Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)
Vol. 1, No. 2, Oktober 2014 hlm. 91-97

SEMANTIC CLUSTERING DAN PEMILIHAN KALIMAT REPRESENTATIF UNTUK PERINGKASAN MULTI DOKUMEN

Pasnur[1], Putu Praba Santika[2], Gus Nanang Syaifuddin[3]

[1,2,3] Jurusan Teknik Informatika, Institut Teknologi Sepuluh Nopember
Kampus ITS Keputih, Sukolilo, Surabaya 60111, Jawa Timur, Indonesia
Email: [1]pasnur13@mhs.if.its.ac.id, [2]praba13@mhs.if.its.ac.id, [3]gusnanang13@mhs.if.its.ac.id

(Naskah masuk: 11 Juni 2014, diterima untuk diterbitkan: 22 Juli 2014)

Abstrak

Coverage dan *saliency* merupakan masalah utama dalam peringkasan multi dokumen. Hasil ringkasan yang baik harus mampu mampu mencakup (*coverage*) sebanyak mungkin konsep penting (*salient*) yang ada pada dokumen sumber. Penelitian ini bertujuan untuk mengembangkan metode baru peringkasan multi dokumen dengan teknik *semantic clustering* dan pemilihan kalimat representatif *cluster*. Metode yang diusulkan berdasarkan prinsip kerja *Latent Semantic Indexing (LSI)* dan *Similarity Based Histogram Clustering (SHC)* untuk pembentukan *cluster* kalimat secara semantik, serta mengkombinasikan fitur *Sentence Information Density (SID)* dan *Sentence Cluster Keyword (SCK)* untuk pemilihan kalimat representatif *cluster*. Pengujian dilakukan pada dataset *Document Understanding Conference (DUC) 2004 Task 2* dan hasilnya diukur menggunakan *Recall-Oriented Understudy for Gisting Evaluation (ROUGE).* Hasil pengujian menunjukkan bahwa metode yang diusulkan mampu mencapai nilai *ROUGE-1* rata-rata sebesar 0,395 dan nilai *ROUGE-2* rata-rata sebesar 0,106.

Kata kunci: *peringkasan multi dokumen, latent semantic indexing, similarity based histogram clustering, sentence information density, sentence cluster keyword*

Abstract

Coverage and saliency is a major problem in multi-document summarization. The good summary should be able to cover (coverage) as much as possible the important concepts (salient) that exist in the source document. This research aims to develop a new method for multiple document summarization with semantic clustering techniques and the selection of representative clusters sentence. The proposed method is based on the principles of Latent Semantic Indexing (LSI) and Similarity Based Histogram Clustering (SHC) for clustering sentences semantically, and combine features of Sentence Information Density (SID) and Sentence Cluster Keyword (SCK) for selecting a representative sentence cluster. Tests are performed on Document Understanding Conference (DUC) 2004 Task 2 dataset and the results are measured using the Recall-Oriented Understudy for Gisting Evaluation (ROUGE). The results show that the proposed method is able to achieve ROUGE-1 value by an average of 0.395 and the ROUGE-2 value by an average of 0.106.

Keywords: *multiple document summarization, latent semantic indexing, similarity based histogram clustering, sentence information density, sentence cluster keyword*

1. PENDAHULUAN

Perkembangan *World Wide Web (WWW)* diikuti dengan pertumbuhan jumlah dokumen digital yang sangat pesat. Hal tersebut menimbulkan permasalahan dalam pencarian dan peringkasan informasi dari berbagai sumber (Sarkar, 2009). Peringkasan yang dilakukan secara manual oleh manusia tidak efisien karena jumlah dokumen yang sangat besar (Gupta & Lehal, 2010).

Peringkasan multi dokumen secara otomatis menjadi perhatian pada beberapa penelitian (Sarkar, 2009; Kogilavani & Balasubramani, 2010; Ouyang, Li, Zhang, Li, & Lu, 2012) sebagai sebuah solusi dalam peringkasan dengan kondisi jumlah dokumen yang sangat banyak (Sarkar, 2009). Peringkasan multi dokumen secara otomatis menghasilkan bentuk dokumen yang lebih ringkas tanpa kehilangan kandungan informasi yang penting (Gupta & Lehal, 2010).

Coverage dan *saliency* merupakan masalah utama dalam peringkasan dokumen. Hasil Ringkasan yang baik adalah ringkasan yang mampu mencakup (*coverage*) sebanyak mungkin konsep penting (*salient*) yang ada pada dokumen sumber (Ouyang, Li, Zhang, Li, & Lu, 2012). Hasil ringkasan harus mampu memilih kalimat-kalimat utama (penting) dan terhindar dari redundansi (*redundancy*) sehingga mampu mencakup banyak konsep.

