

PERINGKASAN MULTI-DOKUMEN MENGGUNAKAN METODE PENGELOMPOKKAN BERBASIS HIRARKI DENGAN MULTI-LEVEL DIVISIVE COEFFICIENT

Muhamad Mustamiin¹, Ahmad Lubis Ghozali², dan Muhammad Lukman Sifa³

¹²³Teknik Informatika, Politeknik Negeri Indramayu, Jl. Raya Lohbener Lama no.8, Indramayu, 45252
E-mail: ¹mustamiin@polindra.ac.id, ²lubis@polindra.ac.id, ³lukmanpolindra@yahoo.co.id

(Naskah masuk: 16 Oktober 2018, diterima untuk diterbitkan: 06 November 2018)

Abstrak

Peringkasan merupakan salah satu bagian dari perolehan informasi yang bertujuan untuk mendapatkan informasi secara cepat dan efisien dengan membuat intisari dari suatu dokumen. Dokumen-dokumen khususnya dokumen laporan setiap hari semakin bertambah seiring dengan bertambahnya pelaksanaan suatu kegiatan atau acara. Kebutuhan informasi yang semakin cepat, jumlah dokumen yang semakin bertambah banyak membuat kebutuhan akan adanya peringkasan dokumen semakin tinggi. Peringkasan yang digunakan untuk meringkas lebih dari satu dokumen disebut peringkasan multi-dokumen. Untuk mencegah adanya penyampaian informasi yang berulang pada peringkasan multi-dokumen, maka proses pengelompokan diperlukan untuk menjamin bahwa informasi yang disampaikan bervariasi dan mencakup semua bagian dari dokumen-dokumen tersebut. Pengelompokan hirarki dengan multi-level *divisive coefficient* dapat digunakan untuk mengelompokkan suatu bagian/kalimat dalam dokumen-dokumen dengan bervariasi dan mendalam yang disesuaikan dengan tingkat kebutuhan informasi dari pengguna. Berdasarkan dari tingkat kompresi peringkasan yang berbeda-beda, peringkasan menggunakan pengelompokan hirarki dengan multi-level *divisive coefficient* dapat menghasilkan hasil peringkasan yang cukup baik dengan nilai *f*-measure sebesar 0,398 sementara nilai *f*-measure peringkasan dengan satu level *divisive coefficient* hanya mencapai 0,335.

Kata kunci: Perolehan Informasi, Pengelompokan, Peringkasan, Multi-dokumen

MULTI-DOCUMENTS SUMMARIZATION USING HIERARCHICAL CLUSTERING METHODS WITH MULTI-LEVEL DIVISIVE COEFFICIENT

Abstract

Summarization is one part of the information retrieval that aims to obtain information quickly and efficiently by making the essence of a document. Documents, especially document reports every day increasing as the implementation of an event. The need for information is getting faster, the increasing number of documents makes the need for document summaries is getting higher. Summarization used to summarize more than one document is called multi-document summarization. To prevent repetitive information from being submitted to multi-document summarization, the grouping process is necessary to ensure that the information submitted varies and covers all parts of the documents. Hierarchical clustering with multi-level *divisive coefficient* can be used to group a part / sentence in documents with varying and depth adjusted to the level of information needs of the user. Based on different compression levels of summarization, summarization using hierarchical clustering with multi-level *divisive coefficient* can produce a fairly good summary result with *f*-measure value of 0.398 while the *f*-measure summarization value with one level of *divisive coefficient* only reaches 0.335.

