Ekstraksi Frasa Kunci pada Penggabungan Klaster berdasarkan *Maximum-Common-Subgraph*

Adhi Nurilham[1], Diana Purwitasari[2], Chastine Fatichah[2]

***Abstract*—Document clustering based on topic similarities helps users in searching from a collection of scientific articles. Topic labels are necessesary for describing subjects of the document clusters. Clusters with related subjects or contextual similarities can be merged to produce more descriptive labels. Relations between those words in one context can be modelled as a graph. Instead of single word, this paper proposed cluster labeling of phrases from scientific articles with cluster merging based on graph. The proposed method begins with K-Means++ for clustering the scientific articles. Then, the candidates of word phrases from document clusters are extracted using Frequent Phrase Mining which inspired by Apriori algorithm. Each cluster result has a representation graph from those extracted word phrases. An indicator value from each graph shows any similarities of graph structures which is calculated with Maximum Common Subgraph (MCS). Those clusters are merged if there are any structure similarities between them. Topic labels of clusters are keyword phrases extracted from a representation graph of previous merged clusters using TopicRank algorithm. The merging process which becomes the contribution of this paper is considering topic distribution within clusters for phrase extraction. The proposed method evaluation is performed based on topic coherence of the merged clusters label. The results show that proposed method can improve topic coherence on the merged clusters with MCS graph size percentage as the key factor. Further observation shows that merged cluster labels consistent to MCS graph.**

***Intisari*—Klasterisasi dokumen berdasarkan kemiripan topik memudahkan pencarian pada koleksi artikel ilmiah yang banyak dan pelabelan klaster diperlukan untuk memberikan gambaran topik bahasan dalam artikel. Beberapa kelompok dokumen yang masih memiliki kemiripan kontekstual atau topik bahasan perlu digabung untuk menghasilkan label klaster lebih baik. Relasi dari kata-kata dalam satu konteks dapat direpresentasikan sebagai model graf. Makalah ini mengusulkan pelabelan kelompok artikel ilmiah dengan kontribusi penggabungan klaster berbasis graf untuk memberikan label topik yang lebih representatif. Usulan metode diawali dari pengelompokan artikel ilmiah berdasarkan topik dengan *K-Means++*. Kemudian, kandidat frasa kata dari kelompok dokumen hasil klasterisasi diekstraksi menggunakan adopsi algoritme Apriori yaitu *Frequent Phrase Mining*. Setiap klaster memiliki representasi graf dari kandidat frasa kata. Dari graf tersebut dihitung nilai indikator sebagai penanda adanya struktur *node* sama dengan *Maximum Common Subgraph* (MCS). Penggabungan klaster dilakukan jika terdapat kesamaan struktur graf representasi. Label topik bahasan adalah frasa kunci sebagai hasil ekstraksi dari graf klaster gabungan berdasarkan distribusi topik menggunakan algoritme *TopicRank*. Evaluasi usulan metode dilakukan berdasarkan koherensi topik label klaster yang dihasilkan. Hasil pengujian menunjukkan bahwa usulan metode dapat meningkatkan koherensi topik pada klaster hasil penggabungan dengan faktor yang memengaruhi, yaitu persentase ukuran graf MCS terhadap graf klaster. Pengamatan lebih lanjut menunjukkan bahwa terdapat konsistensi label klaster hasil penggabungan terhadap isi graf MCS.**

***Kata Kunci*— pelabelan klaster, penggabungan klaster, *Frequent Phrase Mining*, *Maximum Common Subgraph*, *TopicRank*.**

I. PENDAHULUAN

Banyaknya artikel ilmiah yang diterbitkan per tahun menjadi tantangan peneliti untuk mengikuti perkembangan teknologi terbaru pada bidangnya. Klasterisasi dokumen berdasarkan kemiripan topik mengatasi permasalahan tersebut sehingga memudahkan peneliti dalam mencari kelompok artikel ilmiah [1], maupun pencarian pada dokumen medis dengan identifikasi asosiasi penyakit secara konseptual [2]. Klasterisasi juga digunakan dalam pengelompokan kalimat untuk peringkasan dokumen berita [3], [4], dan mengetahui tren informasi [5]. Pelabelan klaster diperlukan untuk memudahkan pengguna memahami topik yang dibahas, misalnya pemanfaatan konsep hierarki dalam kombinasi proses pengklasteran dan pelabelan [6]. Label klaster dapat dihasilkan dengan pemodelan topik berbasis statistik untuk memperjelas topik laten/implisit dalam kelompok dokumen [7] atau menggunakan asumsi ditribusi multinomial [8]. Oleh karena itu, ekstraksi topik menjadi tahap penting dalam pelabelan klaster [9], [10].

Kata tunggal sebagai label klaster dianggap kurang intuitif sehingga frasa kata diutamakan karena lebih deskriptif bagi representasi topik dengan gabungan pendekatan linguistik serta statistik [11]. Klasterisasi memberikan hasil kurang optimal jika beberapa kelompok dokumen masih memiliki kemiripan kontekstual seperti sinonim, polisemi, atau ambiguitas [12]. Hal tersebut disebabkan oleh ketidakadaan informasi semantik atau mengabaikan hubungan antar kata sehingga diperlukan penggabungan kelompok dokumen sebelum pelabelan klaster. Semantik antar kata diabaikan karena asumsi suatu kata tidak bergantung pada kata lain dalam representasi klaster dokumen dengan fitur kata sebagai *bag-of-words*. Namun, relasi kata dapat meningkatkan kinerja proses klasterisasi artikel ilmiah [13].

[1] *Mahasiswa, Departemen Informatika, Fakultas Teknologi Informasi dan Komunikasi (FTIK), Institut Teknologi Sepuluh Nopember (ITS), Jln. Teknik Kimia,Gedung Informatika, Kampus ITS, Sukolilo, Surabaya, Jawa Timur, 60111 Indonesia (tlp: 031-5939214; fax: 031-5913804; email: adhi16@mhs.if.its.ac.id)*

[2] *Dosen, Departemen Informatika, FTIK, ITS, Jln. Teknik Kimia,Gedung Informatika, Kampus ITS, Sukolilo, Surabaya, Jawa Timur, 60111 Indonesia (tlp: 031- 5939214; fax: 031-5913804; email: diana@if.its.ac.id, chastine@if.its.ac.id)*

Model graf menjadi alternatif dalam representasi informasi semantik pada teks [14]. Selain memetakan relasi antar teks seperti kata, frasa, atau kalimat, model graf dapat menyimpan informasi kemunculan bersama antar kata [15]. Terdapat peningkatan akurasi dengan penggunaan model graf sebagai representasi teks dalam permasalahan klasifikasi dokumen [16]. Representasi teks dengan model graf tidak memerlukan pengetahuan spesifik tentang linguistik atau domain tertentu, tetapi tetap memungkinkan adanya integrasi pengetahuan eksternal seperti *Wordnet* [17].

Makalah ini memberikan kerangka kerja baru dengan menggabungkan beberapa metode yang telah ada untuk menyelesaikan permasalahan pelabelan kelompok artikel ilmiah. Usulan metode adalah pelabelan dalam bentuk frasa dengan penggabungan klaster berbasis graf untuk menghindari label klaster yang sulit dibedakan secara konseptual. Setelah klasterisasi artikel ilmiah dilakukan menggunakan algoritme *K-Means++* (proses P2), ekstraksi frasa menyaring kata yang tidak signifikan secara konseptual menggunakan *Frequent Phrase Mining* (FPM) (proses P3) [18]. Representasi graf setiap klaster dibentuk dengan *word2vec* (proses P4) [19]. Penggabungan klaster dilakukan jika terdapat kemiripan yang diidentifikasi dengan pendekatan berbasis graf *Maximum Common Subgraph* (MCS) (proses P5) [20]. Setiap klaster baru terbentuk diberikan label frasa dengan algoritme *TopicRank* (proses P6) [21]. Evaluasi dilakukan melalui perbandingan nilai koherensi topik dari label klaster yang dihasilkan dengan dan tanpa penggabungan klaster. Nilai koherensi topik menentukan tingkat interpretabilitas label yang berkorelasi positif terhadap penilaian manusia [22].

