

**TESIS – KI 42502**

**PELABELAN KLASTER ARTIKEL ILMIAH MENGGUNAKAN *TOPICRANK* DAN *MAXIMUM COMMON SUBGRAPH***

**Adhi Nurilham**

**NRP. 5116201047**

**DOSEN PEMBIMBING**

**Dr. Eng. Chastine Fatichah, M.Kom.**

**NIP: 197512202001122002**

**PROGRAM MAGISTER**

**DEPARTEMEN INFORMATIKA**

**FAKULTAS TEKNOLOGI INFORMASI DAN KOMUNIKASI**

**INSTITUT TEKNOLOGI SEPULUH NOPEMBER**

**SURABAYA**

**2018**

**Pelabelan Klaster Artikel Ilmiah Menggunakan TopicRank dan Maximum Common Subgraph**

Nama Mahasiswa : Adhi Nurilham

NRP : 5116 201 047

Pembimbing : Dr. Eng. Chastine Fatichah, M.Kom.

# ABSTRAK

Metode klasterisasi dapat memudahkan pengelompokkan kumpulan artikel ilmiah yang tidak terstruktur berdasarkan topik. Pelabelan klaster pada hasil klasterisasi diperlukan untuk mengetahui frasa kunci yang merepresentasikan topik yang dibahas. Beberapa hasil klaster masih perlu dilakukan penggabungan karena masih memiliki kemiripan topik untuk memberikan hasil label klaster yang lebih baik. Selain kesamaan kata, kemiripan topik juga dapat diwakili dengan kesamaan relasi kata yang dapat dimodelkan dengan graf. Penelitian ini mengusulkan pelabelan klaster artikel ilmiah dengan tahap penggabungan klaster berbasis graf sebagai kontribusi penelitian untuk memberikan label topik yang lebih representatif.

Usulan metode terdiri dari beberapa tahapan. Pertama, metode klasterisasi *K-Means* dilakukan pada kumpulan artikel ilmiah. Kedua, pada setiap klaster, ekstraksi kandidat frasa dilakukan menggunakan *Frequent Phrase Mining* (FPM) untuk mendapatkan kata - kata yang memiliki potensi untuk membentuk frasa topik. Kata – kata pembentuk frasa digunakan sebagai masukkan untuk pembuatan graf representasi klaster. Setelah itu, penggabungan klaster dilakukan dengan mengukur similaritas graf klaster menggunakan metode *Maximum Common Subgraph* (MCS). Lalu, proses pelabelan klaster dilakukan pada kumpulan klaster yang telah digabung dengan menggunakan metode TopicRank.

Tedapat 2 skenario pengujian untuk mengevaluasi metode usulan. Skenario pertama bertujuan untuk mengevaluasi keefektifan tahap penggabungan klaster, dengan membandingkan nilai *silhouette index* pada klaster yang dilakukan penggabungan dan pada klaster yang tidak melakukan penggabungan. Skenario kedua bertujuan untuk mengevaluasi kinerja metode pelabelan klaster dengan menggunakan kuisioner.

**Kata kunci***: Pelabelan Klaster, Teori Graf, TopicRank, Frequent Phrase Mining, Maximum Common Subgraph*

*[Halaman ini sengaja dikosongkan]*

**Cluster Labelling on Scientific Article Using TopicRank and Maximum Common Subgraph**

Student Name : Adhi Nurilham

NRP : 5116 201 047

Supervisor : Dr. Eng. Chastine Fatichah, M.Kom.

# ABSTRACT

Unstructured scientific articles can benefited by clustering method to group scientific articles based on topic similarity. Cluster labeling on the yielded cluster is required to discover key phrases that best represent the topics covered. Several clusters still need to be bundled because they still have similar topics to give better cluster labels results. In addition to word occurences, the similarity of the topic can also be represented by word semantic relation that can be modeled with the graph. This research proposes labeling clusters of scientific articles with cluster merging as research contribution to provide a more representative label of cluster topics.

This research proposed cluster labeling method with cluster merging process using graph model. Graph model approach is choosen because it can map the relationship between words, hence representing text semantic information. There are several stages in the proposed method. First, *K-Means* clustering method is applied on a collection of scientific articles. Second, for each cluster, phrase extraction is executed using *Frequent Phrase Mining* to get word tokens that capable to constitute representative phrase for cluster topics. Acquired word tokens used as input to constructing graph representation of a cluster. After that, cluster merging is done based on cluster graph similarity using *Maximum Common Subgraph* (MCS) method. Then, the cluster labeling process is performed on clusters that have been merged using the *TopicRank* method.

There are two test scenarios to evaluate the proposed method. The first scenario aims to evaluate the effectiveness of the cluster merging stage, by comparing the silhouette index values on clusters with cluster merging process and without cluster merging process. The second scenario aims to evaluate the performance of cluster labeling methods by comparing the value of expert rating on the cluster label results between proposed methods and other methods.