Beberapa penelitian telah mengusulkan metode untuk mengatasi persoalan *coverage* dan *saliency*. Penelitian (Sarkar, 2009) menggunakan pendekatan *clustering* kalimat untuk menghasilkan cakupan ringkasan yang baik. Penelitian (He, Li, Shao, Chen,

& Ma, 2008) menggunakan fitur kepadatan informasi untuk menentukan kalimat yang menjadi representatif *cluster* untuk menyusun hasil ringkasan. Penelitian (Suputra, Arifin, & Yuniarti, 2013) mengkombinasikan metode pada penelitian (Sarkar, 2009) dan (He, Li, Shao, Chen, & Ma, 2008) untuk menghasilkan peringkasan multi dokumen yang lebih baik.

Clustering kalimat merupakan suatu metode yang mampu memberikan good *coverage* pada ringkasan (Sarkar, 2009). Pencapaian *good coverage* pada ringkasan tidak terlepas dari koherensi *cluster* yang baik. Salah satu metode yang dapat menjamin koherensi *cluster* adalah *SHC*. Metode tersebut pertama kali diusulkan oleh penelitian (Hammouda & Kamel, 2003) kemudian diterapkan pada penelitian (Sarkar, 2009) dan (Suputra, Arifin, & Yuniarti, 2013).

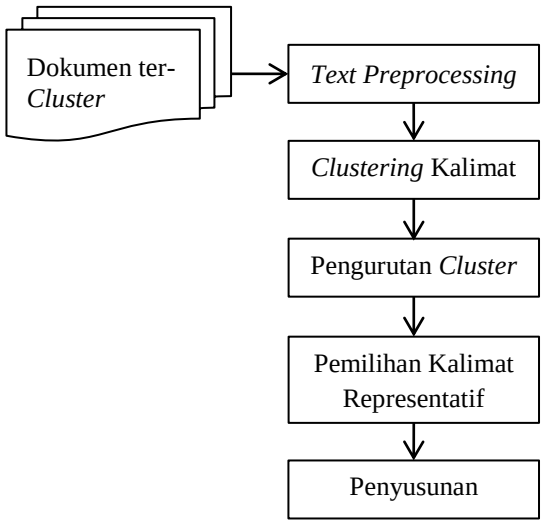
Strategi pemilihan kalimat representatif menjadi sangat penting untuk memecahkan masalah *saliency*. Kalimat yang terpilih harus mampu mewakili topik dari suatu *cluster* tertentu (Sarkar, 2009). Kalimat penting penyusun ringkasan harus memiliki kepadatan informasi yaitu mengandung informasi sebanyak mungkin dari dokumen sumber. Fitur *SID* merupakan sebuah usulan untuk menentukan tingkat kepadatan informasi tersebut (He, Li, Shao, Chen, & Ma, 2008).

Penelitian (Suputra, Arifin, & Yuniarti, 2013) mengatasi permasalahan *coverage* dan *saliency* sekaligus dengan mengusulkan sebuah *framework* untuk peringkasan multi dokumen. Pembentukan *cluster* kalimat dilakukan dengan *SHC*, sedangkan pemilihan kalimat represetatif *cluster* dilakukan dengan kombinasi fitur *SID* dan *SCK*.

Perhitungan similaritas kalimat dalam penentuan keanggotaan *cluster* yang dilakukan pada penelitian (Suputra, Arifin, & Yuniarti, 2013) menggunakan *Uni Gram Matching Based Similarity*. Metode tersebut mempertimbangkan kesamaan *term* yang dipakai secara bersama pada dua buah kalimat yang akan dihitung kemiripannya.

Pada kenyataannya, terdapat kemungkinan kemunculan kata-kata yang sinonim pada kedua kalimat yang dibandingkan. Salah satu contoh yang dapat menjelaskan konsep sinonim adalah kata *car* dan *automobile*. Kedua kata tersebut memiliki makna yang sama tetapi bentuk kata yang berbeda. Jika kedua kata tersebut diukur menggunakan metode usulan penelitian (Suputra, Arifin, & Yuniarti, 2013), maka nilai similaritasnya rendah, padahal keduanya memiliki hubungan semantik yang tinggi. Sehingga perhitungan *similarity* kalimat perlu mempertimbangkan kemungkinan hubungan semantik tersebut. Penelitian (Song & Park, 2009) mengusulkan penggunaan *LSI* untuk menemukan relasi semantik yang tersembunyi pada kasus *text clustering*.

Penelitian ini bertujuan untuk mengembangkan metode baru peringkasan multi dokumen dengan teknik *semantic clustering* dan pemilihan kalimat representatif *cluster*. Metode yang diusulkan berdasarkan prinsip kerja *LSI* dan *SHC* untuk pembentukan *cluster* kalimat secara semantik, serta mengkombinasikan fitur *SID* dan *SCK* untuk pemilihan kalimat representatif *cluster*.



