Jurnal Linguistik Komputasional
Indonesian Association for Computational Linguistics
Vol 1, No 2
September
2018
e-ISSN :
2621-9336

Pembobotan Kata berdasarkan Kluster untuk Peringkasan Otomatis Multi Dokumen

Lukman Hakim[#1], Fadli Husein Wattiheluw[#2], Agus Zainal Arifin[#3], Aminul Wahib[#4]

[#]*Departement of Informatics, Institut Teknologi Sepuluh Nopember*
Surabaya, Indonesia
[1]lukman.its.tif2017@gmail.com
[2]fadliwattiheluw1994@gmail.com
[3]*agusza@cs.its.ac.id*
[4]wahib13@mhs.if.its.ac.id

Abstract*— *Multi-document summarization is a technique for getting information. The information consists of several lines of sentences that aim to describe the contents of the entire document relevantly. Several algorithms with various criteria have been carried out. In general, these criteria are the preprocessing, cluster, and representative sentence selection to produce summaries that have high relevance. In some conditions, the cluster stage is one of the important stages to produce summarization. Existing research cannot determine the number of clusters to be formed. Therefore, we propose clustering techniques using cluster hierarchy. This technique measures the similarity between sentences using cosine similarity. These sentences are clustered based on their similarity values. Clusters that have the highest level of similarity with other clusters will be merged into one cluster. This merger process will continue until one cluster remains. Experimental results on the 2004 Document Understanding Document (DUC) dataset and using two scenarios that use 132, 135, 137 and 140 clusters resulting in fluctuating values. The smaller the number of clusters does not guarantee an increase in the value of ROUGE-1. The method proposed using the same number of clusters has a lower ROUGE-1 value than the previous method. This is because in cluster 140 the similarity values in each cluster experienced a decrease in similarity values.

Keywords*— *cluster, cosine similarity, multi-document, summarization

Abstrak*— Peringkasan multi-dokumen merupakan teknik untuk mendapatkan informasi. Informasi tersebut terdiri dari beberapa baris kalimat yang bertujuan untuk menggambarkan isi dari keseluruhan dokumen secara relevan. Beberapa algoritma dengan berbagai macam kriteria telah dilakukan. Secara umum, kriteria tersebut yaitu tahap praproses, kluster, dan pemilihan kalimat yang representatif untuk menghasilkan ringkasan yang memiliki relevansi tinggi. Dalam beberapa kondisi, tahap kluster merupakan salah satu tahap yang penting untuk menghasilkan peringkasan. Penelitian yang ada tidak bisa menentukan jumlah kluster yang akan dibentuk. Oleh karena itu, kami mengusulkan teknik klusterisasi menggunaan hiraki kluster. Teknik ini mengukur kemiripan antar kalimat menggunakan cosine similarity. Kalimat-kalimat tersebut dikluster berdasarkan nilai kemiripannya. Kluster yang memiliki tingkat kemiripan tertinggi dengan kluster lain akan digabung menjadi satu kluster. Proses penggabungan ini akan terus dilakukan sampai tersisa satu kluster. Hasil eksperimen pada dataset Document Understanding Document (DUC) 2004 dan menggunakan dua skenario yaitu menggunaan 132, 135, 137 dan 140 kluster menghasilkan nilai yang fluktuatif. Semakin kecil jumlah kluster tidak menjamin peningkatan nilai ROUGE-1. Metode yang diusulkan menggunakan jumlah kluster yang sama memiliki nilai ROUGE-1 lebih rendah dibandingkan metode sebelumnya. Hal ini dikarenakan pada kluster 140 nilai similarity pada masing-masing kluster banyak yang mengalami penurunan nilai similarity.

