasangan kalimat dalam *P*, jika nilai *similarity* suatu pasangan kalimat memenuhi *threshold* α maka *edge* dibentuk dan bobot pasangan kalimat tersebut adalah nilai *similarity* yang dimilikinya. Ketika *graph* telah dibangun, fitur *sentence information density* dihitung dengan Persamaan 6.

$$F_{sid}(s_{kj}) = \frac{W_{s_{kj}}}{\max_{l\in\{1,2,..n\}} W_{s_{lj}}} \quad (6)$$

dimana jumlah kalimat *s* pada *cluster* ke-*j* ditunjukkan dengan *n*, adalah penjumlahan bobot dari semua *edge* yang datang dari kalimat *s* ke-*k* pada *cluster* ke-*j*, sedangkan adalah penjumlahan bobot *edge* maksimum diantara semua pasangan *similarity* kalimat yang ada pada *cluster* ke-*j*.

2.5 Penyusunan Ringkasan

Setelah *clustering* kalimat, *cluster*-*cluster* yang terbentuk diurutkan berdasarkan *cluster importance*. Sebuah kalimat representatif dipilih dari setiap *cluster*. Pemilihan kalimat dimulai dari *cluster* yang memiliki bobot *cluster importance* paling tinggi. Kemudian pemilihan dilanjutkan pada *cluster* berikutnya sesuai dengan daftar urutan *cluster* berdasarkan bobot *cluster importance* secara *descending*. Pemilihan kalimat tersebut terus dilakukan hingga panjang ringkasan yang diharapkan terpenuhi.

3. EVALUASI RINGKASAN

Memang sulit untuk menilai apakah ringkasan yang dihasilkan baik atau buruk. Evaluasi manual biasanya bersifat subjektif dan umumnya membutuhkan banyak usaha dari manusia (ahli). Dewasa ini pengukuran evaluasi ringkasan secara otomatis tengah populer. Salah satu pengukuran otomatis yang terkenal adalah ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*). ROUGE adalah metode evaluasi hasil ringkasan yang mengkukur kualitas hasil ringkasan berdasarkan kesesuaian antara unit-unit ringkasan hasil sistem dengan unit-unit ringkasan referensi yang dibuat secara manual. Pada paper ini digunakan metode ROUGE-*N*. Pengukuran ROUGE-*N* mengukur perbandingan *N-gram* dari dua ringkasan, dan menghitung berapa jumlah yang sesuai. Perhitungan ROUGE-*N* yang diadopsi dari perhitungan Lin (2004) ditunjukkan pada Persamaan 7.

$$ROUGE-N = \frac{\sum_{S\in Summ_{ref}} \sum_{N-gram\in S} Count_{match}(N-gram)}{\sum_{S\in Summ_{ref}} \sum_{N-gram\in S} Count(N-gram)}, \quad (7)$$

dimana *N* menunjukkan panjang dari *N-gram*, $Count_{match}$*(N-gram)* adalah jumlah maksimum dari *N-gram* yang muncul pada ringkasan kandidat dan ringkasan sebagai referensi. *Count(N-gram)* adalah jumlah dari *N-gram* pada ringkasan sebagai referensi. Pada paper ini fungsi ROUGE-*N* yang digunakan adalah ROUGE dengan nilai $N = 1$.

Sesuai dengan karakteristik *dataset task* 2 DUC 2004 yang menggunakan *multiple references summaries* (banyak referensi ringkasan) yaitu empat referensi per *cluster* dokumen, maka perhitungan nilai ROUGE-*N* akhir dihitung berdasarkan Persamaan 8. Nilai akhir dari ROUGE-*N* akhir adalah nilai ROUGE-*N* terbesar yang dihasilkan dari pasangan ringkasan hasil sistem dan ringkasan referensi. Nilai ROUGE-N dihitung pada setiap pasangan ringkasan kandidat *sc* dan ringkasan referensi rs_i. Perhitungan ROUGE-N tersebut diadopsi dari Lin (2004) yang ditunjukkan pada Persamaan 8.

$$ROUGE-N_{multi} = \text{argmax}_i\, ROUGE-N(sc, rs_i) \quad (8)$$

4. HASIL UJI COBA DAN PEMBAHASAN

Pada baigan ini dipaparkan hasil *testing* metode usulan SID dengan metode pembanding yaitu metode *local importance* dan *global importance* (selanjutnya

disebut LIGI). Kedua metode baik SID dan LIGI dijalankan pada *framework* sistem peringkasan yang sama sesuai Gambar 1.