Gambar 1. *Framework* Peringkasan Multi Dokumen.

2. METODE

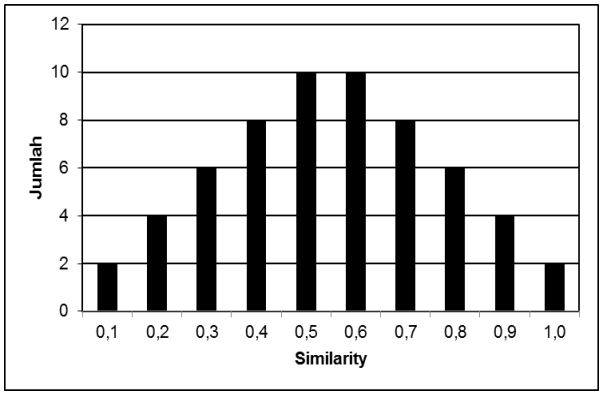
Metode yang digunakan pada penelitian ini menggunakan *framework* yang diadopsi dari penelitian (Sarkar, 2009) seperti yang ditunjukkan pada Gambar 1. Terdapat 5 tahapan yang digunakan, yaitu: *text preprocessing*, *clustering* kalimat, pengurutan *cluster*, pemilihan kalimat representatif, dan penyusunan ringkasan. Dokumen *input* yang akan diolah telah dikelompokkan ke dalam berbagai *cluster*.

2.1 *Text Preprocessing*

Text preprocessing merupakan tahapan pertama yang dilakukan sebelum *input* dokumen diolah lebih lanjut menjadi *cluster-cluster* kalimat. Proses-proses yang dilalui dalam tahap *text preprocessing* adalah *segmentation*, *stopword removal* dan *stemming*. Penelitian ini melakukan segmentasi terhadap kata dan kalimat. Setiap kata yang diperoleh dari hasil segmentasi menjadi *input* pada proses *stopword removal* untuk menghilangkan kata-kata yang tidak memiliki arti signifikan dalam proses pembentukan *cluster*. Pada bagian akhir dilakukan proses *stemming* dengan algoritma *Porter Stemmer* untuk mendapatkan bentuk kata dasar.

2.2 *Clustering* Kalimat

Hasil dari *text processing* akan menjadi *input* untuk melakukan pembentukan *cluster* kalimat. Bagian ini memiliki peran penting dalam sistem peringkasan otomatis. Setiap topik dalam dokumen



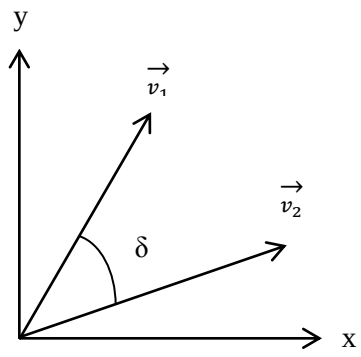
Gambar 2. *Cluster Similarity Histogram.*

harus diidentifikasi secara tepat untuk menemukan *similarity* dan *dissimilarity* yang ada dalam dokumen sehingga menjamin *good coverage* (Sarkar, 2009).

Koherensi *cluster* merupakan sebuah faktor yang sangat penting untuk menjamin kualitas hasil ringkasan. Koherensi *cluster* menunjukkan keterkaitan antar kalimat pada masing-masing *cluster* yang terbentuk dalam proses peringkasan multi dokumen. Derajat koherensi *cluster* yang tinggi sangat sulit dicapai karena memerlukan pengetahuan tentang makna dan struktur kalimat (Sarkar, 2009).

Penelitan ini menggunakan metode *SHC* yang mampu menjaga koherensi dari *cluster* yang terbentuk. Dalam *SHC* diperkenalkan konsep *cluster similarity histogram*. Konsep tersebut merupakan representasi statistik dari suatu distribusi *similarity* pasangan antar anggota yang ada pada suatu *cluster*. Jumlah dari bin dalam histogram menunjukkan interval nilai *similarity* tertentu (Sarkar, 2009; Hammouda & Kamel, 2003).

Derajat koherensi yang tinggi dalam sebuah *cluster* dapat dicapai dengan cara mempertahankan derajat *similarity* antar anggota tetap tinggi (Sarkar, 2009; Hammouda & Kamel, 2003). Dalam konsep *similarity histogram*, hal tersebut berarti menjaga distribusi *similarity* agar cenderung ke kanan. Ilustrasi konsep *SHC* tersebut dapat dilihat pada Gambar 2.



Gambar 3. Representase Vektor Pada Pengukuran *Cosine Similarity*.