Bahasan selanjutnya terbagi dalam beberapa bagian. Bagian kedua menjelaskan tentang penelitian terdahulu yang berhubungan dengan usulan metode. Kemudian bagian ketiga menguraikan tahap P2, P3, P4, P5, dan P6, seperti diperlihatkan pada Gbr. 1, untuk menghasilkan label topik yang lebih representatif. Bagian keempat menjelaskan tentang skenario, persiapan data, dan analisis uji coba. Sebagai penutup, bagian kelima berisi kesimpulan dan juga menjelaskan penelitian lanjutan yang akan dilakukan.

II. PELABELAN KLASTER DAN KLASTERISASI TOPIK

Pelabelan klaster dapat dirumuskan sebagai ekstraksi frasa kunci pada multi-dokumen dengan (i) ekstraksi frasa kandidat dan (ii) pemilihan frasa kandidat. Ekstraksi frasa kandidat melalui pendekatan heuristik memperhitungkan frasa dengan label *part-of-speech* tertentu [23], mengambil *n-gram* dari korpus sumber [24] atau korpus eksternal [25]. Pendekatan heuristik memberikan *recall* yang baik, tetapi jumlah kandidat frasa kunci pada multi-dokumen akan sangat banyak [26]. Pemilihan frasa kandidat dapat menggunakan pendekatan statistik, berbasis graf, atau klasterisasi topik. Pemilihan frasa dengan pendekatan statistik antara lain dengan pembobotan *Term Frequency – Inverse Cluster Frequency* (TF-ICF) [27], menggunakan perhitungan *Markov Chain* [11], dan pemberian nilai frasa kandidat berdasarkan *Pointwise Mutual Information* (PMI) [28]. Pemilihan frasa kandidat berbasis graf sebagai representasi teks seperti *TextRank* [23] mengadopsi algoritme *PageRank* [29] atau *CollabRank* sebagai perbaikan *TextRank* dengan memperhitungkan dokumen lain yang memiliki kemiripan [30]. Pendekatan klasterisasi topik memiliki tujuan agar frasa kunci yang dihasilkan mencakup topik utama dalam kelompok dokumen seperti *KeyCluster* [31]. Pendekatan klasterisasi topik memiliki kekurangan karena seluruh topik atau kelompok frasa dianggap memiliki tingkat kepentingan yang sama, sehingga pada *TopicRank* diidentifikasi distribusi topik untuk mengetahui tingkat kepentingan topik [32]. Namun, terdapat permasalahan efisiensi di *TopicRank* yaitu *TextRank* yang dilakukan pada sejumlah topik. Permasalahan tersebut diatasi dengan menyematkan distribusi topik pada bobot tunggal [21]. Pemilihan frasa dengan pendekatan klasterisasi topik memberikan hasil terbaik dibanding pendekatan lainnya untuk data artikel ilmiah [26].

Pada makalah ini dilakukan pelabelan klaster artikel ilmiah dengan penggabungan klaster untuk menghindari label yang memiliki kemiripan. Pada umumnya, metode penggabungan klaster menggunakan representasi *Vector Space Model* (VSM) dan kemiripan klaster menggunakan perhitungan kemiripan (similaritas/*similarity*) seperti *Euclidean distance* [33] atau *cosine similarity*. Namun, fitur kata yang sangat banyak menyebabkan nilai similaritas berbasis VSM menjadi kurang signifikan [34]. Perhitungan similaritas berbasis VSM juga mengabaikan informasi semantik pada relasi antar kata. Model graf sebagai representasi klaster dokumen menjadi alternatif VSM. Pengukuran similaritas berbasis graf dengan sebuah substruktur MCS menggambarkan kesamaan struktur antara dua graf. Perhitungan dilakukan berdasarkan jumlah *node* dan *edge* pada substrukur MCS [35] serta adanya penambahan bobot *edge* [20].

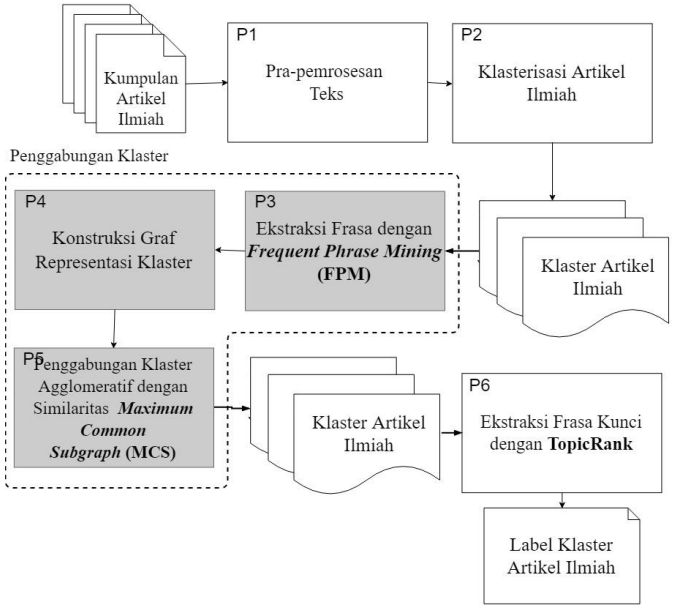
III. USULAN METODE

Tahapan untuk usulan metode pelabelan klaster ditunjukkan pada Gbr. 1, yaitu prapemrosesan teks, ekstraksi frasa, konstruksi graf representasi klaster, penggabungan klaster, dan ekstraksi frasa kunci. Usulan metode diawali dengan tahap prapemrosesan teks untuk menghilangkan kata yang tidak diperlukan. Prapemrosesan teks terdiri atas konversi ke huruf kecil, penghilangan tanda baca, serta penghilangan *stopwords*. Pengklasteran *K-Means++* dengan pembobotan kata berbasis jaringan syaraf tiruan *word2vec* [19] dan ukuran kedekatan *cosine similarity* diimplementasikan pada dokumen hasil prapemrosesan. Jumlah klaster ditentukan dengan nilai *silhouette* tertinggi yang dihitung dari kemiripan data intraklaster dan ketidakmiripan data interklaster [36]. Fokus pada makalah ini, yaitu pelabelan klaster dengan penggabungan klaster, dijelaskan pada bagian-bagian berikut.

A. Klasterisasi Artikel Ilmiah

Dokumen dalam koleksi umumnya direpresentasikan sebagai matriks dengan baris adalah dokumen dan kolom adalah kata-kata yang muncul. Isi matriks dokumen-kata tersebut adalah frekuensi kemunculan kata dalam dokumen. Representasi matriks diubah dengan proses tambahan transformasi frekuensi kata menjadi frekuensi dan relasi kata menggunakan *word2vec* [19], sehingga matriks dokumen-kata

tidak lagi berisi frekuensi kemunculan kata, tetapi berisi bobot yang menggambarkan frekuensi dan relasi kata pada suatu dimensi tertentu. Kemudian, pengklasteran *K-Means++* dilakukan pada matriks representasi baru dengan ukuran kedekatan *cosine similarity*. Jumlah klaster ditentukan berdasarkan analisis *silhouette* yang mencari jumlah klaster dengan nilai *silhouette* tertinggi [36].