Pada paper ini *dataset* yang digunakan adalah *task* 2 DUC 2004. *Dataset task* 2 DUC 2004 merupakan kumpulan dokumen berita dalam bahasa Inggris dari *Associated Press* dan *New York Times*. Dokumen-dokumen tersebut telah terkelompok-kelompok menjadi 50 *cluster* dokumen. Setiap *cluster* dokumen terdiri dari rata-rata 10 dokumen berita.

Pengukuran performa dari masing-masing metode (SID dan LIGI) dilihat dari nilai ROUGE-1. Nilai ROUGE-1 yang lebih besar menunjukkan performa metode yang lebih baik dari segi korelasi ringkasan yang dihasilkan sistem dengan ringkasan secara manual.

4.1 Inisialisai Nilai-nilai Parameter Sistem

Terdapat lima buah parameter yang digunakan dalam paper ini yaitu HR_{min}, ε, S_T, θ, dan α. Parameter HR_{min}, ε, dan S_T adalah parameter yang digunakan pada proses *clustering* kalimat dengan algoritma SHC. Parameter θ adalah parameter yang digunakan pada proses pengurutan *cluster* berdasarkan *cluster importance*. Sedangkan parameter α adalah parameter yang digunakan pada proses pemilihan kalimat representatif. Semua parameter tersebut dikombinasikan kecuali nilai *HRmin* dan ε yang sengaja dibuat berpasangan sesuai dengan kolom yaitu: [*HRmin*, ε] menjadi [0.8, 0.2], [0.7, 0.3], [0.6, 0.4], dan [0.5, 0.5]. Sesuai dengan inisialisasi nilai parameter yang terdapat pada Tabel 1, sehingga dari semua kemungkinan kombinasi akhirnya diperoleh 72 buah kombinasi nilai parameter untuk metode SID.

Metode Pembanding LIGI dijalankan pada *framework* yang sama dengan metode SID. Sehingga akan terdapat empat buah parameter yang dikombinasikan yaitu HR_{min}, ε, S_T, dan θ. Parameter α tidak digunakan karena parameter tersebut hanya digunakan pada perhitungan metode SID. Bobot untuk fitur LI dan fitur GI yang digunakan adalah 0.5 sesuai rekomendasi Sarkar (2009). Sehingga sesuai dengan inisialisasi nilai parameter pada Tabel 1, akhirnya diperoleh 24 buah kombinasi nilai parameter untuk metode LIGI.

Tabel 1. Inisialisasi Nilai Parameter Sistem

Parameter	Inisialisasi Nilai Parameter
HR_{min}	{0.8, 0.7, 0.6, 0.5}
ε	{0.2, 0.3, 0.4, 0.5}
S_T	{0.4, 0.5, 0.6}
θ	{10, 20}
α	{0.4, 0.5, 0.6}

4.2 *Testing* Metode SID dengan metode LIGI

Berikut hasil *testing* metode SID dan LIGI terhadap seluruh *dataset task* 2 DUC 2004. Metode SID dan LIGI sama-sama berjalan pada *framework* peringkasan yang sama yaitu: *text preprocessing*, *clustering* kalimat dengan SHC, dan pengurutan *cluster* yang dilakukan berdasarkan *cluster importance*. Namun untuk metode LIGI pada *text preprocessing* tidak digunakan proses *stemming* sesuai dengan Sarkar (2009).

Tabel 2. Hasil *Testing* Metode SID vs LIGI

Metode Sistem Peringkasan	ROUGE-1
Sentence clustering (SHC) + *cluster ordering* (*cluster importance*) + SID	**0.3915**
Sentence clustering (SHC) + *cluster ordering* (*cluster importance*) + LIGI	0.3905

Tabel 2 menunjukkan perbandingan nilai rata-rata ROUGE-1 yang mampu diraih oleh metode SID dan LIGI yang dijalankan pada *framework* sistem peringkasan yang sama. Nilai rata-rata ROUGE-1 terbesar diraih oleh metode SID yaitu 0.3915 dengan kombinasi nilai parameter HR_{min} = 0.6, ε = 0.4, S_T = 0.4, θ = 10 dan α = 0.4 sedangkan nilai rata-rata ROUGE-1 terbesar yang mampu diraih oleh metode LIGI adalah 0.3905 dengan kombinasi nilai parameter HR_{min} = 0.8, ε = 0.4, S_T = 0.2, dan θ = 10. Nilai ROUGE-1 yang lebih besar menunjukkan performa sistem peringkasan yang dibangun dengan pendekatan *sentence clustering* (SHC) + *cluster ordering* (*cluster importance*) + SID lebih baik dari segi korelasi hasil ringkasan sistem dengan ringkasan secara manual dibandingkan dengan sistem peringkasan yang dibangun dengan pendekatan *sentence clustering* (SHC) + *cluster ordering* (*cluster importance*) + LIGI.