Penentuan *similarity* antar kalimat dalam pembentukan *cluster* menggunakan prinsip kerja dari metode *LSI*, yang berfungsi untuk menemukan relasi semantik yang tersembunyi (Song & Park, 2009). *LSI* mengikuti logika bahwa kata-kata yang digunakan dalam konteks yang sama cenderung memiliki makna semantik yang sama. Salah satu sifat utama *LSI* adalah kemampuannya untuk membangun hubungan antara istilah yang muncul dalam konteks yang serupa.

Metode *LSI* membutuhkan proses *Singular Value Decomposition (SVD)* terhadap matriks yang dihasilkan dari *Vector Space Model (VSM)* antara *term* dan kalimat. Nilai *VSM* yang diwakili oleh matriks X akan didekomposisi menjadi tiga buah matriks U, ∑, dan V, sehingga memenuhi Persamaan (1). Matriks U dan V merupakan matriks *singular* kiri dan kanan, sedangankan matriks ∑ merupakan matriks diagonal yang menunjukkan nilai *singular*.

$$X = U.\sum.V^T. \quad (1)$$

Pengukuran nilai *similarity* yang digunakan adalah *cosine similarity*, yaitu perhitungan tingkat kemiripan berdasar pada besar sudut kosinus antara dua vektor seperti pada Gambar 3. Vektor kalimat pertama dinotasikan dengan vektor v_1 dan vektor kalimat kedua dinotasikan dengan v_2. Kedua vektor kalimat tersebut digambarkan pada bidang datar sumbu x dan sumbu y. Berdasarkan kosinus sudut antara dua vektor (δ), maka nilai *similarity* berkisar antara 0 sampai dengan 1. Nilai 0 menandakan bahwa kedua kalimat yang dibandingkan tidak mirip sama sekali, sedangkan nilai 1 menandakan bahwa kedua kalimat yang dibandingkan benar-benar identik.

Penentuan keanggotaan *cluster* kalimat harus

```
1: N ← Empty List {Cluster List}
2: for each sentence s do
3:   for each cluster c in N do
4:     HRold = HRC
5:     Simulate adding s to c
6:     HRnew = HRC
7:     if (HRnew ≥ HRold) OR
         ((HRnew ≥ HRmin)AND
         (HRold-HRnew < 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 N
16:  end if
17: end for
```

Gambar 4. *Pseudo Code* Algoritma SHC.

memenuhi nilai *threshold* dan kondisi *histogram ratio*. Penentuan keanggotaan *cluster* tersebut mengikuti algoritma *SHC* seperti pada Gambar 4.

2.3 Pengurutan *Cluster*

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).

Proses pembentukan *cluster* akan terus berlangsung mengikuti algoritma pada Gambar 4. Pada penelitian ini tidak digunakan parameter untuk menentukan jumlah *cluster* yang ideal. Seluruh *cluster* yang terbentuk hanya akan diurutkan dan dipilih sejumlah *cluster* yang terbaik.

Cluster importance merupakan 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. *Cluster importance* dihitung sesuai dengan Persamaan (2), di mana bobot dari *cluster c* ke-*j* dinotasikan dengan *Bobot*(c_j).

$$Bobot(c_j) = \sum_{w \in c_j} \log(1 + jumlah(w)). \quad (2)$$

Nilai *jumlah(w)* merupakan jumlah dari kata *w* pada koleksi *input* yang bernilai lebih dari atau sama dengan nilai *threshold θ*. Hasil perhitungan seluruh *cluster importance* akan diurutkan secara *descending* dan nantinya akan dipilih sejumlah *cluster* teratas.

2.4 Pemilihan Kalimat Representatif

Pemilihan kalimat yang menjadi kalimat ringkasan dalam penelitian ini didasarkan pada tingginya skor suatu kalimat di dalam suatu *cluster* tertentu. Penentuan skor kalimat dihitung berdasarkan kombinasi skor fitur *SID* dan *SCK* (Suputra, Arifin, & Yuniarti, 2013). Fitur *SID* menggunakan pendekatan *clustering* kalimat sehingga perhitungannya mengacu pada *positional text* (*sentence*) *graph* dalam suatu *cluster* kalimat (Suputra, Arifin, & Yuniarti, 2013). Sedangkan fitur *SCK* mengacu pada kata yang sering muncul dalam suatu *cluster* kalimat dan jarang/tidak muncul pada *cluster* kalimat lainnya. Fitur *SCK* diadopsi dari konsep *Term Frequency Inverse Document Frequency (TF.IDF)* (Suputra, Arifin, & Yuniarti, 2013). Perhitungan skor fitur *SID*, *SCK*, dan kombinasi keduanya, masing-masing ditunjukkan pada Persamaan (3), Persamaan (4), dan Persamaan (7).