Gbr. 1 Alur proses metode usulan.

Keluaran dari *word2vec* adalah matriks kata yang memiliki baris sejumlah banyaknya kata dan kolom sejumlah dimensi yang telah ditentukan. Tiap baris pada matriks kata merepresentasikan vektor kata. Matriks dokumen-kata dibuat dengan menghitung rata-rata kumpulan vektor kata yang terdapat pada setiap dokumen. Matriks dokumen-kata yang telah dibuat menjadi masukan pada pengklasteran *K-Means++*.

B. Ekstraksi Frasa dengan Frequent Phrase Mining (FPM)

Setiap kata dalam klaster memiliki potensi menjadi label. Untuk mengurangi jumlah kata tersebut, ekstraksi frasa topik dilakukan pada setiap klaster dengan algoritme FPM) [18]. Setiap frasa topik harus memenuhi kriteria *frequent itemset mining* yang ditetapkan algoritme Apriori [37]. Dalam algoritme Apriori, jika frasa *P* tidak sering muncul, maka superfrasa dari frasa *P* yang telah ditambahkan suatu kata juga tidak akan sering muncul.

Tujuan algoritme Apriori adalah mencari asosisasi suatu *item* yang disebut *itemset*. Pada FPM, *item* direpresentasikan oleh kata dan *itemset* direpresentasikan oleh frasa kata. Parameter utama dalam menentukan *frequent itemset* adalah *support* yang merupakan nilai batas minimal kemunculan *item*. Pada makalah ini, parameter tersebut merupakan batas frekuensi kemunculan kata dan frasa kata, dan dinotasikan dengan ε. Jika terdapat frasa kata *P* yang tersusun atas tiga kata dengan notasi $[t_1, t_2, t_3] \in P$, untuk suatu nilai ambang ε, maka $freq(t_1), freq(t_2), freq(t_3) \geq \varepsilon$ yang menyatakan minimal kemunculan kata. Kemudian, untuk superfrasa atau gabungan kata dari frasa *P*, maka syarat $freq(t_1, t_2), freq(t_1, t_3) \geq \varepsilon$ juga harus terpenuhi.

C. Konstruksi Graf Representasi Klaster

Model graf merupakan alternatif yang banyak digunakan untuk merepresentasikan semantik teks. Model graf dibuat untuk setiap klaster artikel ilmiah c_x. Graf dinotasikan dengan $G_x = (V_x, E_x, W_x^v, W_x^e)$ dengan kumpulan *vertex* atau *node* V_x, kumpulan *edge* E_x, kumpulan bobot *vertex* W_x^v, dan kumpulan bobot *edge* W_x^e yang menjadi representasi klaster c_x.

Kata yang menyusun frasa hasil FPM pada klaster c_x menjadi *vertex* yang disimpan dalam V_x. Jika $vx_i \in V_x$, merepresentasikan *vertex* untuk kata t_i pada klaster c_x, maka *edge* $ex_{ij} \in E_x$ dan bobot $wx_{ij}^e \in W_x^e$ dibentuk dari hasil *cosine similarity* antara vektor *word2vec* t_i serta vektor *word2vec* t_j.

D. Penggabungan Klaster dengan Maximum Common Subgraph (MCS)

Setelah graf dibuat untuk setiap klaster artikel ilmiah, perhitungan kemiripan klaster berbasis graf dapat dilakukan menggunakan MCS [20]. MCS adalah substruktur umum antara dua graf. Struktur MCS antara dua graf representasi klaster dapat digunakan sebagai pengukuran similaritas klaster.

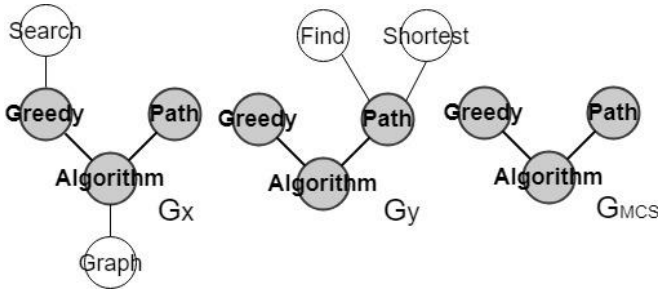
Misalnya terdapat graf $Gx = (Vx, Ex, Wx^v, Wx^e)$ dan graf $Gy = (Vy, Ey, Wy^v, Wy^e)$ yang merepresentasikan klaster c_x dan klaster c_y secara berurutan. Graf Gx dan Gy memiliki substruktur graf MCS yang merepresentasikan kesamaan struktur graf, $Gmcs = (Vmcs, Emcs, Wmcs^v, Wmcs^e)$.

Pada Gbr. 2 terlihat contoh graf MCS yang dihasilkan graf Gx dan Gy. Graf Gx memiliki lima *node* dan graf Gy juga memiliki lima *node*. Kesamaan substruktur graf Gx dan Gy atau $Gmcs$ ada pada *node-node* "*greedy*, *path*, *algorithm*". Pembuatan graf MCS dapat dilihat pada Algoritme 1. Penjelasan notasi yang digunakan pada makalah ini disajikan pada Tabel I.

Algoritme 1: Membuat *Maximum Common Graph*
Masukan: Graf Gx dan Gy
Keluaran: Graf $Gmcs$

1. Cari kumpulan *vertex* yang sama antara Gx dan Gy, dan jadikan kumpulan *vertex* tersebut sebagai *vertex* dalam $Gmcs$.
2. Buat daftar kombinasi pasangan *vertex* pada kumpulan *vertex* $Gmcs$.
3. Jika pasangan *vertex* memiliki *edge* di Gx dan Gy, maka sebuah *edge* akan dibuat pada $Gmcs$ yang menghubungkan pasangan *vertex* tersebut. Bobot *edge* dihitung dengan (1).
4. Ulangi langkah 3 untuk seluruh kombinasi pasangan *vertex*.

Setelah graf MCS antara Gx dan Gy diperoleh, maka nilai similaritas antara kedua graf dihitung dengan (4). Kemiripan graf dipengaruhi oleh perhitungan kemiripan antar *vertex* (2) dan *edges* (3) serta bobot kata dalam graf MCS (1). Pada penggabungan klaster, relasi kata pada graf MCS merepresentasikan konteks dari similaritas antar klaster.



Gbr. 2 Contoh Graf *Maximum Common Subgraph* (MCS).

$$wmcs_{ij}^{e} = \frac{\min\left(wx_{mn}^{e},\ wy_{pd}^{e}\right)}{\max\left(wx_{mn}^{e}, wy_{pd}^{e}\right)},$$

$$vmcs_i = vx_m = vy_p, vmcs_j = vx_n = vy_d \quad (1)$$

$$VSim(Gx, Gy) = \frac{|Vmcs|}{\max(|Vx|,|Vy|)} \quad (2)$$

$$ESim(Gx, Gy) = \frac{\sum_{\forall Wmcs^e} wmcs_{ij}^{e}}{\max(|Ex|,|Ey|)} \quad (3)$$

$$Sim(Gx, Gy) = \alpha * VSim(Gx, Gy) + (1 - \alpha) * ESim(Gx, Gy). \quad (4)$$

Setelah graf dibuat untuk setiap klaster, matriks similaritas klaster artikel ilmiah dibuat dengan menggunakan perhitungan similaritas MCS antar graf klaster. Matriks similaritas klaster digunakan sebagai masukan penggabungan klaster secara aglomeratif seperti pada Algoritme 2. Nilai batas minimum similaritas min_{sim} digunakan untuk membatasi penggabungan klaster. Nilai batas min_{sim} diperoleh melalui perhitungan selisih antara nilai rata-rata dan standar deviasi dari seluruh nilai similaritas antar graf klaster [34].

