Nusantara Journal of Computers and Its Applications
Volume 2 No 1 Juni 2017

PENGGUNAAN DICTIONARY-BASED DAN CORPUS-BASED THESAURUS UNTUK PEMBOBOTAN TERM PADA PENGELOMPOKAN DOKUMEN BERITA BERBAHASA INDONESIA

Amelia Sahira Rahma[1)], Vit Zuraida[2)], Dimas Fanny Hebrasianto Permadi[3)]
[1, 2, 3)]Departemen Teknik InformatikaInstitut Teknologi Sepuluh Nopember
Jl. Raya ITS Kampus Sukolilo, Surabaya, 60111, Indonesia
Email : amelia.sahira.rahma16@mhs.if.its.ac.id[1)], vit16@mhs.if.its.ac.id[2)], dimas15@mhs.if.its.ac.id[3)]

ABSTRAK

Banyaknya dokumen berita berbahasa Indonesia menimbulkan kebutuhan akan document clustering berdasarkan topik untuk mempermudah pembaca mendapatkan berita terkait pada topik yang sama. Namun terdapat permasalahan yang sering dihadapi dalam document clustering yaitu rendahnya relevansi pada hasil cluster sehingga dokumen belum terkelompokkan dalam topik yang sesuai. Paper ini mengusulkan suatu metode pembobotan baru dengan menggunakan kombinasi corpus-based thesaurus dan dictionary-based thesaurus yang mempertimbangkan similaritas konseptual antar term. Metode ini diuji dengan algoritma K-Means terhadap 253 dokumen berita berbahasa Indonesia sebagai data uji. Hasil uji coba membuktikan bahwa penggunaan corpus-based thesaurus dan dictionary-based thesaurus menujukkan performa yang baik.

***Kata Kunci**:Pembobotan term, Pengelompokan dokumen, Tesaurus, Tesaurus berdasarkan korpus, Tesaurus berdasarkan kamus*

ABSTRACT

Huge numbers of digital news document in Indonesian Language led to the need for automatic document clustering based on topic so readers would have an easier access to news articles in the same topic. One of the major problems in document clustering is low relevancy in the clustering result so the documents are not grouped based on their appropriate topic. This paper proposed a new term weighting method that employs combination of corpus-based thesaurus and dictionary-based thesaurus to consider conceptual similarity between terms. This method is evaluated using K-Means algorithm to 253 news document in Indonesian language. Experimental results show that the proposed term weighting method is able to achieve good performance.

***Keywords**: Corpus-based thesaurus, Dictionary-Based Thesaurus, Document Clustering, Term weighting, Thesaurus*

I. PENDAHULUAN

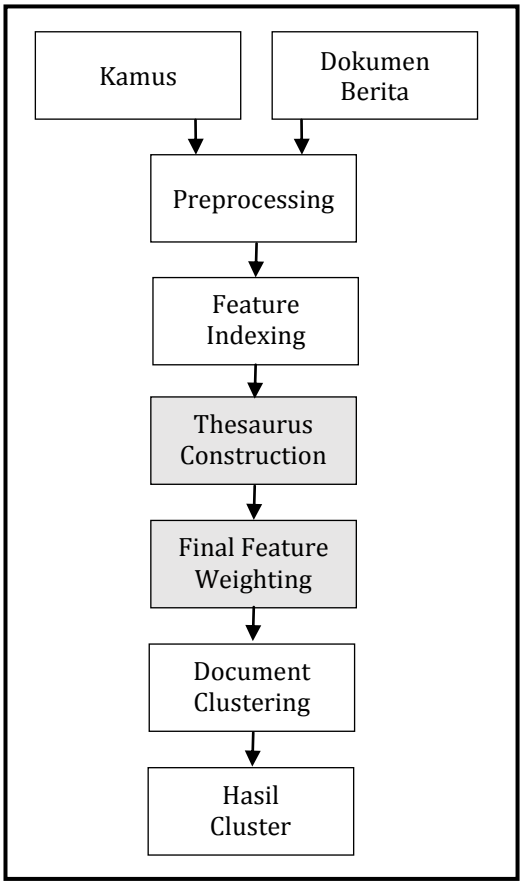
Seiring perkembangan teknologi saat ini, pengelompokan dokumen (*document clustering)*merupakan salah satu bidang riset penting dalam *data mining*. Hal ini disebabkan makin banyaknya dokumen yang tersedia dalam bentuk digital sehingga muncul kebutuhan untuk mengorganisasi dokumen-dokumen tersebut. *Document clustering* adalah sebuah metode pengelompokan dokumen yang dilakukan melalui proses ekstraksi topik dan pengambilan informasi dari sekumpulan dokumen. *Document clustering* bertujuan untuk membagi kumpulan dokumen menjadi beberapa grup kohesif yang disebut *cluster*. Masing-masing *cluster* berisi dokumen-dokumen yang mirip dengan dokumen pada *cluster* yang sama dan tidak miripdengan dokumen pada *cluster* yang lain.*Document clustering*berperandalammemfasilitasi pengorganisasian dokumen digital, pengelompokan hasil pada *search engine*, peringkasan dokumen, dll. Pada dokumen berita berbahasa Indonesia, *document clustering* berguna untuk mempermudah pembacauntuk mencari berita terkait dengan topik yang sama. Beberapa algoritma*document clustering* yang sering digunakan di antaranya K-Means,*agglomerative hierarchical clustering,* dan *ant-based document clustering*.