$$F_{sid}(s_{kj}) = \frac{W_{s_{kj}}}{\max\limits_{l \in \{1,2,..,n\}} W_{s_{lj}}}. \quad (3)$$

Skor fitur *SID* dinotasikan dengan $F_{sid}(s_{kj})$ merupakan hasil bagi antara penjumlahan bobot dari semua *edge* yang datang dari kalimat *s* ke-*k* pada *cluster* ke-*j* (dinotasikan dengan $W_{s_{kj}}$) dengan bobot *edge* maksimum diantara semua kalimat yang ada pada *cluster* ke-*j* (dinotasikan dengan $\max_{l \in \{1,2,..,n\}} W_{s_{lj}}$).

$$F_{sck}(s_{kj}) = \frac{1}{len(s_{kj})} \sum_{w_{ij} \in s_{kj}} tf\_iscf_{w_{ij.}} \quad (4)$$

Skor fitur *SCK* untuk setiap kata penyusun kalimat s ke-*k* yang berada pada *cluster* kalimat ke-*j* dinotasikan dengan $F_{sck}(s_{kj})$. Nilai skor fitur *SCK* merupakan perkalian antara panjang kalimat s ke-*k* pada *cluster* kalimat ke-*j* (dinotasikan dengan $len(s_{kj})$) dan penjumlahan bobot kata-kata kunci yang didapat dari perhitungan *tf_iscf* pada setiap kalimat. Nilai *tf_iscf* dapat diketahui berdasarkan Persamaan (5).

$$tf\_iscf_{w_{ij}} = \frac{tf_{w_{ij}} * iscf_{w_{ij}}}{\sqrt{\sum_{i=1}^{M} (tf_{w_{ij}} * iscf_{w_{ij}})^2}}. \quad (5)$$

Pada Persamaan (5) nilai *tf_iscf* harus memperhitungkan jumlah kemunculan *term w* ke-*i* pada *cluster* ke-*j* (dinotasikan dengan $tf_{w_{ij}}$) serta *inverse sentence cluster frequency term w* ke-*i* pada *cluster* ke-*j* (dinotasikan dengan $iscf_{w_{ij}}$). Nilai $iscf_{w_{ij}}$ dapat dihitung berdasarkan Persamaan (6).

$$iscf_{w_{ij}} = \log(\frac{N}{scf_{w_{ij}}}). \quad (6)$$

Nilai $iscf_{w_{ij}}$ merupakan hasil logaritma antara jumlah dokumen N dengan jumlah *cluster* kalimat yang mengandung kata *w* ke-*i* pada *cluster* kalimat ke-*j* (dinotasikan dengan $scf_{w_{ij}}$).

$$Skor_{Komb}(S_{kj}) = \lambda . F_{sid}(S_{kj}) + (1 - \lambda) . F_{sck}(S_{kj}) \quad (7)$$

Nilai skor kombinasi antara fitur *SID* dan fitur *SCK* dinotasikan dengan $Skor_{Komb}(s_{kj})$ dan dihitung berdasarkan Persamaan (7). Pada perhitungan skor kombinasi tersebut, notasi λ menunjukkan nilai bobot untuk fitur *SID*, sedangkan

nila bobot untuk fitur *SCK* adalah 1-λ. Nilai parameter λ berada pada rentang 0 hingga 1, dan diatur sesuai pembagian persentase bobot antara fitur *SID* dan *SCK*. Semakin besar nilai λ menunjukkan bahwa bobot penggunaan fitur *SID* lebih dominan dibandingkan dengan fitur *SCK*.

2.5 Penyusunan Ringkasan

Penyusunan hasil ringkasan pada penelitian ini dilakukan dengan cara melakukan pemilihan kalimat representatif berdasarkan bobot *cluster importance* yang paling tinggi. Kemudian pemilihan dilanjutkan pada *cluster* berikutnya sesuai dengan daftar urutan *cluster*. Pemilihan kalimat tersebut terus dilakukan hingga panjang ringkasan yang diharapkan terpenuhi.

2.6 Pengukuran Hasil Ringkasan

Pengukuran hasil peringkasan otomatis yang digunakan pada penelitian ini adalah *ROUGE*. Metode ini mengukur kualitas hasil ringkasan berdasarkan kesesuaian antara unit-unit ringkasan hasil sistem dengan unit-unit ringkasan referensi yang dibuat secara manual. Pada penelitian 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 (Lin, 2004) ditunjukkan pada Persamaan (8).

$$ROUGE - N = \frac{\sum_{S \in Summ_{ref}} \sum_{N-gram \in S} Count_{match}(N-gram)}{\sum_{S \in Summ_{ref}} \sum_{N-gram \in S} Count(N-gram)} \quad (8)$$

Pada Persamaan (8), notasi 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 dan N=2. *ROUGE* dengan nilai N=1 berarti membandingkan kesamaan hasil ringkasan dengan ringkasan referensi untuk setiap satu kata. *ROUGE* dengan nilai N=2 berarti membandingkan kesamaan hasil ringkasan dengan ringkasan referensi untuk setiap dua kata.