Algoritme 2: Penggabungan Klaster secara aglomeratif
Masukan: Matriks Similaritas Klaster
Keluaran: Klaster Gabungan

1. Cari pasangan klaster *a* dan *b* yang memiliki nilai similaritas tertinggi pada matriks similaritas klaster.
2. Jika nilai similaritas klaster *a* dan *b* lebih dari min_{sim} nilai batas yang ditentukan, maka klaster *a* dan *b* digabung menjadi klaster *c*.
3. Perbarui nilai similaritas antara klaster *c* dengan klaster lainnya.
4. Ulangi langkah 1-3 untuk melakukan penggabungan klaster.

E. Ekstraksi Frasa Kunci dengan TopicRank

Ekstraksi frasa kunci dilakukan pada klaster-klaster hasil penggabungan dengan menggunakan *TopicRank*. *TopicRank* memanfaatkan struktur graf representasi teks dan distribusi topik dalam ekstraksi frasa kunci. Awalnya pada *TopicRank*, *TextRank* harus dijalankan sebanyak jumlah topik yang ada [32]. Lalu, *TopicRank* telah dikembangkan sehingga hanya perlu menjalankan *TextRank* satu kali dengan menyematkan distribusi topik pada bobot tunggal [21]. Pada makalah ini, *TopicRank* dengan penyematan distribusi topik pada bobot tunggal digunakan untuk tahap ekstraksi frasa kunci.

Pertama-tama, distribusi topik dari klaster artikel ilmiah didapatkan dengan metode *Latent Dirchlet Allocation* (LDA) pada kumpulan dokumen di klaster *d* [27]. Metode LDA menghasilkan model topik $Z = [z_1, \ldots\ldots, z_k]$, dengan z_k merupakan topik ke-*k*. Model topik Z memiliki dua jenis probabilitas topik yang direpresentasikan dalam bentuk vektor, yaitu probabilitas kata terhadap topik dan probabilitas topik terhadap klaster dokumen seperti pada (5).

$$wx_i^v = \frac{\vec{P}(vx_i \mid Z)\ .\ \vec{P}(Z \mid d)}{||\vec{P}(vx_i \mid Z)||\ .\ ||\vec{P}(Z \mid d)||} \quad (5)$$

TABEL I
DESKRIPSI NOTASI PERSAMAAN

No	Notasi	Deskripsi
1	Gx	graf representasi klaster ke-*x*
2	Vx	kumpulan *vertex* pada graf *x*
3	Ex	kumpulan *edge* pada graf *x*
4	$wmcs_{ij}^{e}$	bobot *edge* antara vertex *i* dan *j* pada graf *Maximum Common Subgraph* (MCS)
5	α	koefisien tingkat kepentingan *vertex* pada pengukuran similaritas graf
6	vx_i	*vertex* ke-*i* pada graf *x*
7	wx_i^v	bobot *vertex i* pada graf *x*
8	wx_{ij}^{e}	bobot *edge* antara *vertex i* dan *vertex j* pada graf *x*
9	Z	model topik *Z*
10	z_k	topik ke-*k*
11	d	kumpulan dokumen pada klaster artikel ilmiah
12	$\vec{P}(v_i \mid Z)$	vektor probabilitas kata pada *vertex i* terhadap model topik Z
13	$\vec{P}(Z \mid d)$	vektor probabilitas model topik Z terhadap klaster dokumen *d*
14	$P(v_i \mid z_k)$	probabilitas kata pada *vertex i* terhadap topik *k*
15	$P(z_k \mid d\ )$	probabilitas topik *k* pada klaster dokumen *d*
16	$TPR_z(vx_i)$	nilai *TopicRank* pada *vertex i* dengan model topik *Z pada graf* x
17	δ	*damping factor*, merepresentasikan probabilitas *random surfer* untuk pergi ke *vertex* acak
18	$In(vx_i)$	kumpulan *vertex* yang mengarah ke *vertex i* pada graf *x*
19	$Out(vx_j)$	kumpulan *vertex* dengan *vertex j* mengarah pada graf *x*

Terdapat graf klaster $Gx = (Vx, Ex, Wx^v, Wx^e)$ dengan vertex $vx_i \in Vx$, pada kata yang menjadi vertex dalam graf memiliki $\vec{P}(vx_i \mid Z) = [\ P(vx_i \mid z_1), \ldots, P(vx_i \mid z_k)\ ]$ sebagai vektor probabilitas kata terhadap topik Z, dengan $P(vx_i \mid z_k)$ merupakan probabilitas kata pada vertex ke-*i* terhadap topik ke-*k* dalam graf klaster *x*. Vektor probabilitas topik terhadap suatu klaster berisi dokumen d yang dinotasikan dengan $\vec{P}(Z \mid d) = [\ P(z_1 \mid d), \ldots, P(z_k \mid d\ )]$, dengan $P(z_k \mid d)$ probabiltas topik ke-*k* terhadap kumpulan dokumen *d* di klaster *x*, sehingga bobot vertex vx_i dinotasikan dengan wx_i^v dihitung dengan (6).

$$a = \frac{wx_i^v}{\sum_{k \in |Wx^v|} wx_k^v} \quad (6)$$

$$b = \sum_{vx_j \in In(vx_i)} \frac{wx_{ji}^e}{\sum_{vx_k \in Out(vx_j)} wx_{jk}^e} TPR_z(vx_j) \quad (7)$$

$$TPR_z(vx_i) = (1-\delta) * a + \delta * b \quad (8)$$

Setelah bobot tiap *vertex* pada graf *x* dihitung, nilai skor *TopicRank* dapat dihitung. Nilai skor *TopicRank vertex* $vx_i \in Vx$ pada model topik *Z* dinotasikan dengan $TPR_Z(vx_i)$, dihitung dengan (8).

Pada (8), δ merupakan koefisien *damping*, yang memiliki nilai antara 0 dan 1. Koefisien *damping* mewakili kemungkinan loncatan dari sebuah *vertex* ke *vertex* acak. Dalam konteks penelusuran *web*, koefisien *damping* menggambarkan probabilitas, sebesar δ, pengguna yang disebut *random surfer* untuk memilih *link* yang tersedia pada halaman tersebut, dan probabilitas, sebesar $(1-\delta)$, *random surfer* untuk pergi ke halaman *web* yang bena-benar acak. Konsep implementasi koefisien *damping* dapat disebut juga *Random Surfer Model* [29]. Konvergensi tercapai jika nilai $TPR_z(vx_i)$ sudah tidak banyak mengalami perubahan. Kandidat frasa diberi skor sesuai dengan akumulasi skor *TopicRank* dari kata-kata penyusunnya. Setelah itu, *top-n* kandidat frasa akan terpilih menjadi label klaster artikel ilmiah.

F. Koherensi Topik

Koherensi topik label klaster mengukur tingkat keterkaitan kata-kata label dalam merepresentasikan topik bahasan dari dokumen dalam klaster [22]. Koherensi topik label klaster yang baik menandakan sebuah topik dapat diinterpretasi dengan mudah oleh penilaian pengguna. Evaluasi kinerja usulan dalam makalah ini dilakukan dengan pengujian koherensi topik antara klaster asli (sebelum digabung) dan klaster gabungan.