***Kata kunci*— kluster, cosine similarity, multi-dokumen, peringkasan**

I. PENDAHULUAN

Beberapa topik dalam dokumen memiliki jumlah teks dan informasi yang panjang. Informasi tersebut bisa diperoleh dengan melakukan pencarian dari satu atau beberapa dokumen. Peningkatan jumlah dokumen yang sangat pesat menimbulkan sebuah permasalahan yaitu bagaimana menemukan informasi yang berguna secara cepat dan akurat. Oleh karena itu, diperlukan solusi untuk memecahkan permasalahan tersebut. Baru-baru ini, salah satu solusi yang diakui untuk menemukan informasi yang berguna adalah peringkasan dokumen. Peringkasan dokumen adalah proses untuk mendapatkan informasi dari satu atau beberapa dokumen yang berisi kalimat yang pangjang disaring menjadi beberapa baris kalimat tanpa mengurangi relevansi isi dari keseluruhan dokumen [1]. Melakukan peringkasan dapat memberikan pengetahuan kepada pengguna tentang permasalahan utama dari keseluruhan teks pada dokumen. Dengan demikian, pengguna bisa menghemat waktu untuk memahami esensi diskusi tanpa harus membaca teks secara menyeluruh [2].

Metode peringkasan teks dapat dibagi menjadi dua jenis, yaitu: (1) Metode ekstraktif, dan (2) Metode abstraktif. Metode ekstraktif merupakan teks pertama

yang dianggap sebagai sumber yang mewakili topik utama teks. Sedangkan metode abstraktif merupakan teks sumber kedua yang dapat menghasilkan teks baru. Selanjutnya, dokumen sumber yang akan diringkas tergantung jumlah dari dokumen, ringkasan teks dapat berupa ringkasan dokumen tunggal atau ringkasan multi dokumen. Ringkasan dokumen langsung menghasilkan ringkasan singkat dari beberapa dokumen dalam satu set [3]. Tantangan ringkasan multi dokumen untuk mengekstrak kalimat penting lebih tinggi dari pada ringkasan dokumen tunggal karena memiliki ruang pencarian yang lebih besar dibandingkan dengan kumpulan dokumen tunggal [2].

Beberapa penelitian tentang peringkasan multi dokumen telah dipelajari untuk menghasilkan ringkasan optimal berdasarkan metode peringkasan abstraktif. Beberapa diantaranya menggunakan algoritma pengoptimalan seperti Cuckoo Search [2], Cat Swarm [3], Local Sentence Spread [4].

Beberapa kriteria yang dipertimbangkan oleh algoritma optimasi adalah menghasilkan ringkasan yang lebih baik dengan melakukan pengelompokan menggunakan kemiripan berdasarkan Kluster Histogram dan membandingkan kesamaan masing-masing kluster. Kemudian mempertimbangkan kata-kata yang tersebar di seluruh kluster dan memberi bobot pada masing-masing kluster. Tujuannya untuk mengatasi masalah kata dengan jumlah frekuensi yang sama dalam dokumen tapi memiliki nilai yang sangat berbeda sesuai dengan topik utama keseluruhan dokumen yang telah diringkas.

Namun, dalam kondisi tertentu, dokumen lebih dari satu dan terdiri dari banyak kategori yang perlu dikelompokkan terlebih dahulu. Perumusan teks bisa diimplementasikan dalam proses pengelompokan dokumen. Akibatnya, penelitian sebelumnya pada bagian kluster tidak dapat menentukan jumlah kluster yang akan dibentuk sehingga menambahkan fungsi Cluster Importance untuk mengurutkan kluster secara descending [4].

Oleh karena itu, kami mengusulkan sebuah analisis kluster berbasis Hirarki. Prosesnya yaitu membandingkan kluster-kluster secara berpasangan dengan kluster-kluster yang berdekatan, dan menggabungkan kluster dengan kluster lain yang memiliki nilai kemiripan tertinggi untuk menghasilkan sejumlah kecil kluster. Prosesnya yaitu membandingkan kemiripan antar kluster yaitu kluster pertama dengan kluster kedua, kluster kedua dengan kluster ketiga, kluster ketiga dengan kluster keempat dan seterusnya [5]. Proses pengklasteran akan berhenti sampai jumlah kluster sama dengan satu. Selanjutnya, menentukan berapa kluster yang akan digunakan. Sedangkan untuk memperoleh nilai kemiripan antar kluster pada peringkasan multi dokumen menggunakan metode Cosine Measure.