Berdasarkan karakteristik *dataset DUC 2004 Task 2* yang menggunakan *multiple references summaries* (banyak referensi ringkasan) yaitu empat referensi per-*cluster* dokumen, maka perhitungan nilai *ROUGE-N* akhir dihitung berdasarkan persamaan (9). 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 *rsi*. Perhitungan *ROUGE-N* tersebut diadopsi dari (Lin, 2004) dan ditunjukkan pada Persamaan (9).

$$ROUGE - N_{multi} = argmax_i ROUGE - N(sc, rs_i) \quad (9)$$

3. HASIL UJI COBA

Pengujian dilakukan pada *dataset DUC 2004 Task 2* yang dapat diunduh pada alamat http://duc.nist.gov/duc2004/tasks.html. *Dataset DUC 2004 Task 2* merupakan kumpulan dokumen berita dalam bahasa Inggris dari *Associated Press* dan *New York Times*. Dokumen-dokumen tersebut telah terbagi ke dalam kelompok-kelompok menjadi 50 *cluster* dokumen. Setiap *cluster* dokumen terdiri dari rata-rata 10 dokumen berita.

Proses pengujian terdiri atas *training* dan *testing*. Proses *training* dilakukan untuk mengetahui kombinasi parameter yang optimal. Parameter-parameter yang diperhitungkan beserta nilai optimalnya ditunjukkan pada Tabel 1.

Proses *testing* dilakukan untuk menguji kualitas hasil ringkasan berdasarkan kombinasi nilai parameter optimal yang telah dihasilkan pada proses *training*. Tabel 2 menunjukkan perbandingan hasil pengukuran *ROUGE-1* dan *ROUGE-2* rata-rata terhadap metode peringkasan multi dokumen yang dikembangkan oleh (Suputra, Arifin, & Yuniarti, 2013), metode *Local Importance Global Importance (LIGI)*, dan metode yang diusulkan *(Semantic Clustering)*. Hasil pengujian pada Tabel 2 merupakan rata-rata nilai *ROUGE* dengan parameter λ diatur secara bertahap dari 0 sampai 1.

Pada penelitian ini, juga didapatkan hasil bahwa penggunaan jumlah elemen diagonal *k* pada matriks *singular* $\sum$ memberikan pengaruh pada hasil pengujian nilai *ROUGE-1* dan *ROUGE-2*. Pengujian dilakukan pada beberapa dokumen dengan nilai *k*=75, *k*=100, *k*=125, dan *k*=150. Hasil

Tabel 1. Kombinasi Parameter Uji Optimal

Parameter	Keterangan	Nilai Optimal
HR_{min}	Batas nilai minimum *Histogram Ratio*	0,7
ε	Batas selisih maksimum antara HR_{old} dengan HR_{new}	0,3
S_T	Batas *similarity bin* pada perhitungan *histogram ratio*	0,5
θ	Batas frekuensi minimal kata *w* dalam proses pengurutan *cluster*	10
α	Nilai *threshold* untuk menentukan pembentukkan *edge* antar kalimat pada fitur *SID*	0,4
λ	Bobot untuk fitur *SID* dan fitur *SCK*	0-1

Tabel 2.Hasil Pengukuran *ROUGE-1* dan *ROUGE-2*

Metode	**ROUGE-1**	**ROUGE-2**
Suputra	0,390	0,104
LIGI	0,374	0,096
Semantic Clustering	0,395	0,106

pengujian rata-rata pada beberapa dokumen dengan nilai *k* yang beragam ditunjukkan pada Gambar 5.

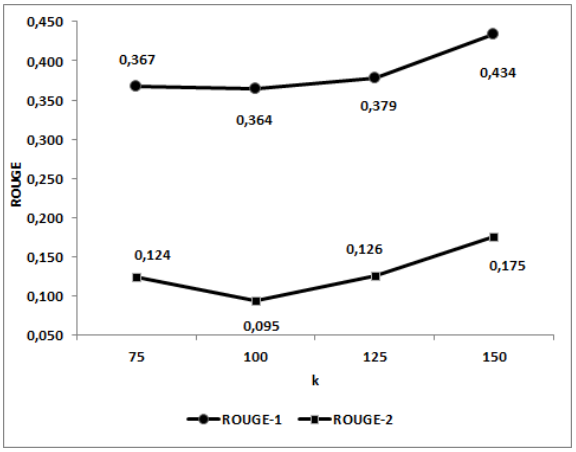
4. PEMBAHASAN HASIL UJI COBA

Hasil pengujian yang dilakukan pada proses *training* menghasilkan kombinasi parameter yang optimal, masing-masing HR_{min}=0,7, ε=0,3, S_T=0,5, θ=10, dan α=0,4. Sedangkan nilai λ diujikan pada nilai 0 sampai dengan 1 terhadap kombinasi parameter optimal tersebut. Kombinasi parameter tersebut memberikan nilai *ROUGE-1* dan *ROUGE-2* tertinggi untuk peringkasan multi dokumen pada penelitian ini.