Dalam *document clustering,* salah satu isu utama adalah cara untuk merepresentasikan dokumen.

Metode yang umum digunakan yaitu dengan membuat vektor dokumen berdasarkan frekuensi kemunculan term pada dokumen yang dikenal dengan *bag of words* (BOW). Namum demikian, *clustering dokumen* yang hanya mengandalkan BOW saja tidak cukup memberikan performa yang baik karena dokumen hanya dikelompokan berdasarkan term yang benar-benar identik tanpa mempertimbangkan similaritas konseptual antar term[1]. Salah satu solusi yang populer dipakai untuk mengatasi masalah ini adalah penggunaan tesaurus dalam pembobotan term.

Tesaurus adalah himpunan kata yang berhubungan secara semantik dan hirarkikal yang dikelompokan menurut kedekatan makna. Tesaurus dapat digunakan untuk menerjemahkan bahasa sehari-hari ke dalam bahasa indeks di area ilmu pengetahuan tertentu [2].Terdapat dua jenis tesaurus, yaitu tesaurus yang dibangun secara manual oleh ahli bahasa dan tesaurus yang dibangun secara otomatis.Tesaurus yang dibangun secara manual seperti WordNet (tesaurus Bahasa Inggris) berpedoman pada kata yang disusun dalam taksonomi berdasarkan hubungan antar term, di antaranya sinonim, hipernim, dan moronim. Kelebihan dari metode ini adalah pembobotan term tidak hanya didasarkan pada frekuensi kemunculan term yang identik secara leksikal saja, tetapi juga dekat dalam hal makna. Jenis tesaurus berikutnya adalah tesaurus otomatis yang umumnya dibuat berdasarkan korpusdokumen (*corpus-based thesaurus*).*Corpus-based thesaurus* dibangun dengan menghitung nilai similaritas antar term berdasarkan *co-occurrence* (kemunculan term secara bersamaan) di suatu kumpulan dokumen [3]. Kelebihan dari metode ini adalah dipertimbangkannya similaritas term-term yang tidak terdapat dalam WordNet, misalnya angka, nama, istilah baru, dan singkatan. Di sisi lain, *corpus-based thesaurus* rentan terhadap term yang sering muncul bersamaan tetapi tidak memiliki makna yang berkaitan. Masalah ini seharusnya dapat diatasi dengan membentuk tesaurus otomatis berdasarkan Kamus Besar Bahasa Indonesia sebagai korpus (*dictionary-based thesaurus*) karena kalimat-kalimat di dalamnya merupakan kalimat standar dengan isi yang relatif relevan sehingga term-term dengan similaritas tinggi benar-benar terkait dalam hal semantik. Kombinasi tesaurus yang similaritas antar term dihitung berdasarkan *co-occurrence* pada korpus dan kamus akan

.Paper ini mengusulkan suatu metode pembobotan baru dengan menggunakan kombinasi *corpus-based thesaurus* dan *dictionary-based thesaurus* yang mempertimbangkan similaritas konseptual antar term. Metode ini akan dievaluasi dengan algoritma K-Means terhadap 253 dokumen berita berbahasa Indonesia. Desain umum dari metode ini ditunjukkan pada Gambar 1.



Gambar 1. Desain Sistem

II. KAJIAN PUSTAKA

Konsep dasar dari tesaurus yang dibangun secara otomatis adalah pengukuran similaritas antar term berdasarkan jumlah kemunculanya di dalam dokumen secara bersamaan. Pendekatan ini didasarkan pada hipotesa bahwa term-term yang berkaitan memiliki tendensi untuk muncul secara bersamaan [4]. Pada bidang *information retrieval*,tesaurus awalnya digunakan untuk melakukan *query expansion* tetapi penelitian selanjutnya membuktikan bahwa tesaurus juga efektif dipakai dalampengelompokan dokumen [2].