II. PERINGKASAN DOKUMEN

Peringkasan multi-dokumen merupakan ringkasan singkat dari beberapa dokumen tanpa kehilangan informasi yang berguna. Ada lima langkah utama seperti tahap input dokumen, tahap pra-proses, tahap kluster, tahap pengoptimalan ringkasan, dan rangkuman akhir. Kerangka umum yang diusulkan dijelaskan pada Gambar 3. Beberapa dokumen dengan topik yang berbeda diberikan sebagai masukan pada metode yang diusulkan. Kemudian, dokumen dikelompokkan berdasarkan topik. Setelah itu, hasilnya akan melalui tahap pra-proses seperti tokenisasi, stopword, dan stemming. Akhirnya, optimasi ringkasan diterapkan untuk mengekstrak ringkasan akhir.

A. Peringkasan Dokumen Tunggal vs Multidokumen

Metode peringkasan dokumen dibedakan berdasarkan jumlah dokumen yang diringkas. Penelitian awal, dokumen sumber yang akan diringkas hanya dokumen tunggal. Pada awal perkembangannya, metode yang paling sering dipakai adalah dengan memanfaatkan fitur dokumen untuk menghitung significant value sebagai acuan pemilihan kalimat ringkasan. Beberapa fitur yang digunakan dintaranya word frequency, posisi kalimat, cue word, dan kategori kerangka dokumen [6]–[8]. Metode yang lebih maju dengan menggunakan machine learning untuk memilih kalimat, diantaranya dengan menggunakna naive bayes classifier dan decision tree [9][10].

Metode peringkasan multidokumen memiliki perbedaan dengan peringkasan dokumen tunggal dikarenakan adanya informasi yang sama, saling melengkapi, atau berlawanan. Informasi tersebut tersebar pada dokumen-dokumen yang akan diringkas. Oleh karena itu, tujuan dari peringkasan dokumen secara otomatis tidak hanya menghindari pengulangan informasi dalam ringkasan namun juga mengenali informasi baru dan memastikan hasil akhir ringkasan lengkap dan koheren [11].

McKeown dan Radev [12] mengusulkan metode peringkasan multi dokumen dengan konsep clustering kalimat dengan memanfaatkan similaritas antar kalimat dari beberapa dokumen yang akan diringkas. Carbonel dan Goldstein [13] mengusulkan metode dengan memanfaatkan Maximal Marginal Relevance (MMR) untuk menilai kebaruan informasi pada suatu kalimat. Meskipun peringkasan otomatis dokumen telah banyak dikembangkan.

B. Metode Peringkasan Ekstraktif vs Abstraktif

Pada metode peringkasan ekstraktif, kalimat dalam hasil ringkasan adalah sebagian kalimat dari dokumen yang akan diringkas. Metode ekstraktif dilakukan dengan memilih kalimat dari dokumen asal untuk dimasukkan ke dalam ringkasan. Sedangkan pada metode peringkasan abstraktif, kalimat ringkasan dibuat dari nol berdasarkan informasi yang didapat dari dokumen yang diringkas (dengan memanfaatkan metode pembentukan kalimat).

Meskipun pada awalnya metode ekstraktif dianggap kurang baik untuk peringkasan multi dokumen, anggapan ini berubah ketika sistem peringkasan MEAD yang dikembangkan Radev et.al [14] memberikan performa yang baik dengan metode ekstraktif. Metode abstraktif akan memberikan ringkasan dalam susunan yang lebih baik, namun sangat sulit diimplementasikan tanpa tersedianya metode pembentukan kalimat yang baik [9].

C. Kata Pencarian dalam Peringkasan Dokumen

Artikel (selain artikel tanya jawab) biasanya ditulis untuk memberikan sejumlah informasi kepada pembaca. Informasi tersebut biasanya tidak memperhatikan informasi yang berguna bagi pembaca. Kata pencarian mempengaruhi metode peringkasan. Pengaruh tersebut tergantung pada tujuan suatu sistem peringkasan yang akan dibuat. Dengan tidak adanya kata pencarian, ringkasan akan merefleksikan informasi yang ingin disampaikan penulis dari dokumen yang akan diringkas [11]. Kata pencarian menjadi penting dalam metode peringkasan ketika ringkasan perlu disajikan sebagai jawaban terhadap suatu pertanyaan oleh pencari informasi.

$$tf = (t_i, D) \quad (1)$$

$$idf(t_i) = \log \frac{N}{df(t_i, D)} + 1 \quad (2)$$

$$w_D(t_i) = \frac{tf(t_i, D) \times log \frac{N}{df(t_i)} + 1}{\sqrt{\sum_{t_i}(tf(t_i, D) \times log \frac{N}{df(t_i)} + 1)^2}} \quad (3)$$

$$similarity(Q, D) = \sum_{r=1}^{M} w_Q(t_i) \times w_D(t_i) \quad (4)$$

PERSAMAAN 1, dimana tf_i merupakan frekuensi kemunculan term t_i, sedangkan d_i merupakan jumlah frekuensi seluruh kata pada dokumen.