Sebuah kumpulan label dalam bentuk frasa pada klaster dapat dinotasikan dengan $L = [lb_1, \ldots., lb_p]$. Kumpulan kata $T = [t_1, \ldots., t_q]$ didapatkan dari kata unik penyusun seluruh label frasa pada kumpulan label klaster *L*. Vektor konteks kata t_i yang dinotasikan dengan $\overrightarrow{vt_i}$ memiliki ruang dimensi sebesar jumlah kata *q*. Elemen vektor $\overrightarrow{vt_i}$ ke-*j* yang dinotasikan oleh vt_{ij} dihitung dengan (11), dengan ϵ adalah nilai *bias* yang ditentukan secara manual. $P(t_i)$ merupakan probabilitas kemunculan kata t_i pada rentang jendela kata (*sliding window*) yang dinotasikan dengan n_{win}. $P(t_i, t_j)$ merupakan probabilitas kemunculan bersama kata t_i dan t_j pada n_{win}. Probabilitas kemunculan kata dihitung berdasarkan korpus Wikipedia yang terdiri atas artikel Wikipedia dari tahun 2009. Nilai n_{win} dan ϵ yang digunakan adalah 110 dan 1, sesuai dengan penelitian sebelumnya [22].

$$PMI(t_i, t_j) = \log \frac{P(t_i,t_j)+\epsilon}{P(t_i).P(t_j)} \quad (9)$$

$$NPMI(t_i, t_j) = \frac{PMI(t_i,t_j)}{-\log(P(t_i,t_j)+\epsilon)} \quad (10)$$

$$vt_{ij} = NPMI(t_i, t_j) \quad (11)$$

Jika kumpulan vektor konteks kata $VT = [\overrightarrow{vt_1}, \ldots., \overrightarrow{vt_q}]$, maka kombinasi pasangan vektor konteks kata disimpan dalam $qC_2(VT)$. Koherensi topik label klaster *L* dihitung menggunakan (12), dengan $CosSim(S)$ merupakan *cosine similarity* dari pasangan vektor konteks kata *S* .

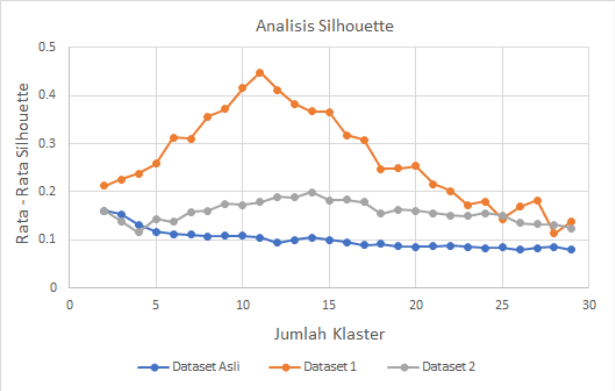
$$Coh(L) = \frac{\sum_{S \in qC_2(VT)} CosSim(S)}{|qC_2(VT)|} \quad (12)$$

IV. HASIL DAN PEMBAHASAN

A. Persiapan Data

Data set diperoleh dari basis data publikasi ilmiah *AMiner* pada *data set Citation* (2018). *Data set* terdiri atas ±12.200 artikel ilmiah yang diambil dari jurnal *IEEE Transaction on Computers* beserta artikel-artikel kutipannya. Seterusnya, *data set* tersebut disebut *Data set* Asli. Observasi awal dilakukan dengan menggunakan analisis *silhouette* pada *Data set* Asli [36]. Nilai *silhouette* pada data merepresentasikan tingkat kemiripan data intraklaster dan ketidakmiripan data interklaster dalam rentang -1 dan 1. Analisis *silhouette* mencari jumlah klaster yang menghasilkan rata-rata *silhouette* tertinggi sebagai jumlah klaster optimal.

Hasil analisis *silhouette* observasi awal menunjukkan bahwa *Data set* Asli memiliki banyak derau sehingga tidak memiliki jumlah klaster yang optimal seperti pada Gbr. 3. Artikel ilmiah yang bersifat derau dapat diketahui dengan melihat nilai *silhouette*-nya. Pada jumlah klaster sebelas, terdapat 51,16% data artikel ilmiah di *Data set* Asli yang memiliki nilai *silhouette* kurang dari 0,1. Oleh karena itu, perlu dilakukan penyaringan data pada *Data set* Asli, sehingga *data set* dapat digunakan untuk evaluasi usulan metode.



Gbr. 3 Analisis *Silhouette* pada *Data set* Asli, *Data set* 1, dan *Data set* 2.

Berikut adalah penyaringan data guna mendapatkan dua *data set* untuk meningkatkan variasi karakteristik *data set* dalam evaluasi. Cara Pertama adalah dengan mengambil lima puluh artikel ilmiah dengan nilai *silhouette* tertinggi dari tiap klaster hasil klasterisasi *Data set* Asli dengan jumlah klaster sebelas secara acak. Sebagai catatan, *Data set* Asli tidak memiliki jumlah klaster optimal. Penyaringan data tersebut menghasilkan *data set* dengan 550 artikel ilmiah yang disebut sebagai *Data set* 1.

TABEL II
HASIL RATA-RATA KOHERENSI TOPIK SKENARIO 1

No. Pengujian	*Data set*	Parameter *Support* FPM	Klaster Asli		Klaster Hasil Penggabungan	
			Jumlah Klaster	Rata-Rata Koherensi Topik	Jumlah Klaster	Rata-Rata Koherensi Topik
1	1 (550 artikel)	3	11	0,**443**	8	0,426
2		2	11	0,443	6	**0,450**
3	2 (±5.500 artikel)	3	14	0,449	8	**0,452**
4		2	14	**0,449**	7	0,441

TABEL III
PERBANDINGAN UKURAN GRAF DENGAN RATA-RATA KOHERENSI TOPIK

No. Pengujian	*Data set*	FPM *support*	Rata-Rata Ukuran Graf Klaster Asli		Rata-Rata Ukuran Graf MCS		Jumlah *Vertex* Gmcs (terkecil, terbesar)	Rata-Rata Koherensi Topik Label	
			Jumlah *vertex*	Jumlah *edge*	Jumlah *vertex*	Jumlah *edge*		Klaster Asli	Klaster Gabungan
1	1	3	72	1.123	5,29 % (4 *vertex*)	0,15 % (2 *edge*)	(0 , 11)	0,443	0,426
2		2	203	7.137	12,65 % (26 vertex)	0,60 % (43 *edge*)	(8 , 45)	0,443	0,450
3	2	3	379	34.305	18,89 % (72 vertex)	1,40 % (479 *edge*)	(21 , 211)	0,449	0,452
4		2	422	40.051	33,88% (143 vertex)	3,91% (1.565 *edge*)	(26, 247)	0,449	0,441

Cara kedua adalah dengan mengambil artikel ilmiah yang memiliki nilai *silhouette* > 0,1 dari tiap klaster hasil klasterisasi *Data set* Asli dengan jumlah klaster tujuh belas. Jumlah klaster juga ditentukan secara acak karena *Data set* Asli tidak memiliki jumlah klaster optimal dan untuk meningkatkan variasi karakteristik *data set* yang dihasilkan. Penyaringan cara kedua menghasilkan data dengan ±5.500 data artikel ilmiah, disebut sebagai *Data set* 2.

Analisis *silhouette* pada Gbr. 3 dilakukan pada *Data set* 1 dan *Data set* 2. *Data set* 1 bersifat homogen karena memiliki rata-rata nilai *silhouette* tinggi dan jumlah data sedikit. Sedangkan *Data set* 2 lebih heterogen karena nilai *silhouette* lebih rendah dan jumlah data banyak.

B. Skenario Uji Coba

Usulan metode dievaluasi menggunakan dua skenario. Skenario 1 menguji usulan metode dengan beberapa nilai parameter *support* pada tahap ekstraksi kandidat frasa dengan metode FPM. Parameter *support* menentukan ketatnya penyaringan kata untuk pembuatan graf berdasarkan prinsip *association rule mining*. Semakin rendah parameter *support* FPM, semakin banyak kata yang akan menyusun graf representasi teks klaster, sehingga graf representasi klaster akan semakin besar.