Berbagai riset telah banyak dilakukan untuk mengidentifikasi efektifitas tesaurus dalam *document clustering*, baik tesaurus yang dibangun secara manual maupun tesaurus yang dibangun secara otomatis (*corpus-based thesaurus*). Salah satunya adalah riset mengenai penggunaan tesaurus otomatis dalam klasifikasi maupun pengelompokan dokumen [5][6]. Di samping itu, riset lain yang menggunakan kombinasi tesaurus manual dan tesaurut otomatis juga banyak dilakukan, salah satunya adalah riset Li et al [3]yang menggunakan kombinasi WordNet dan *corpus-based thesaurus* dalam pengelompokan dokumen.

Selain itu, Steinberger et al [7]juga melakukan riset lain mengenai tesaurus multilingual yang dibangun dari korpus dokumen dengan berbagai macam bahasa. Nilai similaritas term dihitung terhadap dokumen yang direpresentasikan indepeneden terhadap bahasa yang digunakan.Berbagai penelitian tersebut membuktikan bahwa penggunaan tesaurus dapat memperbaiki performa*document clustering*.

III. METODE

Metode yang diajukan dalam paper ini memiliki lima tahapan utama, yaitu *preprocessing*, *feature indexing*, *thesaurus construction*, *final feature weighting*, dan *document clustering*.

A. Preprocessing

Pada tahap ini, proses yang dilakukan untuk memperoleh term dalam bentuk dasar dari setiap dokumen adalah sebagai berikut:

1. *Case folding*, yaitu mengubah setiap huruf dalam dokumen menjadi *lowercase*.
2. *Filtering*, yaitu menghilangkan karakter ilegal (angka dan simbol).
3. *Stoplist removal*, yaitu mengeliminasi kata berfrekuensi tinggi, misalnya "dan", "yang", dan "atau".
4. *Stemming*, yaitu mengubah setiap term ke dalam bentuk kata dasar. Pada penelitian ini, *stemmer* yang digunakan adalah *enhanced confix stripping stemmer* [8].

B. Feature Indexing

Sebelum proses *clustering* dilakukan, term dari dokumen ditransformasi menjadi *vector space model*. Pada tahap ini, setiap dokumen k direpresentasikan sebagai vektor dari bobot term $i$pada dokumen$k$ (w_{ik}).

$$d_k = (w_{1k}, w_{2k}, \ldots, w_{ik}) \quad (1)$$

Metode pembobotan yang dipilih sebagai bobot inisial adalah *term frequency-inverse document frequency* (TF-IDF) *weigthing*. Pada korpus dengan sejumlah N dokumen dandf_iadalah jumlah dokumen yang mengandung term$i$,bobot term i padadokumen $k$dikalkulasikan sebagai w_{ik} dengan formula (2).

$$w_{ik} = tf_{ik} \times \log_2\left(\frac{N}{df_i}\right) \quad (2)$$

C. Thesaurus Construction

Tahap awal pembentukan tesaurus adalah pengukuran nilai similaritas antar term. Pada *corpus-based thesaurus*, dengan w_{ik} sebagai bobot term $i$pada dokumen k dan w_{jk} sebagai bobot term j pada dokumen k, similaritas term i dan$j$ pada korpusdihitung dengan formula (3).

$$sim_{CT}(t_i, t_j) = \frac{\sum_{\forall d_k} w_{ik} \times w_{jk}}{\sqrt{\sum_{\forall d_k}(w_{ik})^2} \times \sqrt{\sum_{\forall d_k}(w_{jk})^2}} \quad (3)$$

Untuk*dictionary-based thesaurus*, satu kata dasar beserta seluruh definisinya dianggap sebagai satu dokumen. Dengan w_{im} sebagai bobot term$i$ pada dokumen m dan w_{jm} sebagai bobot term j pada dokumen m, similaritas term i dan j pada kamus dihitung dengan formula (4).

$$sim_{DT}(t_i, t_j) = \frac{\sum_{\forall d_m} w_{im} \times w_{jm}}{\sqrt{\sum_{\forall d_m}(w_{im})^2} \times \sqrt{\sum_{\forall d_m}(w_{jm})^2}} \quad (4)$$