5. KESIMPULAN

Pada paper ini telah diusulkan sebuah strategi pemilihan kalimat representatif *cluster* sebagai kalimat ringkasan dengan menggunakan konsep *sentence information density* (metode SID) yang dibangun menggunakan pendekatan *positional text graph*. Pemilihan kalimat representatif *cluster* menggunakan pendekatan *positional text graph* mampu menghasilkan nilai rata-rata ROUGE-1 sebesar 0.3915 dan berhasil mengungguli metode LIGI dimana nilai rata-rata ROUGE-1 yang dihasilkan adalah 0.3905. Hal tersebut mengindikasikan bahwa metode SID mampu memilih kalimat ringkasan lebih baik dan memberikan hasil yang lebih baik berdasarkan korelasi hasil ringkasan sistem dengan ringkasan secara manual dari segi kesesuaian *unigram* dibandingkan dengan metode LIGI.

6. DAFTAR PUSTAKA

Boros, E. Kantor, P. B. dan Neu, D. J. (2001), "A *Clustering* Based Approach to Creating Multi-Document Summaries". In Proceedings of the 24th ACM SIGIR Conference, Eds: Kraft, D. H. et al., ACM, New Orleans, Los Angeles, hal. 1-4.

Ge, S. S., Zhang Z., dan He, H. (2011), "Weighted Graph Model Based Sentence *Clustering* and Ranking for Document Summarization" Proceeding of 2011 4th International Conference on Interaction Sciences (ICIS), National University of Singapore, Singapore, hal. 90-95.

Gupta, V. (2010), "A Survey of *Text* Summarization Extractive Tecniquies", Journal of Emerging Technologies in Web Intelligence, Vol. 2, No. 3, hal. 258-268.

Hammouda, K. M. dan Kamel, M. S. (2003), "Incremental Document *Clustering* Using Cluster Similarity Histograms" Proceeding of the 2003 IEEE/WIC International Conference on Web Intelligence, Eds: Liu, J. et al., University of Waterloo, Halifax, Canada, hal. 597-601.

He, T., Li F., Shao, W., Chen, J., dan Ma, L. (2008), "A New Feature-Fusion Sentence Selecting Strategy for Query-Focused Multi-document Summarization", Proceeding of International Conference Advance Language Processing and Web Information Technology, Eds: Ock C. et al., University of Normal, Wuhan, China, hal. 81-86.

Kogilavani, A. dan Balasubramani, P. (2010), "*Clustering* and Feature Sprecific Sentence Extraction Based Summarization of Multiple Documents", International Journal of Computer Science & Information Technology (IJCSIT), Vol. 2, No. 4, hal. 99-111.

Lin, C. Y. (2004), "ROUGE: a Package for Automatic Evaluation of Summaries", In Proceedings of Workshop on *Text* Summarization Brances Out, Eds: Moens, M. F. dan Szpakowicz, S., Association for Computational Linguistics, Barcelona, hal. 74-81.

Ouyang, Y., Li W., Zhang R., Li S., dan Lu Q. (2013), "A Progressive Sentence Selection Strategy for Document Summarization", Journal of information Precessing and Management. Vol. 49, Issue 1, hal. 213-221.

Randev, D. R., Jing, H., Stys, M., dan Tam, D. (2004), "Centroid-Based Summarization of Multiple Documents", Journal Information Processing and Management: an International Journal, Vol. 40 Issue 6, hal. 919-938.

Sarkar, K. (2009), "Sentence *Clustering*-based Summarization of Multiple *Text* Documents", International Journal of Computing Science and Communication Technologies, Vol. 2, No. 1, hal. 325-335.

Wan, X. dan Yang, J. (2008), "Multi-Document Summarization Using Cluster-Based Link Analysis", Proceedings of the 31st annual international ACM SIGIR conference on Research and Development in Information Retrieval, Eds: Chua T. S. et al., Association for Computational Linguistics, New York, hal. 181-184.