Skenario 2 menguji usulan metode dengan metode klasterisasi yang berbeda. Hal ini dilakukan untuk menguji ketahanan tahap penggabungan klaster pada penggunaan metode klasterisasi yang beragam. Metode klasterisasi yang diuji coba adalah *K-Means*++, DBSCAN, dan BIRCH.

C. Hasil & Pembahasan Skenario 1

Pengujian 1 (dengan *Data set* 1) menggunakan parameter jumlah klaster optimal sebelas dan parameter minimal *support* 3 pada metode FPM. Tabel II memperlihatkan bahwa koherensi topik pada klaster asli (sebelum dilakukan penggabungan) lebih tinggi daripada koherensi topik pada klaster yang telah mengalami penggabungan. Pengamatan lebih lanjut dilakukan dengan perbandingan ukuran graf klaster dan persentase ukuran graf MCS terhadap koherensi topik pada Tabel III.

Pengujian 2 menggunakan *Data set* 1 dengan nilai parameter minimal *support* 2 pada metode FPM sehingga ukuran graf klaster semakin besar. Sejumlah empat belas klaster digunakan dalam Pengujian 2 berdasarkan analisis *silhouette* pada Gbr. 3. Hasil rata–rata koherensi topik Pengujian 2 pada Tabel III menunjukkan peningkatan rata-rata koherensi topik terhadap Pengujian 1. Pengaruh perubahan nilai *support* FPM terhadap koherensi topik terlihat pada Tabel III melalui perbandingan rata-rata ukuran graf klaster dan rata-rata persentase ukuran graf MCS.

Pengujian 3 menggunakan nilai parameter minimal *support* 3 pada metode FPM dan jumlah klaster empat belas berdasarkan analisis *silhouette*. Pengujian ini menggunakan *Data set* 2 yang lebih bersifat heterogen, berbeda dengan *Data set* 1. Hasil rata-rata koherensi topik pada Pengujian 3 ditunjukkan pada Tabel II dan menunjukkan peningkatan koherensi topik label klaster gabungan jika dibandingkan dengan klaster asli.

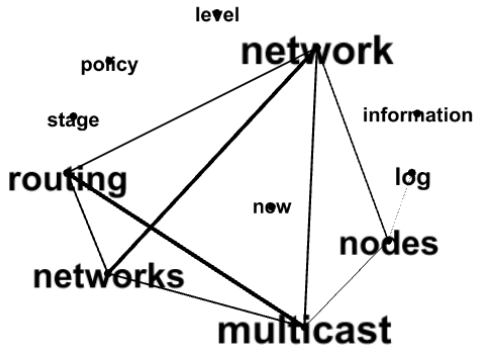
Pengujian 4 menggunakan nilai parameter *support* 2 pada metode FPM, lebih rendah dari nilai parameter *support* pada Pengujian 3. Penurunan nilai parameter *support* FPM pada *Data set* 2 menyebabkan penurunan koherensi topik pada label hasil penggabungan klaster. Hal ini disebabkan oleh *data set* yang memiliki banyak derau seperti pada hasil analisis *silhouette* pada Gbr. 3, sehingga jika semakin besar ukuran graf klaster (yang dipengaruhi oleh penurunan nilai parameter *support* FPM), maka graf klaster akan memiliki lebih banyak kata yang bersifat derau dan tidak merepresentasikan klaster.

TABEL IV
CONTOH PENGGABUNGAN KLASTER PENGUJIAN 1 (*DATA SET* 1)

Klaster Hasil Penggabungan G_{mcs}		Koherensi Topik	Klaster Asli (G_x atau G_y)		Koherensi Topik	Rata-Rata Koherensi Topik Klaster Asli
Klaster ID	**Label Klaster**		**Klaster ID**	**Label Klaster**		
3 (ada 11 *vertex* G_{mcs})	• *new multicast network* • *new interconnection network* • ***wireless data networks*** • ***fault-tolerant interconnection network*** • *multicast network*	0,482	2	***wireless data networks***, *network performance, wireless data broadcast, network traffic, network weather service*	0,403	0,454
			10	***fault-tolerant interconnection networks***, *virtual interconnection networks, arbitrary interconnection networks, multistage interconnection networks, fault-tolerant interconnection network*	0,505	

TABEL V
CONTOH PENGGABUNGAN KLASTER PENGUJIAN 3 (*DATA SET* 2)

Klaster Hasil Penggabungan G_{mcs}		Koherensi Topik	Klaster Asli (G_x atau G_y)		Koherensi Topik	Rata-Rata Koherensi Topik Klaster Asli
Klaster ID	**Label Klaster**		**Klaster ID**	**Label Klaster**		
5 (ada 80 *vertex* G_{mcs})	• *network flow algorithm* • *optimal routing algorithm* • *network algorithm* • ***linear time algorithm*** • ***polynomial time algorithm***	0,499	2	*fault-tolerant interconnection networks, fault-tolerant interconnection network, binary hypercube networks, efficient interconnection networks, arbitrary interconnection networks*	0,487	0,518
			11	***linear time algorithm***, ***polynomial time algorithm***, *time algorithm, n algorithm, n* log *n*	0,550	



Gbr. 4 Visualisasi Graf MCS pada penggabungan klaster-klaster asli 2 & 10 di Pengujian 1 Skenario 1 (*Data set* 1).

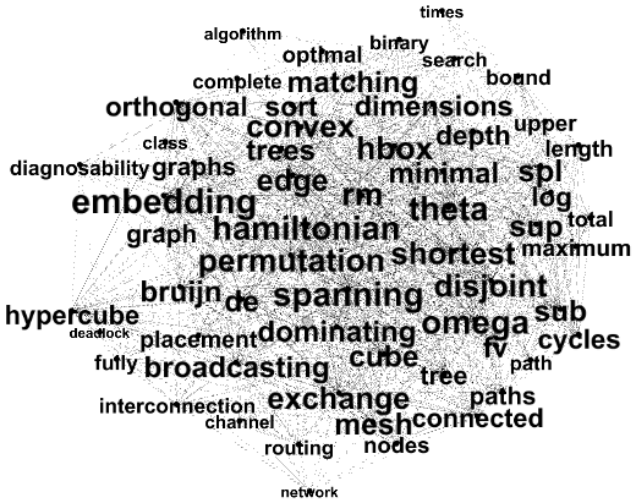
Tabel III memperlihatkan perbandingan ukuran graf terhadap koherensi topik label pada Pengujian 1. Terlihat bahwa graf MCS dalam Pengujian 1 sangat kecil dengan rata-rata persentase ukuran graf MCS hanya 5,29% untuk jumlah *vertex* dan 0,15% untuk jumlah *edge*. Artinya rata-rata jumlah *vertex* graf MCS hanya empat *vertex* dan rata-rata jumlah *edge* graf MCS hanya dua *edge*. Kecilnya ukuran graf MCS menyebabkan jumlah informasi semantik yang digunakan semakin sedikit. Informasi semantik tersebut tersimpan dalam relasi kata pada graf MCS. Oleh karena itu, pada Pengujian 2, ukuran graf klaster asli diperbesar dengan mengatur parameter *support* FPM menjadi 2. Penurunan parameter *support* FPM menyebabkan banyaknya kata yang digunakan dalam pembuatan graf klaster semakin tinggi. Peningkatan ukuran graf klaster menyebabkan peningkatan persentase ukuran graf MCS yang dihasilkan.