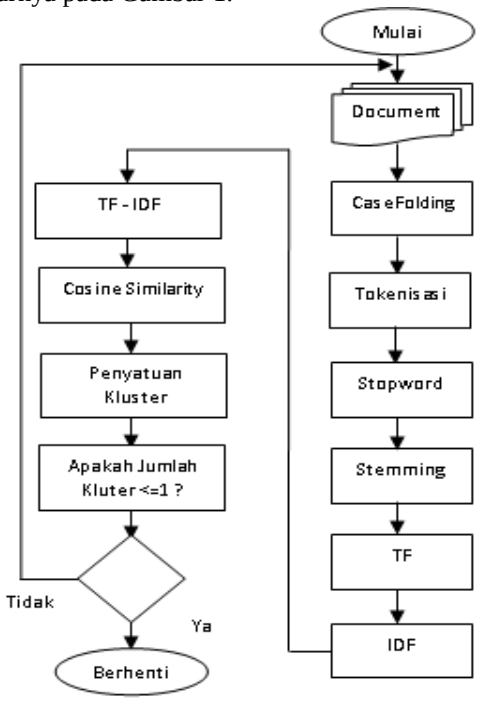
PERSAMAAN 2, dimana n_i dokumen yang mengandung term t_i dalam sejumlah koleksi N dokumen. Term Frekuensi saja tidak cukup untuk memberikan indikasi bahwa term yang dihasilkan memiliki posisi yang sesuai dengan dokumen atau query teks. Ini dikarenakan ada term yang memiliki frekuensi tinggi terkonsentrasi pada satu atau beberapa dokumen, tidak ditemukan di beberapa dokumen lain, itu akan mempengaruhi keakuratan hasil pencarian. Oleh karena itu IDF memperhitungkan faktor-faktor yang melibatkan penyebaran istilah dalam satu set dokumen.

PERSAMAAN 3, W_{ij} merupakan bobot dari word / term t_j terhadap dokumen ke d_i, tf_{ij} merupakan jumlah word / term kemunculan t_j pada tf_{ij}, N jumlah semua dokumen dalam database, n jumlah dokumen yang mengandung word / term.

PERSAMAAN 4, untuk mengukur kemiripan antar kluster yaitu dengan menggunakan metode cosine measure.

III. METODOLOGI

Penelitian ini mengusulkan metode hirarki kluster yang diterapkan pada tahap kluster. Secara menyeluruh terdapat beberapa tahapan proses yang dilakukan pada penelitian ini dan dapat dilihat alurnya secara menyeluruh pada Gambar 3. Sedangkan kontribusi pada penelitian ini dapat dilihat alurnya pada Gambar 1.



Gambar 1 Proses pembentukan cluster

A. Preprocessing

Preprocessing digunakan untuk menghitung bobot setiap kalimat dengan menggunakan hasil proses menjadi beberapa kata. Tahapan *preprocessing* ini terdiri dari 4 (empat) tahap yaitu casefolding, tokenisasi, stopword, dan stemming.

a. Casefolding

Merubah semua huruf menjadi huruf kecil (*lowercase*).
Contoh:
Input: Software Engineering and Civil Engineering.
Output: software engineering and civil engineering.

b. Tokenisasi

Pada tahap ini, input kalimat kebutuhan dipecah menjadi unit terkecil.
Contoh:
Input: software engineering and civil engineering.
Output: "software", "engineering", "and", "civil", "engineering".

c. Stopword

Stopwords memiliki peran untuk menghapus kata yang dianggap tidak penting.
Contoh:
Input: "software","engineering","and","civil", "engineering".
Output: "software","engineering","civil", "engineering".

d. Stemming

Stemming memiliki peran untuk menjadikan teks menjadi kata dasar. Proses kata stemming dalam bahasa inggris memiliki karakteristik tersendiri dengan menggunakan algoritma stemming Porter yang tidak lepas dari pengaruh tata bahasanya.
Contoh:
"*maintainable*" akan diubah menjadi "*maintan*".

