041509241.pdf

by

Submission date: 26-Sep-2018 08:24AM (UTC+0700)
Submission ID: 1008467719
File name: 041509241.pdf
Word count: 1829
Character count: 12259

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

Muhamad Mustamiin [1)], Ahmad Lubis Ghozali [2)], dan Muh. Lukman Sifa [3)]

[123] Teknik Informatika, Politeknik Negeri Indramayu, Jl. Raya Lohbener Lama no.8, Indramayu, 45252
E-mail: m.mustamiin@gmail.com

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*

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 pengelompokkan diperlukan untuk menjamin bahwa informasi yang disampaikan bervariasi dan mencakup semua bagian dari dokumen-dokumen tersebut. Pengelompokkan 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. Bedasarkan dari tingkat kompresi peringkasan yang berbeda-beda, peringkasan menggunakan pengelompokkan 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, Pengelompokkan, Peringkasan, Multi-dokumen*

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.

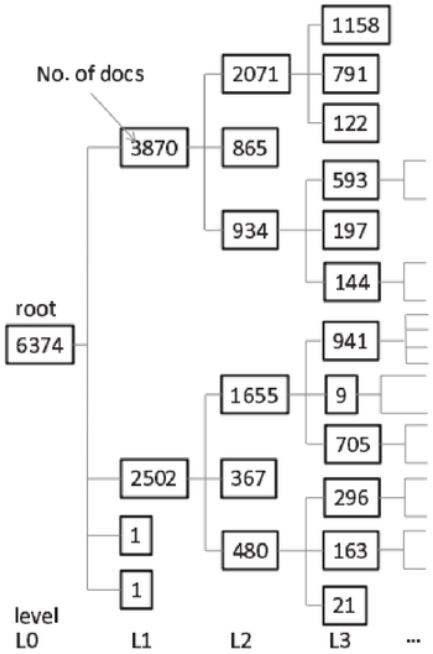
Terkadang dokumen-dokumen yang ada khususnya dokumen laporan memiliki format serta isi yang tidak jauh berbeda. Salah satu dokumen laporan yang memiliki format serta isi yang pembahasannya tidak jauh berbeda adalah dokumen laporan Program Praktik Industri (PPI) yang setiap tahun jumlahnya bertambah cukup 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 pengelompokkan diperlukan untuk menjamin bahwa informasi yang disampaikan bervariasi dan mencakup semua bagian dari dokumen-dokumen tersebut. Pengelompokkan berbasis hirarki memiliki performa yang lebih baik jika dibandingkan dengan pengelompokkan berbasis *flat* atau K-Means khususnya untuk data yang masih belum diketahui jumlah pasti kelompoknya (Mustamiin M., Budi, I., & Santoso, H. B., 2018).

Pengelompokkan berbasis hirarki memiliki informasi yang lebih detil sehingga bisa dijadikan sebagai bahan untuk analisis data (Pandit, S. R. & Potey, M. A., 2013). Pengelompokkan berbasis hirarki untuk mengelompokkan dokumen dengan Algoritma *Hierarchical Multi-way Divisive Clustering* (HMDC) menunjukkan hasil yang baik dan efektif (KISHIDA, 2017). Pengelompokkan berbasis hirarki dengan *multi-level divisive coefficient* digunakan dalam penelitian ini karena selain dapat memberikan detil informasi juga dapat digunakan untuk mengelompokkan suatu bagian/kalimat dalam dokumen-dokumen dengan bervariasi dan mendalam yang disesuaikan dengan tingkat kebutuhan informasi dari pengguna.

Pengelompokkan dokumen dengan menggunakan algoritma *Hierarchical Multi-*

way Divisive Clustering (HMDC) menunjukkan hasil yang baik dan efektif (KISHIDA, 2017). Dalam pengelompokkan dokumen tersebut setiap level dari hasil pengelompokkan digambarkan dengan jumlah kelompok yang berbeda pada setiap levelnya. Gambar 1 menjelaskan bagaimana ilustrasi pengelompokkan yang digambarkan dalam bentuk dendogram serta level dan jumlah dokumen dari setiat kelompoknya (KISHIDA, 2017).

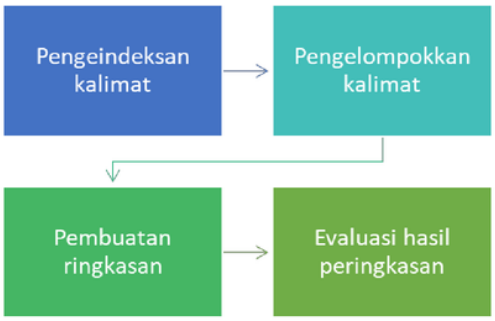


Gambar 1. Dendogram oleh HMDC (KISHIDA, 2017)

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 mengevalusi hasil peringkasan dokumen.

METODE PENELITIAN

Secara umum penelitian ini memiliki beberapa tahapan utama yaitu: (1) Pengindeksan kalimat; (2) Pengelompokkan 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

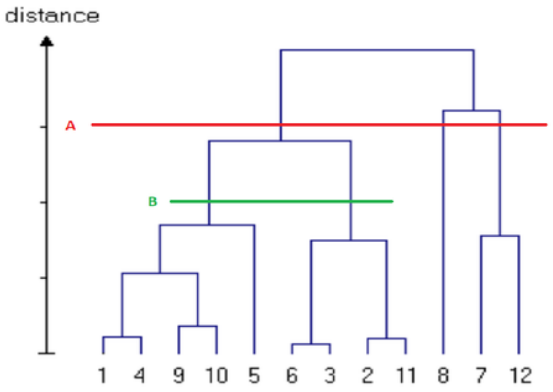
1) Pengindeksan kalimat

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.