Perbandingan ukuran graf pada Tabel III menunjukkan bahwa penurunan nilai *support* FPM meningkatkan persentase ukuran graf MCS. Peningkatan persentase ukuran graf MCS juga diikuti dengan peningkatan koherensi topik label klaster gabungan pada *data set* 1, tercermin pada hasil Pengujian 1 dan 2. Namun, pada *data set* 2 peningkatan persentase ukuran graf MCS menyebabkan penurunan koherensi topik label klaster gabungan, tercermin pada hasil Pengujian 3 dan 4.

Contoh hasil penggabungan klaster *data set* 1 pada Pengujian 1 di Tabel IV menunjukkan bahwa klaster gabungan 3 merupakan hasil gabungan dari klaster asli 2 dan 10. Terlihat pada Tabel IV, label '*wireless data network*' pada klaster asli 2 dan label '*fault-tolerant interconnection network*' pada klaster asli 10 muncul kembali pada label klaster gabungan 3 yang ditandai dengan tulisan yang dipertebal. Dari pengamatan terhadap graf MCS yang dihasilkan klaster asli 2 dan 10 pada Gbr. 4, terlihat bahwa terdapat kata '*network*' dan '*multicast*' yang menunjukkan bahwa klaster asli 2 dan 10 memiliki kesamaan konteks pada topik komunikasi pada jaringan, sehingga frasa tersebut menjadi label klaster gabungan 3. Hal tersebut menunjukkan konsistensi graf MCS terhadap label klaster hasil penggabungan.

TABEL VI
HASIL RATA-RATA KOHERENSI TOPIK SKENARIO 2

Metode Klasterisasi	*Data set*	Klaster Asli		Klaster Hasil Penggabungan	
		Jumlah Klaster	Rata-Rata Koherensi Topik	Jumlah Klaster	Rata - Rata Koherensi Topik
K-Means++	1	11	0,443	6	**0,45**
	2	14	0,449	8	**0,452**
DBSCAN	1	10	0,495	8	**0,51**
	2	27	0,413	14	**0,426**
BIRCH	1	11	0,443	9	**0,444**
	2	14	**0,446**	9	0,443



Gbr. 5 Visualisasi graf MCS pada penggabungan klaster asli 2 & 11 di Pengujian 3 skenario 1 (*Data set* 2).

Contoh hasil penggabungan klaster *data set* 2 pada Pengujian 3 ditunjukkan pada Tabel V. Pada penggabungan tersebut, klaster gabungan 5, yang tersusun atas klaster asli 2 dan 11, memiliki koherensi topik label klaster yang lebih rendah daripada rata-rata koherensi topik label klaster penyusunnya. Graf MCS yang dihasilkan antara klaster asli 2 dan 11 diperlihatkan pada Gbr. 5. *Vertex* yang memiliki jumlah *edge* kurang dari sepuluh dihilangkan pada gambar tersebut untuk memudahkan visualisasi. Pada graf MCS tersebut terlihat bahwa terdapat kata-kata notasi perhitungan yang tidak memiliki konteks, seperti "*hbox*", "*theta*", dan "*rm*", sehingga parameter *support* FPM 3 kurang tepat untuk digunakan pada *data set* yang memiliki jumlah kata yang besar seperti *data set* 2.

D. Hasil & Pembahasan Skenario 2

Pada hasil skenario 1, dapat diketahui bahwa pada *data set* 1 dan 2 hasil terbaik diberikan oleh parameter *support* FPM 2 dan 3 berturut-turut. Parameter tersebut digunakan pada skenario 2 untuk masing-masing *data set* dengan menggunakan metode klasterisasi DBSCAN dan BIRCH. Hasil dari skenario 2 ditunjukkan pada Tabel VI.

Pada Tabel VI terlihat bahwa penggunaan klasterisasi DBSCAN menghasilkan koherensi topik label klaster gabungan yang lebih baik daripada label klaster sebelum penggabungan, pada kedua *data set*. Hal ini disebabkan oleh metode klasterisasi DBSCAN yang menghilangkan data-data yang bersifat derau. Pada metode klasterisasi BIRCH hanya terjadi peningkatan 0,001 pada *data set* 1 dan penurunan sebesar 0,004 pada *data set* 2. Penurunan terjadi pada *data set* 2 karena sifat *data set* 2 yang heterogen dan memiliki banyak data derau dibandingkan *data set* 1. Hasil skenario 2 menunjukkan bahwa derau memengaruhi hasil penggabungan klaster, sehingga penggunaan metode klasterisasi DBSCAN dapat memberikan hasil yang lebih baik daripada metode klasterisasi lainnya.

V. KESIMPULAN

Dalam makalah ini metode pelabelan klaster dengan penggabungan klaster berbasis graf diusulkan. Penggabungan klaster dilakukan secara aglomeratif dengan pengukuran similaritas menggunakan MCS. Pelabelan klaster diimplementasikan pada hasil klaster gabungan dengan menggunakan *TopicRank*.

Pengujian dilakukan dalam dua skenario terhadap dua jenis *data set* dengan perbedaan karakteristik yaitu *data set* 1 yang bersifat homogen dan *data set* 2 yang bersifat heterogen. Hasil pengujian pada skenario 1 menunjukkan bahwa ukuran graf memengaruhi hasil koherensi topik label klaster gabungan. Koherensi topik label klaster menurun jika ukuran graf yang dihasilkan terlalu besar, seperti pada *data set* 2, sehingga parameter *support* pada metode FPM harus disesuaikan dengan ukuran korpus *data set*. Pengamatan lebih lanjut terhadap graf MCS memperlihatkan konsistensi isi graf MCS terhadap label klaster hasil penggabungan. Hal ini menunjukkan informasi semantik yang terkandung pada relasi kata dalam graf MCS berpengaruh terhadap koherensi topik label klaster hasil penggabungan.

Hasil skenario 2 menunjukkan bahwa penggunaan DBSCAN dalam penggabungan klaster memberikan hasil yang lebih baik dibandingkan metode klasterisasi lainnya. Hal ini disebabkan oleh metode klasterisasi DBSCAN, yang dapat menghilangkan data yang bersifat derau. Oleh karena itu, berdasarkan hasil pengujian dapat disimpulkan bahwa koherensi topik label klaster gabungan sensitif terhadap tingkat derau pada *data set*.

UCAPAN TERIMA KASIH

Terima kasih disampaikan kepada Institut Teknologi Sepuluh Nopember (ITS) atas dukungan dana yang telah

diberikan dalam penelitian ini melalui program beasiswa Fresh Graduate berdasarkan SK Rektor ITS nomor: 046993/IT2/HK.00.01/2016.

REFERENSI

[1] H. Park, K. Kwon, A. i. Z. Khiati, J. Lee, dan I. J. Chung, "Agglomerative Hierarchical Clustering for Information Retrieval Using Latent Semantic Index," *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, 2015, hal. 426–431.

[2] S. Shah dan X. Luo, "Exploring Diseases Based Biomedical Document Clustering and Visualization Using Self-Organizing Maps," *2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)*, 2017, hal. 1–6.

[3] A. Wahib, A. Z. Arifin, dan D. Purwitasari, "Improving Multi-Document Summary Method Based on Sentence Distribution," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, Vol. 14, No. 1, hal. 286, 2016.

[4] A. Zaini, M. A. Muslim, dan W. Wijono, "Pengelompokan Artikel Berbahasa Indonesia Berdasarkan Struktur Laten Menggunakan Pendekatan Self Organizing Map," *J. Nas. Tek. Elektro dan Teknol. Inf.*, Vol. 6, No. 3, hal. 259-267, 2017.

[5] D. Purwitasari, C. Fatichah, I. Arieshanti, dan N. Hayatin, "K-medoids Algorithm on Indonesian Twitter Feeds for Clustering Trending Issue as Important Terms in News Summarization," *Proc. 2015 Int. Conf. Inf. Commun. Technol. Syst. ICTS 2015*, 2015, hal. 95–98.