Pengujian pada proses *testing* dengan menggunakan kombinasi parameter optimal, menunjukkan bahwa metode *Semantic Clustering* memiliki nilai *ROUGE-1* lebih tinggi dari pada metode lainnya. Hasil yang dicapai oleh metode *Semantic Clustering* memiliki selisih sebesar 0,00446 dengan metode Suputra atau meningkat sebesar 1,14%. Hasil tersebut juga memiliki selisih sebesar 0,02056 dengan metode *LIGI* atau meningkat sebesar 5,49%. Perbandingan hasil yang diperoleh menunjukkan bahwa metode yang diusulkan mampu melakukan perbaikan pada metode terdahulu untuk pengukuran kualitas dengan *ROUGE-1*.

Pengukuran dengan menggunakan *ROUGE-2* juga menunjukkan keunggulan metode *Semantic Clustering* dibanding dengan metode lain. Hasil yang dicapai oleh metode *Semantic Clustering* memiliki selisih sebesar 0,00205 dengan metode Suputra atau meningkat sebesar 1,97%. Hasil tersebut juga memiliki selisih sebesar 0,00955 dengan metode *LIGI* atau meningkat sebesar 9,89%.

Hasil pengujian menunjukkan bahwa



Gambar 5. Pengaruh Nilai *k* Terhadap Nilai *ROUGE*.

pengukuran kualitas ringkasan yang hanya membandingkan satu kata saja *(uni gram)* memiliki nilai lebih tinggi jika membandingkan dua kata *(bigram).* Hal tersebut ditunjukkan dengan nilai *ROUGE-1* rata-rata yang lebih tinggi dari nilai *ROUGE-2* rata-rata. Hasil ini dapat dengan mudah dipahami, misalnya dengan contoh bahwa lebih banyak hasil yang sama jika menggunakan kata *party* atau *leaders* saja dibanding menggunakan gabungan kata *party leaders.*

Keunggulan nilai yang diperoleh pada metode *Semantic Clustering* disebabkan oleh kemampuan metode tersebut untuk menemukan relasi semantik pada pembuatan *cluster* kalimat. Kemampuan itu dimungkinkan dengan penggunaan *LSI* sebelum mengukur nilai similaritas kalimat. Pada metode Suputra, nilai similaritas diukur dengan menggunakan *Uni Gram Matching Based Similarity*. Metode pengukuran tersebut hanya mempertimbangkan jumlah kata yang dipakai bersama oleh kedua kalimat yang dibandingkan. Sedangkan dalam kenyataannya, banyak kata yang memiliki makna yang sama tetapi bentuk yang berbeda (sinonim). Salah satu contoh yang dapat menjelaskan konsep sinonim adalah kata *car* dan *automobile*. Kedua kata tersebut memiliki makna yang sama tetapi bentuk kata yang berbeda. Jika kedua kata tersebut diukur menggunakan metode *Uni Gram Matching Based Similarity*, maka nilai similaritasnya rendah. Akan tetapi jika diukur menggunakan konsep *semantic similarity*, maka similaritasnya tinggi.

Penggunaan *LSI* pada proses pembentukan *cluster* kalimat mampu menemukan kata-kata yang sering dipakai pada konsep yang sama. Dengan demikian hasil ringkasan yang diperoleh lebih mendekati hasil peringkasan manusia. Kedekatan tersebut ditunjukkan oleh nilai *ROUGE* yang lebih tinggi.

Pada penelitian ini juga didapatkan hasil bahwa penggunaan nilai bobot pada fitur *SID* dan *SCK* tidak mempengaruhi secara signifikan kualitas hasil ringkasan. Hal tersebut ditunjukkan dengan nilai *ROUGE* yang cenderung tetap walaupun menggunakan nilai parameter λ yang bervariasi dari 0 hingga 1. Metode *Semantic Clustering* memungkinkan pemberian nilai bobot untuk fitur *SID* dan *SCK* yang sama untuk keduanya.