Keywords: Information Retrieval, Clustering, Summarization, Multi-documents

1. PENDAHULUAN

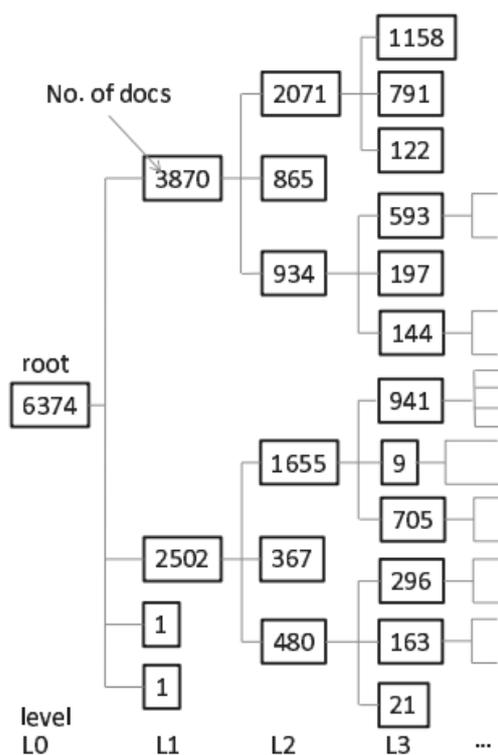
Dokumen-dokumen khususnya dokumen laporan setiap hari semakin bertambah seiring dengan bertambahnya pelaksanaan suatu kegiatan atau acara. Dokumen-dokumen tersebut terkadang hanya dibuat sebagai bentuk formalitas saja tanpa ada kajian dan evaluasi yang mendalam terkait isi

dari dokumen laporan tersebut. Kegiatan kajian dan evaluasi terkait dokumen laporan memang membutuhkan waktu yang tidak sedikit karena jumlahnya yang juga tidak sedikit.

Terkadang dokumen-dokumen yang ada khususnya dokumen laporan memiliki format serta isi yang tidak jauh berbeda. Salah satu dokumen

laporan yang memiliki format serta isi yang pembahasannya tidak jauh berbeda adalah dokumen laporan Program Praktik Industri (PPI) yang setiap tahun jumlahnya bertambah cukup banyak. Pembuatan dokumen laporan PPI sendiri dibuat dengan tujuan untuk mendapatkan gambaran dari perusahaan dan keadaan dunia industri pada umumnya serta laporan kegiatan mahasiswa saat melaksanakan PPI (Politeknik Negeri Indramayu, 2013).

Peringkasan yang digunakan untuk meringkas lebih dari satu dokumen disebut peringkasan multi-dokumen. Untuk mencegah adanya penyampaian informasi yang berulang pada peringkasan multi-dokumen membuat proses pengelompokan diperlukan untuk menjamin bahwa informasi yang disampaikan bervariasi dan mencakup semua bagian dari dokumen-dokumen tersebut. Pengelompokan berbasis hirarki memiliki performa yang lebih baik jika dibandingkan dengan pengelompokan berbasis flat atau K-Means khususnya untuk data yang masih belum diketahui jumlah pasti kelompoknya (Mustamiin M., Budi, I., & Santoso, H. B., 2018).



Gambar 1. Dendrogram oleh HMDC (KISHIDA, 2017)

Pengelompokan berbasis hirarki memiliki informasi yang lebih detail sehingga bisa dijadikan sebagai bahan untuk analisis data (Pandit, S. R. & Potey, M. A., 2013). Pengelompokan berbasis hirarki untuk mengelompokkan dokumen dengan Algoritma Hierarchical Multi-way Divisive Clustering (HMDC) menunjukkan hasil yang baik dan efektif (Kishida, 2017). Pengelompokan berbasis hirarki dengan multi-level divisive

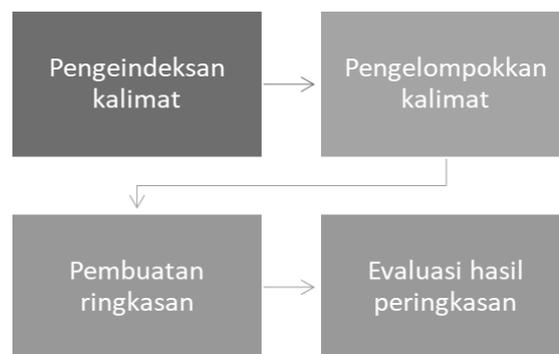
coefficient digunakan dalam penelitian ini karena selain dapat memberikan detail informasi juga dapat digunakan untuk mengelompokkan suatu bagian/kalimat dalam dokumen-dokumen dengan bervariasi dan mendalam yang disesuaikan dengan tingkat kebutuhan informasi dari pengguna.

Pengelompokan dokumen dengan menggunakan algoritma Hierarchical Multi-way Divisive Clustering (HMDC) menunjukkan hasil yang baik dan efektif (Kishida, 2017). Dalam pengelompokan dokumen tersebut setiap level dari hasil pengelompokan digambarkan dengan jumlah kelompok yang berbeda pada setiap levelnya. Gambar 1 menjelaskan bagaimana ilustrasi pengelompokan yang digambarkan dalam bentuk dendrogram serta level dan jumlah dokumen dari setiap kelompoknya (Kishida, 2017).