D. Final Feature Weighting

Tahap berikutnya adalah menghitung nilai similaritas antara setiap dokumen pada korpus berita dengan masing-masing term.Untuk *corpus-based thesaurus*, dengan w_i sebagai bobot term i dan $sim_{CT}(t_i, t_j)$ sebagai *corpus-based similarity* antara term $i$dan term j, similaritas dokumen berita$k$dengan term $j$dikalkulasi dengan formula (5).

$$sim_{CT}(d_k, t_j) = \sum_{t_i \in d_k}^{n} \left(w_i \times sim_{CT}(t_i, t_j)\right) \quad (5)$$

Sedangkan untuk *dictionary-based thesaurus*, dengan w_i sebagai bobot term $i$dan$sim_{DT}(t_i, t_j)$ sebagai *dictionary-based similarity* antara term $i$dan term$j$, similaritas dokumen berita k dengan term $j$dikalkulasi dengan formula (6).

$$sim_{DT}(d_k, t_j) = \sum_{t_i \in d_k}^{n} \left(w_i \times sim_{DT}(t_i, t_j)\right) \quad (6)$$

Selanjutnya, untuk masing-masing dokumen berita *k*, term diurutkan berdasarkan nilai similaritasnya terhadap dokumen *k* lalu ditentukan sejumlah *top-ranked related terms* berdasarkan urutan tersebut. Bobot dari *top-ranked related terms* kemudian digunakan sebagai bobot tambahan pada vektor original.

Nilai bobot tambahan (*expanded weight*) term *j* pada dokumen *k* berdasarkan *corpus-based thesaurus* dihitung dengan formula (7).

$$e_{WC}(d_k, t_j) = sim_{CT}(d_k, t_j) \Big/ \sum_{t_i \in d_k} w_i \qquad (7)$$

Dengan α sebagai parameter penentu tingkat pengaruh *dictionary-based similarity*, nilai bobot tambahan term *j* pada dokumen *k* berdasarkan *dictionary-based thesaurus* dihitung dengan formula (8).

$$e_{WD}(d_k, t_j) = \alpha \, x \, sim_{DT}(d_k, t_j) \Big/ \sum_{t_i \in d_k} w_i \qquad (8)$$

Berikutnya, nilai *expanded weight* dari *corpus-based thesaurus* dan *dictionary-based thesaurus* dijumlahkan untuk mendapatkan nilai total *expanded weight* dari term *j* seperti yang ditunjukkan pada formula (9).

$$w_j' = e_{WC}(d_k, t_j) + e_{WD}(d_k, t_j) \qquad (9)$$

Nilai total *expanded weight* w_j' kemudian ditambahkan pada bobot awal w_j untuk mendapatkan vektor dokumen yang baru seperti pada formula (10).

$$d_{ek} = (w_{k1} + w_1', w_{k2} + w_2', \ldots, w_{kn} + w_n') \qquad (10)$$

E. Document Clustering

Vektor dokumen baru dengan bobot tambahan dari *corpus-based thesaurus* dan *dictionary-based thesaurus* selanjutnya dijadikan input pada algoritma K-Means untuk mendapatkan hasil *cluster* dokumen berdasarkan topik.

K-Means clustering adalah salah satu algoritma pengelompokan umum yang digunakan dalam *document clustering*. Langkah awal yang dilakukan adalah inisialisasi jumlah *cluster* (k) dan pemilihan dokumen yang dipilih secara random sebagai *centroid* pada masing-masing cluster.

Dokumen *d* direpresentasikan sebagai vektor d = {w_{d1}, w_{d2}, w_{d3},..., w_{di}} dengan w_{di} adalah bobot term *i* pada dokumen *d*, sedangkan *centroid* pada masing-masing *cluster* direpresentasikan sebagai vektor c = {w_{c1}, w_{c2}, w_{c3}, ..., w_{ci}} dengan w_{ci} adalah bobot term ke *i* pada *centroid c*. Jarak antara dokumen *d* ke *centroid c* dikalkulasi dengan *Euclidian distance* seperti pada formula (11).

$$distance_{(d,c)} = \sqrt{\sum_{i=1}^{n} (w_{di} - w_{ci})^2} \qquad (11)$$

Setiap dokumen kemudian dimasukkan ke dalam cluster dengan jarakterdekat. Selanjutnya nilai *centroid* pada masing-masing cluster diperbaharui berdasarkan nilai rata-rata bobot term w_{di} dari setiap dokumen *d* yang merupakan anggota *cluster c*. Nilai bobot term ke *i* pada centroid *c* dengan jumlah n_c anggotaakan dihitung berdasarkan formula (12).

$$w_{ci} = \frac{1}{n_c} \sum_{d=1}^{n_c} w_{di} \qquad (12)$$

Setelah didapatkan nilai *centroid* baru, jarak masing-masing dokumen terhadap centroid yang baru dihitung kembali menggunakan formula (11). Setiap dokumen kemudian dikelompokan kembali terhadap *cluster* berdasarkan jarak baru yang terdekat. Proses ini dilakukan pada setiap iterasi hingga anggota pada masing-masing cluster tidak mengalami perubahan.