B. Sinonim

Perluasan kesamaan kata atau sinonim dilakukan pada term-term yang ada pada dokumen. Pertimbangan utamanya adalah kata-kata yang memiliki makna yang sama dapat dinyatakan dalam satu atau lebih kata yang berbeda dalam suatu kalimat kebutuhan. Sebagai contoh term pada dokumen yaitu *accident* memiliki sinonim `*disaster*`, `*collision*`, `*blow*`,`*fluk*`' yang terdapat dalam dokumen yang dianalisis.

C. Post Tagging

Pada penelitian ini melakukan penentuan part of speech terbaik untuk mencari sinonim dari kata tersebut. POST Tagging ini sangat penting karena kata yang sama dengan part of speech yang berbeda akan menghasilkan daftar sinonim yang relatif berbeda.
Contoh:
N: Noun, V: Verb, R: Adverb, and A: Adjective.

D. Term Weighting

Pada tahapan term weighting ini kami mengajukan beberapa metode pembobotan yaitu:

a. TF (*Term Frequency*)
Term Frequency adalah proses untuk menghitung jumlah kata yang muncul dalam satu kalimat.

b. IDF (Inverse Document Frequency)
Inverse Document Frequency adalah proses untuk menghitung jumlah kata yang muncul pada kedua dokumen.

c. TF-IDF
Metode TF-IDF merupakan cara untuk memberikan bobot hubungan ke dokumen. Metode ini menggabungkan dua konsep untuk perhitungan bobot yaitu TF dan IDF. Frekuensi kemunculan kata pada dokumen yang diberikan menunjukkan betapa pentingnya kata itu dalam dokumen. Jumlah frekuensi dokumen yang mengandung kata menunjukkan seberapa umum kata itu. Jadi bobot hubungan antara kata dan dokumen akan tinggi jika frekuensi kata yang ada dalam dokumen tinggi dan akan rendah jika frekuensi kata yang ada dalam dokumen rendah.

E. Similarity

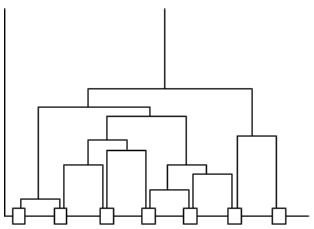
Cosine similarity adalah proses perhitungan untuk menemukan kesamaan antar kalimat. Penyatuan kluster adalah proses penggabungan cluster yang memiliki tingkat kemiripan tertinggi diantara kluster lainnya. Penyatuan kluster tersebut akan berhenti sampai jumlah kluster sama dengan satu. Selanjutnya, menentukan berapa jumlah kluster yang akan digunakan.

F. Tahap Kluster

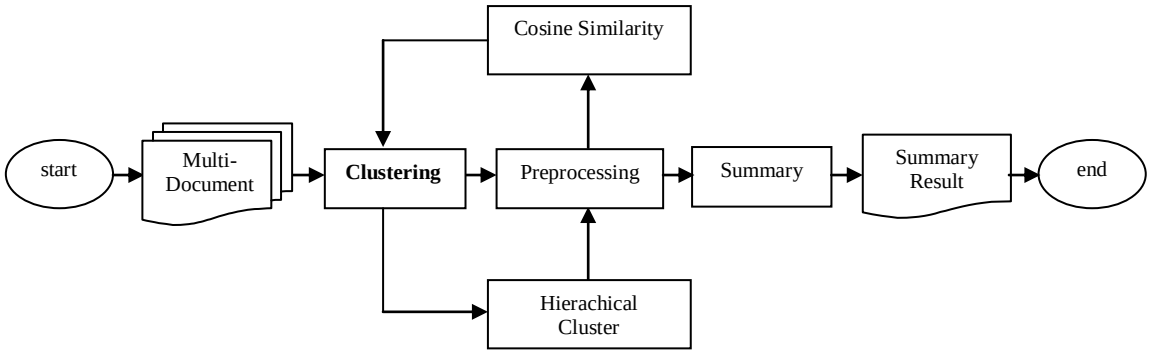
Penyusunan dokumen ke dalam kelompok memiliki tujuan agar dokumen yang memiliki kemiripan dengan dokumen lain akan dijadikan satu kelompok. Tahap kluster merupakan tahapan yang penting dari sistem ringkasan, karena setiap topik dalam kumpulan dokumen harus diidentifikasi dengan benar untuk menemukan kemiripan dan ketidakmiripan.