Metode evaluasi digunakan untuk menilai performa dari sebuah sistem perolehan informasi diantaranya adalah precision, recall, f-measure (Manning, C. D., Raghavan, P. & Schütze, H., 2002). Metode-metode tersebut dapat dijadikan alat ukur untuk mengevaluasi hasil peringkasan dokumen.

2. METODE PENELITIAN

Secara umum penelitian ini memiliki beberapa tahapan utama yaitu: (1) Pengindeksan kalimat; (2) Pengelompokan kalimat; (3) Pembuatan ringkasan; dan (4) Evaluasi hasil peringkasan. Gambar 2 menunjukkan alur proses peringkasan yang merupakan rancangan dari proses penelitian yang dilakukan.



Gambar 2. Alur Proses Peringkasan

a) Pengindeksan kalimat

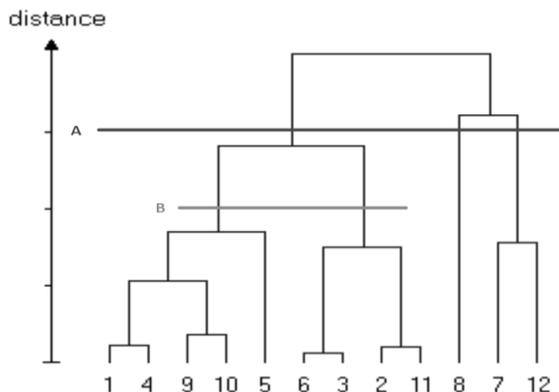
Pada tahap ini terlebih dahulu dilakukan pemilahan terhadap teks seperti menghilangkan tanda baca, menghilangkan kata imbuhan (*stemming*) dan juga menghilangkan kata tidak penting (*stopwords*). Kemudian dilakukan proses pengindeksan kalimat dari dokumen-dokumen yang akan dibuat ringkasannya. Dokumen dalam penelitian ini adalah dokumen laporan Program Praktek Industri (PPI) pada jurusan Teknik Informatika Politeknik Negeri Indramayu tahun 2016.

Dalam proses pengindeksan kalimat ini juga dilakukan perhitungan jarak antara satu kalimat

dengan kalimat lainnya berdasarkan nilai dari perbedaan kalimat, yaitu nilai maksimal 1 (satu) dikurangi nilai cosine similarity antar kalimat tersebut. Sebelum menghitung jarak antar kalimat terlebih dahulu dilakukan perhitungan bobot kata dengan pembobotan TF-IDF pada kata yang terdapat pada tiap kalimat yang telah diindeks. Setelah itu, dilakukan perhitungan cosine similarity terhadap kalimat yang dihitung jaraknya.

b) Pengelompokkan kalimat

Kalimat-kalimat yang sudah di indeks pada tahap sebelumnya kemudian dikelompokkan menggunakan pengelompokkan berbasis hirarki berdasarkan jarak antar kalimat-kalimat yang sudah diindeks. Hasil pengelompokkan tersebut kemudian akan menjadi bahan untuk pembuatan ringkasan. Gambar 3 menggambarkan bagaimana ilustrasi dari nilai division coefficient yang berbeda-beda, pada gambar tersebut dicontohkan pemilihan nilai division coefficient dengan dua level berbeda yaitu pada nilai A dan untuk subkelompok pada nilai B.



Gambar 3. Pengelompokkan Hirarki dengan Multi-Level Division Coefficient.

c) Pembuatan ringkasan

Pembuatan ringkasan dilakukan dengan mengambil hasil dari pengelompokkan kalimat yang sudah dibuat pada tahap sebelumnya. Masing-masing kelompok akan menjadi bahan dalam ringkasan yang dibuat, dalam proses ini tidak serta merta setiap perwakilan kelompok akan menjadi bagian ringkasan melainkan ada aturan khusus dimana setiap perwakilan kalimat dari kelompok hasil nilai division coefficient level pertama (A) akan menjadi kelompok parent yang kemudian kelompok tersebut dilakukan pemotongan kembali pada kelompok didalamnya dengan nilai division coefficient level kedua (B).