IV. EKSPERIMEN

A. Data Uji

Data uji yang digunakan adalah 253 dokumen berita berbahasa Indonesia dengan 12 topik berbeda dari situs berita online www.kompas.com danwww.detik.com.Daftar topik dan jumlah dokumen berita pada masing-masing topik ditunjukkan pada Tabel 1. Sementara untuk *dictionary-based thesaurus*, digunakan Kamus Besar Bahasa Indonesia sebagai korpus kamus. Suatu kata dasar beserta seluruh definisinya diasumsikan sebagai satu dokumen pada korpus kamus.

TABEL 1
DAFTAR TOPIK PADA DATASET

ID	Topik	Jumlah Dokumen
1	Kasus sodomi Anwar Ibrahim	26
2	Kasus suap Artalyta (Ayin)	27
3	Bencana gempa bumi di China	19
4	FPI dalam insiden Monas	35
5	Konflik Israel-Palestina	25
6	Dua tahun tragedi lumpur Lapindo	21
7	Badai nargis di Myanmar	24
8	Pengunduran jadwal Pemilu	10
9	Perkembangan harga emas	18
10	Kiprah Khofifah dalam Pilkada Jatim	15
11	Kematian mantan presiden Soeharto	28
12	Festival Film Cannes	5
	Jumlah	**253**

B. Evaluasi

Evaluasi dilakukan dengan menggunakan F-Measure berdasarkan kelas atau topik pada dokumen yang telah diidentifikasi sebelumnya. Untuk setiap kelas i sejumlah n_i dokumen dan *cluster* j sejumlah n_j dokumen, nilai *recall* dan *precision* dihitung dengan formula (13).

$$Recall(i,j) = R_{ij} = n_{ij}/n_i$$
$$Precision(i,j) = P_{ij} = n_{ij}/n_j \quad (13)$$

F-Measure dihitung untuk setiap kelas i dan cluster j pada formula (14) sehingga didapatkan nilai keseluruhan F-Measure pada formula (15).

$$F_{ij} = \left(2 \times R_{ij} \times P_{ij}\right)/\left(R_{ij} + P_{ij}\right) \quad (14)$$

$$F = \sum_i \frac{n_i}{n} max\{F_{ij}\} \quad (15)$$

C. Hasil Eksperimen

Evaluasi dilakukan terhadap empat metode *document clustering*, yaitu:

1. K-Means tanpa tesaurus
2. K-Means dengan *corpus-based thesaurus*
3. K-Means dengan *dictionary-based thesaurus*
4. K-Means dengan kombinasi *corpus-based thesaurus* dan *dictionary-based thesaurus*

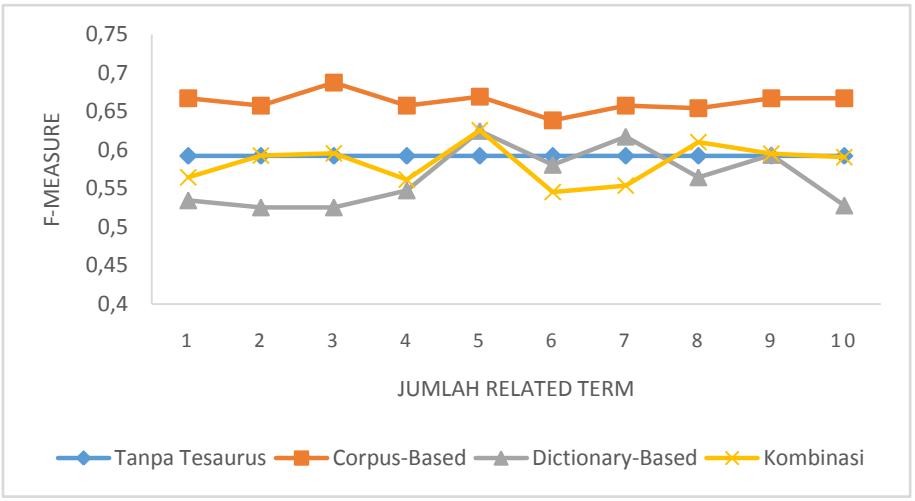
Jumlah cluster diset 12 sesuai dengan jumlah topik pada data uji. Nilai α ditentukan 10 sebagai perbandingan nilai rata-rata *document-term similarity* dari *corpus-based thesaurus* terhadap *dictionary based thesaurus*. Eksperimen dilakukan dengan variasi jumlah *top-ranked related terms* dari 1 hingga 10.