[6] P. Hennig, P. Berger, C. Steuer, C. Wuerz, dan C. Meinel, "Cluster Labeling for the Blogosphere," *2014 IEEE Fourth International Conference on Big Data and Cloud Computing*, 2014, hal. 416–423.

[7] P. Xie dan E. P. Xing, "Integrating Document Clustering and Topic Modeling," *Proc. 29th Conf. Uncertain. Artif. Intell.*, 2013, hal. 694-703.

[8] Q. Mei, X. Shen, dan C. Zhai, "Automatic Labeling of Multinomial Topic Models," *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '07*, 2007, hal. 490-499.

[9] T. L. Griffiths dan M. Steyvers, "Finding Scientific Topics," *Proc. Natl. Acad. Sci.*, Vol. 101, No. Supplement 1, hal. 5228–5235, 2004.

[10] C. Aalla dan V. Pudi, "Mining Research Problems from Scientific Literature," *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2016, hal. 351–360.

[11] Z. Li, J. Li, Y. Liao, S. Wen, dan J. Tang, "Labeling Clusters from Both Linguistic and Statistical Perspectives: A Hybrid Approach," *Knowledge-Based Syst.*, Vol. 76, hal. 219–227, 2015.

[12] N. Y. Saiyad, H. B. Prajapati, dan V. K. Dabhi, "A Survey of Document Clustering Using Semantic Approach," *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, 2016, hal. 2555–2562.

[13] J. Jayabharathy, S. Kanmani, dan A. A. Parveen, "Document Clustering and Topic Discovery Based on Semantic Similarity in Scientific Literature," *2011 IEEE 3rd International Conference on Communication Software and Networks*, 2011, hal. 425–429.

[14] S. S. Sonawane dan P. A. Kulkarni, "Graph based Representation and Analysis of Text Document: A Survey of Techniques," *Int. J. Comput. Appl.*, Vol. 96, No. 19, hal. 1–8, Jun. 2014.

[15] S. Sonawane dan P. Kulkarni, "Graph based Representation and Analysis of Text Document: A Survey of Techniques," *Int. J. Comput. Appl.*, Vol. 96, No. 19, hal. 1–8, 2014.

[16] N. Shanavas, H. Wang, Z. Lin, dan G. Hawe, "Centrality-Based Approach for Supervised Term Weighting," *IEEE Int. Conf. Data Min. Work. ICDMW*, 2017, hal. 1261–1268.

[17] F. Role dan M. Nadif, "Beyond Cluster Labeling: Semantic Interpretation of Clusters' Contents Using a Graph Representation," *Knowledge-Based Syst.*, Vol. 56, hal. 141–155, 2014.

[18] A. El-Kishky, Y. Song, C. Wang, C. Voss, dan J. Han, "Scalable Topical Phrase Mining from Text Corpora," *Proc. VLDB Endow.*, Vol. 8, No. 3, hal. 305–316, 2014.

[19] T. Mikolov, G. Corrado, K. Chen, dan J. Dean, "Efficient Estimation of Word Representations in Vector Space," *Proc. Int. Conf. Learn. Represent. (ICLR 2013)*, 2013, hal. 1–12.

[20] J. Wu, Z. Xuan, dan D. Pan, "Enhancing Text Representation for Classification Tasks with Semantic Graph Structures," *Int. J. Innov. Comput. Inf. Control*, Vol. 7, No. 5, hal. 13–16, 2011.

[21] L. Sterckx, T. Demeester, J. Deleu, dan C. Develder, "Topical Word Importance for Fast Keyphrase Extraction," *Proc. 24th Int. Conf. World Wide Web - WWW '15 Companion*, 2015, No. 2, hal. 121–122.

[22] M. Röder, A. Both, dan A. Hinneburg, "Exploring the Space of Topic Coherence Measures," *Proc. Eighth ACM Int. Conf. Web Search Data Min. - WSDM '15*, 2015, hal. 399–408.

[23] R. Mihalcea dan P. Tarau, "TextRank: Bringing Order into Texts," *Proc. EMNLP*, Vol. 85, hal. 404–411, 2004.

[24] A. Hulth, "Improved Automatic Keyword Extraction Given More Linguistic Knowledge," *Proc. 2003 Conf. Empir. Methods Nat. Lang. Process.*, 2003, No. 2000, hal. 216–223.

[25] M. Grineva, M. Grinev, dan D. Lizorkin, "Extracting Key Terms from Noisy and Multitheme Documents," *Proc. 18th Int. Conf. World wide web - WWW '09*, 2009, hal. 661-670.

[26] K. S. Hasan dan V. Ng, "Automatic Keyphrase Extraction: A Survey of the State of the Art," *Proc. 52nd Annu. Meet. Assoc. Comput. Linguist. (Volume 1 Long Pap.)*, 2014, hal. 1262–1273.

[27] L. H. Suadaa dan A. Purwarianti, "Combination of Latent Dirichlet Allocation (LDA) and Term Frequency-Inverse Cluster Frequency (TFxICF) in Indonesian Text Clustering with Labeling," *2016 4th Int. Conf. Inf. Commun. Technol. ICoICT 2016*, 2016, hal. 1-6.

[28] D. Carmel, H. Roitman, dan N. Zwerdling, "Enhancing Cluster Labeling Using Wikipedia," *Proc. 32nd Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. - SIGIR '09*, 2009, hal. 139–146.

[29] L. Page, S. Brin, R. Motwani, dan T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," *World Wide Web Internet Web Inf. Syst.*, Vol. 54, No. 1999–66, hal. 1–17, 1998.

[30] X. Wan dan J. Xiao, "CollabRank: Towards a Collaborative Approach to Single-Document Keyphrase Extraction," *Proc. 22nd Int. Conf. Comput. Linguist. Coling 2008*, 2008, hal. 969–976.

[31] Z. Liu, P. Li, Y. Zheng, dan M. Sun, "Clustering to Find Exemplar Terms for Keyphrase Extraction," *Proc. 2009 Conf. Empir. Methods Nat. Lang. Process.: Vol. 1*, 2009, hal. 257–266.

[32] Z. Liu, W. Huang, Y. Zheng, dan M. Sun, "Automatic Keyphrase Extraction via Topic Decomposition," *Proc. 2010 Conf. Empir. Methods Nat. Lang. Process.*, 2010, hal. 366–376.

[33] N. F. Azzahra, H. Ginardi, dan A. Saikhu, "Praproses Data Alir ADS-B dari Multi-Receiver dengan Pengelompokan Agglomerasi Berbasis Konsistensi Jarak," *JNTETI*, Vol. 4, No. 1, hal. 39-44, 2015.

[34] A. Krauza, "Extension of Fuzzy Gustafson-Kessel Algorithm Based on Adaptive Cluster Merging," *2015 IEEE MIT Undergrad. Res. Technol. Conf. URTC 2015*, 2016, hal. 1–4.

[35] C. Jin dan Q. Bai, "Text Clustering Algorithm Based on the Graph Structures of Semantic Word Co-occurrence," *2016 Int. Conf. Inf. Syst. Artif. Intell.*, 2016, hal. 497-502.

[36] P. J. Rousseeuw, "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis," *J. Comput. Appl. Math.*, Vol. 20, hal. 53–65, 1987.

[37] R. Gunawan dan K. Mustofa, "Pencarian Aturan Asosiasi Semantic Web untuk Obat Tradisional Indonesia," *J. Nas. Tek. Elektro dan Teknol. Informasi (JNTETI)*, Vol. 5, No. 3, hal. 192–200, 2016.