Gambar 3 menunjukkan pada level pertama (A) akan menghasilkan 3 poin ringkasan, kemudian pada level kedua (B), poin ringkasan dari hasil parent poin pertama menghasilkan dua kelompok lagi yang akan menjadi poin tambahan penjelasan dari hasil poin ringkasan pertama. Pemilihan tingkat pemotongan dendogram yang berbeda ini bertujuan untuk memberikan gambaran umum hasil ringkasan dari level pertama (A) dan untuk level kedua (B)

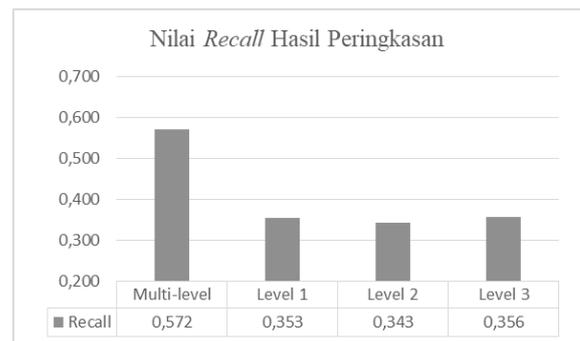
diharapkan dapat menjelaskan lebih detail jika user membutuhkan penjelasan yang lebih lengkap terkait bagian ringkasan tersebut.

d) Evaluasi hasil peringkasan

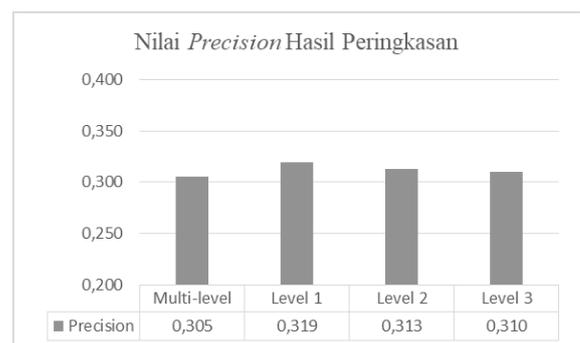
Evaluasi digunakan untuk menilai performa dari sebuah sistem perolehan informasi diantaranya adalah precision, recall, f-measure (Manning, C. D., Raghavan, P. and Schütze, H., 2002). Sebagai acuan dalam evaluasi, peneliti menggunakan gold standard, dimana gold standard dibuat oleh dua Dosen dan satu Mahasiswa Teknik Informatika, Politeknik Negeri Indramayu.

3. HASIL DAN PEMBAHASAN

Dalam penelitian ini, peneliti menggunakan tiga nilai division coefficient yang berbeda yaitu pada nilai jarak 0,8 (Level 1), 0,4 (Level 2), 0,6 (Level 3), dan untuk implementasi multi-level division coefficient sendiri peneliti menggunakan nilai jarak 0,8 dan 0,4. Eksperimen peringkasan kemudian dievaluasi dengan membandingkan hasil peringkasan dengan gold standard yang telah dibuat sebelumnya.



Gambar 4. Nilai Recall Hasil Eksperimen Peringkasan

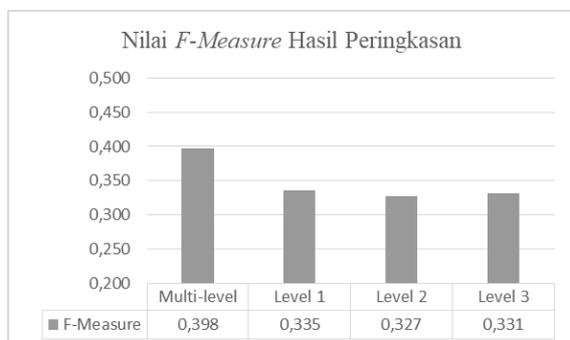


Gambar 5. Nilai Precision Hasil Eksperimen Peringkasan

Gambar 4 menunjukkan nilai recall hasil eksperimen peringkasan, dari gambar tersebut dapat diketahui bahwa nilai recall untuk peringkasan dengan multi-level division coefficient memiliki nilai tertinggi yaitu sebesar 0,572 sedangkan peringkasan dengan satu level division coefficient hanya mencapai nilai 0,356 yang dimiliki oleh peringkasan dengan nilai jarak 0,6.