Pada eksperimen pertama, seperti yang ditunjukkan pada Gambar 1, hanya pembobotan dengan *corpus-based thesaurus* yang menunjukkan akurasi lebih baik daripada pembobotan tanpa tesaurus, sedangkan pembobotan dengan *dictionary-based thesaurus* dan dengan kombinasi kedua jenis tesaurus masih memberikan performa yang lebih buruk.

Hal ini merupakan indikasi bahwa similaritas term berdasarkan kamus belum bisa memberikan kontribusi positif dalam *document clustering*. Oleh karena itu, dilakukan identifikasi mengenai *top-ranked related terms* pada setiap dokumen berita berdasarkan *dictionary-based thesaurus*. Hasilnya, terdapat beberapa term yang terlalu sering muncul sebagai *top-ranked related terms* (TRRT) seperti yang ditunjukkan pada Tabel 2.

TABEL 2
TOP-RANKED RELATED TERMS PADA BANYAK DOKUMEN

Term pada TRRT	Jumlah Dokumen
Orang	237
Buat	210
Proses	166
Jadi	85
Negara	70
Pemerintah	39
Perintah	36
Barang	33
Daerah	32
Perkara	29



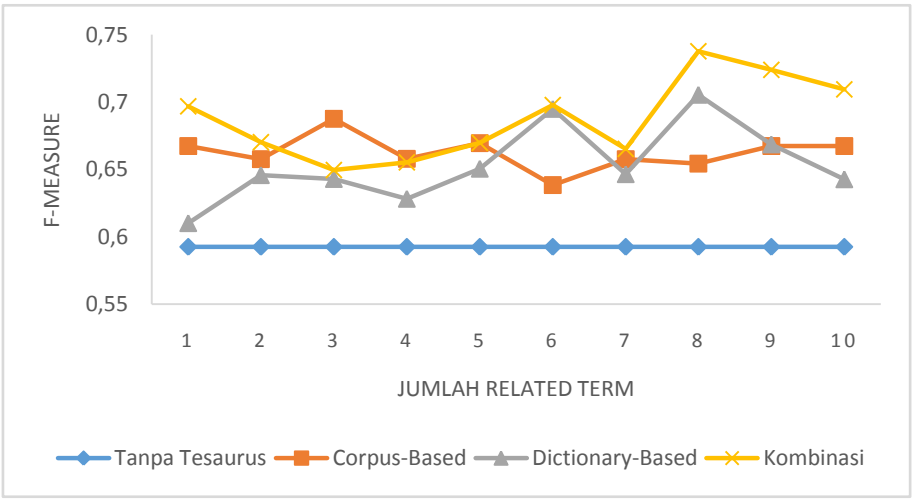
Gambar 2. Eksperimen 1 tanpa eliminasi *top-ranked related terms*

Selanjutnya pada eksperimen kedua, dilakukan eliminasi 5 term dengan frekuensi kemunculan terbanyak dari daftar*top-ranked related terms* pada masing-masing dokumen. Proses eliminasi ini didasarkan pada hipotesa bahwa term yang menjadi *top-ranked related terms* dari banyak dokumen merupakan penyebab *document clustering* menggunakan *dictionary-based thesaurus* memiliki akurasi yang lebih buruk.