Apabila kalimat dikempokkan ke dalam satu kelompok yang telah ditentukan, kluster mungkin tidak koheren karena beberapa kalimat mungkin dipaksa menjadi salah satu anggota kluster walaupun tidak memenuhi keriteria yang sudah ditentukan. Kelompok yang tidak sesuai mungkin berisi unit teks yang diduplikasi pada kelompok yang berbeda. Duplikasi ini menyebabkan hasil ringkasan menjadi redundan atau menurunan tingkat relevansinya. Sebaliknya, jika kluster sangat tinggi, sebagian besar kluster menjadi singletons atau terdiri dari satu dokumen saja pada satu kluster. Dengan demikian, metode kluster dipilih untuk memastikan koherensi dan meminimalkan jarak antar kluster.

Tahap pertama dari proses pengelompokan dokumen adalah dokumen dibagi menjadi beberapa kalimat. Kalimatnya terbagi menjadi kata-kata. Bobot setiap kata bisa dihitung dengan menggunakan kemiripan masing-masing dokumen. Kemudian setiap kalimat dengan distribusi kata dibentuk menjadi sebuah kelompok. Kluster mengukur tingkat kemiripan dengan kelompok lain secara berpasangan untuk menghasilkan kluster tunggal atau sejumlah kluster yang dibutuhkan. Proses pengelompokan yaitu dengan membandingkan kluster pertama dengan kluster kedua, dan seterusnya. Untuk proses pembentukan kluster menggunakan hirarki kluster ditunjukkan pada Gambar 2.



Gambar 2 Hirarki cluster



Gambar 3. Kerangka Metode Usulan

TABEL 1.

CONTOH KATA TERSEBAR

There was an accident between a motorcycle rider with a pedestrian on the bojonegoro highway. The accident did not claim casualties but motorcycle had been unconscious for some time. There were witnesses who witnessed the accident by motorcycle. Hearing that there had been an accident, several police officers drove. The pedestrian claim that the accident happened open because of him but because the bikers motorcycle did not obey then the others.

TABEL 2.

PEMBOBOTAN KATA BERDASARKAN FREKUENSI KUMUNCULAN

Term	**Frequency**
accident	4
motorcycle	5
pedestrian	2
police	1
officers	1
claim	2
etc...	

TABEL 3.

PERBEDAAN PEMBOBTAN MENGGUNAKAN TF DAN KATA SEBARAN

Term	**Frequency**	**W_{d-1}**
accident	4	0,913
motorcycle	5	0,305
pedestrian	2	0,691
police	1	0,297
officers	1	0,297
claim	2	0,691
etc...		

Tabel 1 merupakan contoh dokumen distribusi kata dan dihitung bobot kata berdasarkan frekuensi kejadian dan hasilnya ditunjukkan pada Tabel 2.

Pembobotan term frekuensi pada tabel 2 menunjukkan bahwa term accident dan motorcycle memiliki nilai bobot yang sama sedangkan term accident itu sendiri lebih lebih mewakili isi semua dokumen bila dibandingkan dengan term motorcycle, karena term accident lebih tersebar luas di setiap paragraf[15]. Untuk menghitung distribusi kata-kata dalam dokumen (wd-1) to-j diwakili oleh nilai p yang menunjukkan jumlah paragraf yang mengandung kata, nilai P menunjukkan jumlah paragraf dalam dokumen, menunjukkan frekuensi terjadinya. Dalam paragraf berdasarkan jumlah kata yang berbeda yang membentuk paragraf ke-i(Ci) dibagi dengan jumlah Ci dalam dokumen.

IV. HASIL DAN PEMBAHASAN

Hasil eksperiment telah dilakukan pada DUC (Document Understanding Conference) 2004 dataset dari Nasional Institute of Standards and Technology (NIST) untuk memvalidasi kinerja metode yang kami usulkan. Dataset berisi 50 dokumen, setiap topik terdiri dari 50 topik, dan setiap topik memiliki 10 dokumen. Percobaan diimplementasikan di Java menggunakan Text Editor Netbeans 8.2 dan database menggunakan MySql pada sistem operasi Window 7.