Sementara itu, untuk nilai precision hasil eksperimen peringkasan yang ditunjukkan oleh

Gambar 5 menunjukkan bahwa nilai precision untuk peringkasan dengan multi-level division coefficient memiliki nilai sebesar 0,305 sedangkan peringkasan dengan satu level division coefficient dapat mencapai nilai 0,319 yang dimiliki oleh peringkasan dengan nilai jarak 0,8.



Gambar 6. Nilai F-Measure Hasil Eksperimen Peringkasan

Untuk nilai f-measure hasil eksperimen peringkasan ditunjukkan oleh Gambar 6 menunjukkan bahwa nilai f-measure untuk peringkasan dengan multi-level division coefficient memiliki nilai tertinggi yaitu sebesar 0,398 sedangkan peringkasan dengan satu level division coefficient hanya mencapai nilai 0,335 yang dimiliki oleh peringkasan dengan nilai jarak 0,8.

4. KESIMPULAN

Dari keseluruhan eksperimen yang telah dilakukan diperoleh kesimpulan bahwa metode peringkasan menggunakan pengelompokan berbasis hirarki dengan multi-level divisive coefficient dapat dijadikan sebagai salah satu metode yang cukup baik untuk membuat ringkasan multi-dokumen secara lengkap ini dibuktikan dengan nilai F-Measure mencapai 0,398 dibandingkan dengan pengelompokan berbasis hirarki menggunakan satu nilai divisive coefficient yang hanya mencapai 0,335.

Terkait penelitian selanjutnya peneliti memiliki beberapa saran, diantaranya: penambahan jumlah koleksi dokumen yang lebih banyak, penggunaan metode yang beragam dalam penentuan jarak antar kalimat, dan pemanfaatan template atau format contoh laporan sebagai acuan dalam pembuatan ringkasan.

5. DAFTAR PUSTAKA

Dendograms. 2012. Dendograms. Tersedia di: <http://www.statistics4u.com/fundstat_eng/cc_dendograms.html> [Diakses 20 Juli 2018]

GUPTA, V. K. AND SIDDIQUI, T. J. 2012. Multi-document Summarization using sentence clustering. Kharagpur: Intelligent Human Computer Interaction (IHCI), pp. 1-5.

KISHIDA, KAZUAKI. 2017. "An Experiment on Simple and Practical Methods of Cluster Labeling for Hierarchically Organized Document Subsets".

MANNING, C. D., RAGHAVAN, P. AND SCHÜTZE, H. 2002. An Introduction To Information Retrieval [online]. Tersedia di: <<http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf>> [Diakses 1 Juni 2018]

MEENA, Y. K., JAIN, A. AND GOPALANI, D. 2014. Survey on Graph and Cluster Based approaches in Multi-document Text Summarization. Jaipur: International Conference Recent Advances and Innovations in Engineering (ICRAIE), pp. 1-5.

MUSTAMIIN, M., BUDI, I., SANTOSO, H. B. (2018). Multi-documents summarization based on clustering of learning object using hierarchical clustering. Journal of Physics: Conference Series 978 (1), 012053.

PANDIT, S. R. AND POTEY, M. A. 2013. A Query Specific Graph Based Approach to Multi-document Text Summarization: Simultaneous Cluster and Sentence Ranking. Katra: Machine Intelligence and Research Advancement (ICMIRA), pp. 213-217.

PADMALAHARI, E., KUMAR, D. V. AND PRASAD, S. 2014. Automatic text Summarization with statistical and linguistic features using successive thresholds. Ramanathapuram: IEEE, pp. 1519-1524.

Politeknik Negeri Indramayu. 2013. Panduan PPI Polindra. [online]. Tersedia di: <<http://polindra.ac.id/~download/dokumen/panduan%20PPI%20polindra%20fix.pdf>> [Diakses 1 Juli 2018]

SHEPITSEN, A., GEMMELL, J., MOBASHER, B. AND BURKE, R. 2008. Personalized recommendation in social tagging systems using hierarchical clustering. New York: ACM, pp. 259-266.

WANG, S., LI, W., WANG, F. AND DENG, H. 2010. A Survey on Automatic Summarization. Kunming: IEEE, pp. 193-196.