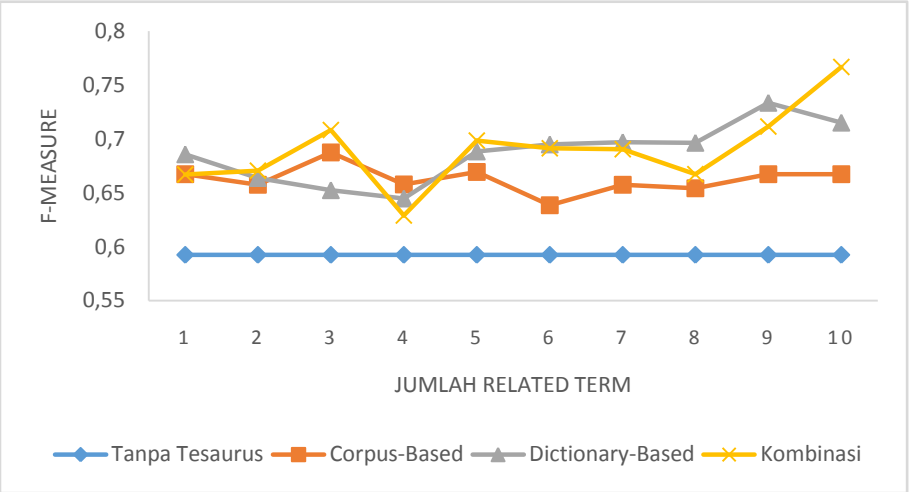
Hasil eksperimen kedua seperti yang ditunjukkan pada Gambar 3 menunjukkan efek positif penggunaan *dictionary-based thesaurus*.Ketiga metode pembobotan dengan tesaurus memiliki akurasi yang lebih baik daripada metode pembobotan tanpa tesaurus,sedangkan *corpus-based thesaurus* menunjukkan akurasi yang relatif lebih stabil daripada *dictionary-based thesaurus*. Sementara itu, pembobotan dengan kombinasi *corpus-based* dan *dictionary-based thesaurus* menunjukkan akurasi yang relatif lebih baik daripada metode lainnya.

Sebagai pembanding, pada eksperimen berikutnya dilakukan eliminasi terhadap 10 term dengan frekuensi kemunculan terbanyak dari daftar*top-ranked related terms*. Hasil eksperimen seperti pada Gambar 4 menunjukkan bahwa senada dengan eksperimen kedua, pada eksperimen ini *dictionary-based thesaurus*, baik yang digunakan sendiri ataupun dikombinasikan dengan *corpus-based thesaurus*, menunjukkan performa yang lebih baik daripada *corpus-based thesaurus* dan pembobotan tanpa tesaurus.

Di samping itu, hasil eksperimen kedua dan ketiga juga menunjukkan bahwa penggunaan dictionary-based thesaurus dengan eliminasi 10 term sedikit lebih baik daripada eksperimen kedua dengan eliminasi 5 term, sedangkan kombinasi corpus-based thesaurus dan dictionary-based thesaurus menunjukkan akurasi yang relatif sama baik dengan eliminasi 5 term maupun 10 term.



Gambar 3. Eksperimen 2 dengan eliminasi 5 *top-ranked related terms*



Gambar 4. Eksperimen 3 dengan eliminasi 10 *top-ranked related terms*

V. ANALISIS

Pada percobaan pertama, konsisten dengan penelitian-penelitian sebelumnya *corpus-based thesaurus*yang dibangun berdasarkan similaritas antar term mampu memberikan efek positif pada *document clustering*. Khususnya pada dokumen berita, *corpus-based thesaurus* sangat penting karena terdapat banyak nama tempat, nama orang, dan singkatan maupun istilah baru sehingga tesaurus yang dibangun berdasarkan*co-occurrence* term pada korpus dapat memperbaiki akurasi *document clustering*.

Sebaliknya, penggunaan *dictionary-based thesaurus* dalam pembobotan term tidak menunjukkan akurasi yang lebih baik dibandingkan dengan *document clustering* tanpa tesaurus. Hal ini kemungkinan disebabkan adanya beberapa term yang menjadi *top-ranked related terms* pada terlalu banyak dokumen. Akibatnya, penambahan bobot pada term yang demikian tidak hanya memperbesar nilai similaritas dokumen dalam topik yang sama, tetapi juga akan memperbesar nilai similaritas dokumen dalam topik yang berbeda. Hasil akhirnya, akurasi *document clustering*akan menurun sebab kemungkinan suatu dokumen dikelompokan dengan dokumen lain yang berbeda topik akan lebih besar.

Setelah dilakukan eliminasi 5 dan 10 term dengan frekuensi kemunculan terbanyak dari daftar*top-ranked related terms*, *dictionary-based thesaurus* mulai menunjukkan efek positif terhadap*document clustering*. Proses eliminasi ini memastikan term yang bobotnya ditambahkan sebagai bobot tambahan (*expanded weight*) merupakan term-term yang merepresentasikan dokumen terkait saja. Efeknya, setiap dokumen akan memiliki nilai similaritas yang lebih tinggi dengan dokumen pada topik yang sama saja.