2) 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*.

3) 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.

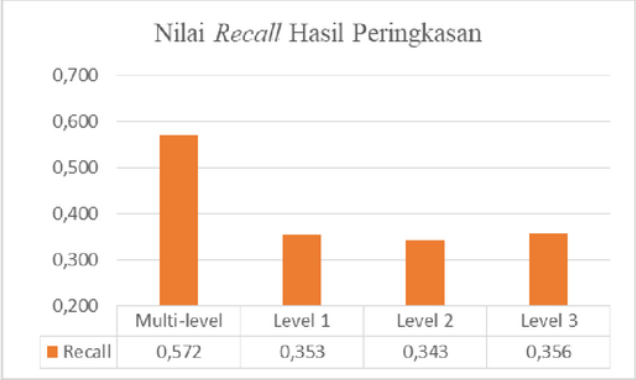
4) Evaluasi hasil peringkasan

Evaluasi digunakan untuk menilai performa dari sebuah sistem perolehan informasi diantaranya adalah *precision*, *recall*, *f-measure* (Manning, C. D.,

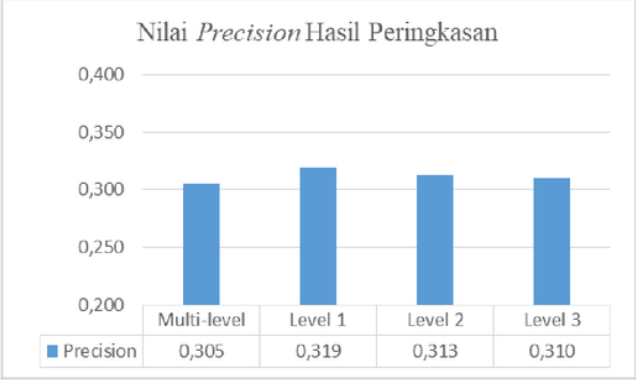
Raghavan, P. and Schütze, H., 2002). Sebagai acuan dalam evaluasi, peneliti menggunakan *gold standard*, dimana *gold standard* dibuat oleh dua Dosen dan satu Mahasiswa Teknik Informatika, Politeknik Negeri Indramayu.

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

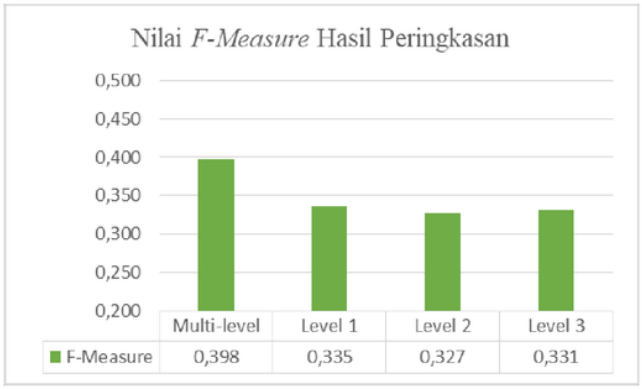


Gambar 5. Nilai *Precision* Hasil Eksperimen Peringkasan

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 peringksan 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 peringksan 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 peringksan dengan nilai jarak 0,8.

SIMPULAN

Dari keseluruhan eksperimen yang telah dilakukan diperoleh kesimpulan bahwa metode peringkasan menggunakan pengelompokkan 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 pengelompokkan 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.

DAFTAR PUSTAKA

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.

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.

Gupta, V. K. and Siddiqui, T. J. (2012). *Multi-document Summarization using sentence clustering*. Kharagpur: Intelligent Human Computer Interaction (IHCI), pp. 1-5.

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.

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.

Wang, S., Li, W., Wang, F. and Deng, H. (2010). *A Survey on Automatic Summarization*. Kunming: IEEE, pp. 193-196.

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.

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

Politeknik Negeri Indramayu. (2013). Panduan PPI Polindra. [PDF document]. Retrieved from http://polindra.ac.id/~download/dokumen/panduan%20PPI%20polindra%20fix.pdf

Manning, C. D., Raghavan, P. and Schütze, H. (2002). *An Introduction To Information Retrieval* [PDF document]. Retrieved from http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf

Dendrograms. (2012). [Graphic Dendogram July 20, 2018]. Dendograms. Retrieved from http://www.statistics4u.com/fundstat_eng/cc_dendrograms.html

041509241.pdf

ORIGINALITY REPORT

15%	13%	12%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	M Mustamiin, I Budi, H B Santoso. "Multi-documents summarization based on clustering of learning object using hierarchical clustering", Journal of Physics: Conference Series, 2018 Publication	9%
2	Submitted to University of Sheffield Student Paper	1%
3	china.iopscience.iop.org Internet Source	1%
4	web.flet.keio.ac.jp Internet Source	1%
5	Sifa, Agus, Badruzzaman, and Dedi Suwandi. "Ply Thickness Fiber Glass on Windmill Drive Salt Water Pump", IOP Conference Series Materials Science and Engineering, 2016. Publication	1%
6	ccsun.nchu.edu.tw Internet Source	1%
7	eprints.ums.ac.id	

	Internet Source	<1%
8	konser.web.id Internet Source	<1%
9	www.slideshare.net Internet Source	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		