Tabel 4 menunjukkan bahwa metode yang diusulkan menggunakan 132, 135, 137, dan 140 kluster menghasilkan nilai yang fluktuatif. Semakin kecil jumlah kluster tidak menjamin peningkatan nilai ROUGE-1. Hal ini disebabkan karena nilai similarity yang dipakai adalah nilai similarity terbaru dari kluster yang baru terbentuk. Nilai kluaster baru tersebut bisa memiliki nilai similarity yang fluktuatif. Terlihat pada Tabel 4, peningkatan terjadi antara kluaster 140 dan kluster 137. Akan tetapi, penurunan terjadi antara kluster 137 dan 135.

Tabel 5 menunjukkan bahwa metode yang diusulkan menggunakan jumlah kluster yang sama memiliki nilai ROUGE-1 lebih rendah dibandingkan metode sebelumnya. Hal ini dikarenakan pada kluster 140 nilai similarity pada masing-masing kluster banyak yang mengalami penurunan nilai similarity. Hasil similarity pada kluster yang digabungkan bisa dilihat pada Tabel 6 dan Tabel 7.

Sedangkan perbandingan hasil peringkasan antara kedua metode bisa dilihat pada Tabel 8.

TABEL 4.

PERBANDINGAN NILAI ROUGE-1 ANTARA TIGA JUMLAH KLUSTER

Jumlah Kluster	Rouge 1
132 kluster	0.24512
135 kluster	0.22136
137 kluster	0.23080
140 kluster	0.23077

TABEL 5.

PERBANDINGAN NILAI ROUGE-1 ANTARA METODE YANG DIUSULKAN DENGAN METODE SEBELUMNYA

Jumlah Kluster	Rouge 1
Hirarki Kluster 140 kluster	0.23077
SHC 140 kluster	0.24511

TABEL 6.

SIMILARITY PADA KLUSTER 137

No	Nama Kluster	Rouge 1
1	Kluster 1	0.32507
2	Kluster 2+Kluster 3	0.41438
3	Kluster 4	0.10376
4	Kluster 5	0.42541
...		
137		

TABEL 7.

SIMILARITY PADA KLUSTER 135

No	Nama Kluster	Rouge 1
1	Kluster 1+ (Kluster 2+Kluster 3)	0.27104
2	Kluster 4+Kluster 5	0.14272
..		
135		

TABEL 8.

PERBANDINGAN HASIL PERINGKASAN

Hiraki Kluster – 140 Kluster
Concerned that crowded shelter conditions could produce outbreaks of hepatitis, respiratory infections and other ailments, the Health Ministry announced an inoculation campaign, especially for children. Mitch _ once, 2nd graf pvs. Countries overwhelmed by the storm's devastation have only just begun to calculate the damage. Numbers still can vary wildly. About 100 victims had been buried around Tegucigalpa, Mayor Nahum Valladeres said. He urged the more than 1.5 million Hondurans affected by the storm to help with the recovery effort. As of Thursday, Mitch had killed 6,076 people in Honduras _ down from officials' earlier estimate of 7,000. At least
SHC – 140 Kluster
At least 231 people have been confirmed dead in Honduras from former-hurricane Mitch, bringing the storm's death toll in the region to 357, the National Emergency Commission said Saturday. A hurricane warning was also in effect for the Caribbean coast of Guatemala. Jerry Jarrell, the weather center director, said Mitch was the strongest hurricane to strike the Caribbean since 1988, when Gilbert killed more than 300 people. At 0900 GMT Tuesday, Mitch was 95 miles (152 kilometers) north of Honduras, near the Swan Islands. In Washington on Thursday, President Bill Clinton ordered dlrs 30 million in Defense Department equipment and services and dlrs 36 million in food, fuel and other aid be sent to Honduras, Nicaragua, El Salvador and Guatemala.