Di samping itu, pada percobaan kedua, metode pembobotan dengan menggunakan kombinasi *corpus-based* dan *dictionary-based thesaurus* menunjukkan akurasi yang relatif lebih baik daripada ketiga metode lain. Hasil ini membuktikan bahwa penggunaan kombinasi kedua jenis tesaurus ini dapat memanfaatkan kelebihan dari *corpus-based thesaurus* yang memastikan bahwa similaritas konseptual antar term dapat dipertimbangkan serta kelebihan dari *dictionary-based thesaurus* yang memastikan bahwa term dengan similaritas tinggi memiliki makna yang berkaitan.

Isu yang perlu diteliti lebih lanjut terkait *dictionary-based thesaurus* adalah mekanisme pemilihan *top-ranked related terms*yang lebih tepat. Hal ini bertujuan agar *expanded weight* yang ditambahkan pada bobot original dapat memperbesar similaritas dokumen dengan dokumen lain pada topik yang sama sekaligus memperkecil similaritas dengan dokumen lain pada topik yang berbeda. Metode pertama adalah dengan mengeliminasi term yang terlalu sering muncul sebagai*top-ranked related terms*seperti yang dipakai pada penelitian ini. Di samping itu, perlu diidentifikasi kemungkinan dilakukannya *feature selection*sebelum penghitungan similaritas antar term sehingga *top-ranked related terms* yang dihasilkan akan lebih baik.

VI. KESIMPULAN

Pembobotan term dalam K-Means *document clustering* menggunakan kombinasi *corpus-based* dan *dictionary-based thesaurus* menunjukkan akurasi yang lebih baik daripada ketiga metode lain yang diuji, yaitu K-Means

tanpa tesaurus, K-Means dengan *corpus-based thesaurus*, dan K-Means dengan *dictionary-based thesaurus*. Hal ini disebabkan pembobotan term tidak hanya memperhatikan frekuensi term yang identik secara leksikal, tetapi juga mempertimbangkan hubungan similaritas konseptual antar term. Di sisi lain, dokumen berita juga rentan terhadap term-term yang muncul bersamaan namun sebenarnya tidak memiliki makna yang berkaitan. *Dictionary-based thesaurus* dipakai sebagai solusi untuk masalah tersebut. Oleh karena itu, penggunaan kombinasi*corpus-based* dan *dictionary-based thesaurus* dapat memberikan akurasi *document clustering* yang lebih baik. Untuk penelitian selanjutnya, perlu dilakukan identifikasi mengenai mekanisme pemilihan *top-ranked related terms* yang lebih tepatpada*dictionary-based thesaurus* sehingga *expanded weight* yang ditambahkan pada bobot original dapat memperbesar similaritas dokumen dengan dokumen lain pada topik yang sama sekaligus memperkecil similaritas dengan dokumen lain pada topik yang berbeda.

VII. REFERENSI

[1] S. Staab dan G. Stumme, "Wordnet Improves Text Document Clustering," dalam *Proceeding of the SIGIR 2003 Semantic Web Workshop*, 2003, hal. 541-544.

[2] S. L. Bang, J. D. Yang dan H. J. Yang, "Hierarchical document categorization with k-NN and concept-based thesauri," *Information Processing and Management*, vol. 42, hal. 387-406, 2006.

[3] C. H. Li, J. C. Yang dan S. C. Park, "Text categorization algorithms using semantic approaches, corpus-based thesaurus and WordNet," *Expert System with Applications*, vol. 39, hal. 765-772, 2012.

[4] J. Xu dan W. B. Croft, "Query expansion using local and global document analysis," dalam *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval*, 1996.

[5] H. Xu dan B. Yu, "Automatic thesaurus construction for spam filtering using revised back propagation neural network," *Expert Systems with Applications*, vol. 37, hal. 18-23, 2010.

[6] C. H. Li, W. Song dan S. C. Park, "An automatically constructed thesaurus for neural network based document categorization," *Expert Systems with Applications*, vol. 36, hal. 10969-10975, 2009.

[7] R. Steinberger , B. Pouliquen dan J. Hagman, "Cross-Lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC," dalam *Third International Conference CICLing*, Mexico City, 2002.

[8] A. Z. Arifin, I. P. A. Kerta Mahendra dan H. T. Ciptaningtyas, "Enhanced Confix Stripping Stemmer and Ants Algorithm for Classifying News Document In Indonesian Language," *The 5th International Conference on Information & Communication Technology and Systems*, 2008.