Penggunaan jumlah elemen diagonal *k* pada matriks *singular* $\sum$ juga memperlihatkan adanya pengaruh pada nilai *ROUGE-1* dan *ROUGE-2*. Nilai *ROUGE-1* dan *ROUGE-2* mengalami penurunan pada penggunaan nilai *k*=75 hingga *k*=100, kemudian mengalami kenaikan pada penggunaan nilai *k*=100 hingga *k*=150. Nilai *k* menentukan jumlah elemen *singular* yang digunakan untuk proses perhitungan *similarity* kalimat pada tahapan pembentukan *cluster* kalimat. Pengurangan nilai *k* dapat mereduksi dimensi data yang diolah, namun pada jumlah tertentu justru menghilangkan informasi

penting yang terkandung pada dokumen yang diolah. Berdasarkan temuan tersebut, nilai *k* perlu diatur sedemikian agar mampu menghasilkan hasil peringkasan dengan nilai *ROUGE* yang tinggi walaupun terjadi reduksi dimensi data.

5. KESIMPULAN

Teknik *semantic clustering* dan pemilihan kalimat representatif *cluster* dapat diimplementasikan pada peringkasan multi dokumen. Implementasi tersebut dilakukan dengan memanfaatkan prinsip kerja *LSI* dan *SHC* untuk pembentukan *cluster* kalimat secara semantik, serta mengkombinasikan fitur *SID* dan *SCK* untuk pemilihan kalimat representatif *cluster*.

Hasil pengujian metode yang diusulkan mampu mencapai nilai *ROUGE-1* sebesar 0,395 dan *ROUGE-2* sebesar 0,106. Nilai *ROUGE-1* metode yang diusulkan lebih tinggi 1,14% dari metode Suputra dan 5,49% dari metode *LIGI*. Sedangkan nilai *ROUGE-2* metode yang diusulkan lebih tinggi 1,97% dari metode Suputra dan 9,89% dari metode *LIGI*. Dengan demikian metode yang diusulkan layak untuk diimplementasikan pada peringkasan multi dokumen.

Metode yang diusulkan mampu menemukan hubungan semantik tersembunyi antar kata. Hal tersebut dapat meningkatkan koherensi *cluster* kalimat. Namun demikian, metode *LSI* yang digunakan untuk tujuan tersebut menggunakan bantuan dekomposisi matriks *SVD*. Pada proses dekomposisi *SVD*, jumlah elemen diagonal *k* matriks *singular* $\sum$ memberikan pengaruh pada nilai pengujian *ROUGE-1* dan *ROUGE-2*. Pengurangan nilai *k* tersebut dapat mereduksi dimensi data, namun pada nilai tertentu menurunkan hasil pengujian *ROUGE-1* dan *ROUGE-2*. Sehingga penelitian ini dapat dikembangkan lebih lanjut untuk menentukan nilai *k* yang paling optimal untuk digunakan agar hasil pengujian *ROUGE-1* dan *ROUGE-2* berada pada nilai tertinggi.

6. DAFTAR PUSTAKA

GUPTA, V. & Lehal, G. S. 2010. A Survey of Text Summarization Extractive Techniques. *Journal of Emerging Technologies in Web Intelligence*, vol. II, no. 3, pp. 258-268.

HAMMOUDA, K. M. & KAMEL, M. S. 2003. Incremental Document *Cluster*ing Using *Cluster* Similarity Histograms. *Proceeding of the 2003 IEEE/WIC International Conference on Web Intelligence*.

HE, T., LI, F., SHAO, W., CHEN, J. & L. MA, L. 2008. A New Feature-Fusion Sentence Selecting Strategy for Query-Focused Multi-Document Summarization. *International Conference on Advanced Language Processing and Web Information Technology*.

KOGILAVANI, A. & BALASUBRAMANI, P. 2010. *Cluster*ing and Feature Spesific Sentence Extraction Based Summarization of Multiple Documents. *International journal of computer science & information Technology*, vol. II, no. 4.

LIN, C.Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. *Proceedings of Workshop on Text Summarization Brances Out*.

OUYANG, Y., Li, W., Zhang, R. , Li, S. & Lu, Q. 2012. A Progressive Sentence Selection Strategy for Document Summarization. *Information Processing and Management*.

SARKAR, K. 2009. Sentence *Cluster*ing-based Summarization of Multiple Text Documents. *International Journal of Computing Science and Communication Technologies*, vol. II, no. 1.

SONG, W. & PARK, S.C., 2004. Genetic Algorithm for Text *Cluster*ing Based on Latent Semantic Indexing. *Computers and Mathematics with Applications*, pp. 1901-1907.

SUPUTRA , I.P.G., H., ARIFIN, A. Z. & YUNIARTI, A. 2013. Pendekatan Positional Text Graph Untuk Pemilihan Kalimat Representatif *Cluster* Pada Peringkasan Multi-Dokumen. *Jurnal Ilmu Komputer*, vol. IV, no. 2.