V. KESIMPULAN

Penelitian ini mengusulkan metode hiraki kluster untuk melakukan pengelompokan dengan cara membandingkan kelompok secara berpasangan. Proses membandingkan antar kluster tersebut dengan mengukur nilai kemiripannya. Bila nilai kemiripannya telah diperoleh, nilai kemiripan tertinggi digabungkan menjadi satu sampai tersisa satu kluster. Kluster tersebut kemudian diberikan pembobotan menggunakan local sebaran sentence untuk menghitung penyebaran kata.

Hasil percobaan menggunakan dataset DUC 2004 menggunakan dua skenario pengujian yaitu menggunakan 132, 135, 137 dan 140 kluster menghasilkan nilai yang fluktuatif. Semakin kecil jumlah kluster tidak menjamin peningkatan nilai ROUGE-1. Hal ini disebabkan karena nilai similarity yang dipakai adalah nilai similarity terbaru dari kluster yang baru terbentuk. Nilai kluster baru tersebut bisa memiliki nilai similarity yang fluktuatif.

Metode yang diusulkan menggunakan jumlah kluster yang sama memiliki nilai ROUGE-1 lebih rendah dibandingkan metode sebelumnya. Hal ini dikarenakan pada kluster 140 nilai similarity pada masing-masing kluster banyak yang mengalami penurunan nilai similarity.

ACKNOWLEDGMENT

Kami mengucapkan terima kasih kepada seluruh peneliti pada bidang *summarization* yang memberikan referensi dan memotivasi dalam penulisan *paper* ini.

REFERENSI

[1] R. M. Aliguliyev, "A new sentence similarity measure and sentence based extractive technique for automatic text summarization," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 7764–7772, 2009.

[2] R. Rautray and R. C. Balabantaray, "An evolutionary framework for multi document summarization using Cuckoo search approach: MDSCSA," *Appl. Comput. Informatics*, vol. 14, no. 2, pp. 134–144, 2018.

[3] R. Rautray and R. C. Balabantaray, "Cat swarm optimization based evolutionary framework for multi document summarization," *Phys. A Stat. Mech. its Appl.*, vol. 477, pp. 174–186, 2017.

[4] A. Wahib, Arifin Z.A, and D. Purwitasari, "Peringkasan Dokumen Berbahasa Inggris Menggunakan Sebaran Local Sentence," *J. Buana Inform.*, vol. 7, pp. 33–42, 2016.

[5] A. Z. Arifin and A. Asano, "Image segmentation by histogram thresholding using hierarchical cluster analysis," *Pattern Recognit. Lett.*, vol. 27, no. 13, pp. 1515–1521, 2006.

[6] H. P. Luhn, "The Automatic Creation of Literature Abstracts," *IBM J. Res. Dev.*, vol. 2, no. 2, pp. 159–165, Apr. 1958.

[7] H. P. Edmundson, "New Methods in Automatic Extracting," *J. ACM*, vol. 16, pp. 264–285, 1969.

[8] P. B. Baxendale, "Machine-Made Index for Technical Literature—An Experiment," *IBM J. Res. Dev.*, vol. 2, no. 4, pp. 354–361, 1958.

[9] E. Liddy, "Advances in Automatic Text Summarization," *Inf. Retr. Boston.*, vol. 4, no. 1, pp. 82–83, Apr. 2001.

[10] C.-Y. Lin, "Training a Selection Function for Extraction," in *Proceedings of the Eighth International Conference on Information and Knowledge Management*, 1999, pp. 55–62.

[11] D. Das and A. F. T. Martins, "A Survey on Automatic Text Summarization," *Eighth ACIS Int. Conf. Softw. Eng. Artif. Intell. Netw. ParallelDistributed Comput. SNPD 2007*, vol. 4, pp. 574–578, 2007.

[12] K. Mckeown and D. R.Radev, "Generating Summaries of Multiple News Articles," *Proc. 18th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, vol. 3, pp. 74–82, 1995.

[13] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," *Proc. 21st Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. - SIGIR '98*, pp. 335–336, 1998.

[14] D. R. Radev, H. Jing, and M. Budzikowska, "Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies," *Inf. Process. Manag. 40.6 919-938.*, vol. 40, no. 6, p. 10, 2000.

[15] T. Xia and Y. Chai, "An improvement to TF-IDF: Term distribution based term weight algorithm," *J. Softw.*, vol. 6, no. 3, pp. 413–420, 2011.