***Keywords****: Cluster Labeling, Graph Theory, TopicRank, Frequent Phrase Mining, Maximum Common Subgraph*

*[Halaman ini sengaja dikosongkan]*

# KATA PENGANTAR

Alhamdulillahirabbil’alamin. Puji dan syukur penulis panjatkan kehadirat  
Allah SWT atas berkat, rahmat dan hidayah-Nya, penyusunan Tesis ini dapat  
diselesaikan. Tesis ini dibuat sebagai salah satu syarat dalam menyelesaikan  
Program Studi Magister di Institut Teknologi Sepuluh November Surabaya.  
Penulis menyadari bahwa Tesis ini dapat diselesaikan karena dukungan dari  
berbagai pihak, baik dalam bentuk dukungan moral dan material.  
Melalui kesempatan ini dengan kerendahan hati penulis mengucapkan  
terima kasih dan penghargaan setinggi-tingginya kepada semua orang untuk  
semua bantuan yang telah diberikan, antara lain kepada:

1. Ayahanda tercinta Bambang Soekanto dan Ibunda tercinta Enna Andari  
   untuk semua doa, pengorbanan dan usaha yang tak kenal lelah telah  
   mendidik dan membimbing dengan penuh ketulusan untuk keberhasilan  
   penulis.
2. Nenek tercinta Toebijati Soetrisno atas dukungan dan doanya selama menempuh pendidikan di Kota Surabaya.
3. Ito Nurarief, Nahla Nur Ardhiani serta seluruh keluarga  
   yang selalu berdoa dan memberikan dukungan.
4. Ibu Dr. Eng. Chastine Fatichah, S.Kom, M.Kom dan Ibu Diana Purwitasari, S.Kom, M.Sc. selaku pembimbing yang senantiasa memberikan arahan dan bimbingan kepada penulis. Semoga Allah SWT senantiasa merahmati Ibu dan keluarga.
5. Bapak Dr. Agus Zainal Arifin, S.Kom, M.Kom., Dr. Darlis Herumurti,  
   S.Kom, M.Kom., dan Bapak Prof. Dr. Ir. Joko Liantio Buliali sebagai tim  
   Penguji Tesis yang memberikan masukan dan kritik yang membangun  
   untuk Tesis ini.
6. Seluruh dosen S2 Teknik Informatika yang telah memberikan ilmu dan  
   pengetahuan kepada penulis selama menempuh studi.
7. Teman seperjuangan lainnya yang tidak dapat disebutkan satu persatu, terima kasih atas bantuan dan motivasi yang telah diberikan.

Akhirnya dengan segala kerendahan hati penulis menyadari masih banyak  
terdapat kekurangan pada Tesis ini. Oleh karena itu, segala tegur sapa dan kritik yang sifatnya membangun sangat penulis harapkan demi kesempurnaan Tesis ini.  
Penulis berharap bahwa perbuatan baik dari semua orang yang dengan tulus  
memberikan kontribusi terhadap penyusunan Tesis ini mendapatkan pahala dari  
Allah. Aamiin Alluhamma Aamiin.

Surabaya,

Adhi Nurilham

*[Halaman ini sengaja dikosongkan]*

# DAFTAR ISI

[ABSTRAK 2](#_Toc518291956)

[ABSTRACT 4](#_Toc518291957)

[KATA PENGANTAR 6](#_Toc518291958)

[DAFTAR ISI 8](#_Toc518291959)

[DAFTAR GAMBAR 11](#_Toc518291960)

[DAFTAR TABEL 13](#_Toc518291961)

[1 BAB 1 PENDAHULUAN 15](#_Toc518291962)

[1.1 Latar Belakang 15](#_Toc518291963)

[1.2 Perumusan Masalah 17](#_Toc518291964)

[1.3 Tujuan 18](#_Toc518291965)

[1.4 Manfaat 18](#_Toc518291966)

[1.5 Kontribusi Penelitian 18](#_Toc518291967)

[1.6 Batasan Masalah 18](#_Toc518291968)

[2 BAB 2 KAJIAN PUSTAKA 19](#_Toc518291969)

[2.1 Klasterisasi Dokumen 19](#_Toc518291970)

[2.1.1 K-Means 19](#_Toc518291971)

[2.1.2 Klasterisasi Hirarkikal 20](#_Toc518291972)

[2.2 Pra-pemrosesan Teks 21](#_Toc518291973)

[2.3 Perhitungan Similaritas Teks menggunakan Maximum Common Subgraph 21](#_Toc518291974)

[2.4 Ekstraksi Frasa menggunakan *Frequent Phrase Mining* (FPM) 22](#_Toc518291975)

[2.5 Representasi Teks dalam Graf 23](#_Toc518291976)

[2.6 TopicRank 24](#_Toc518291977)

[2.7 Koherensi Topik 25](#_Toc518291978)

[3 BAB 3 METODOLOGI PENELITIAN 27](#_Toc518291979)

[3.1 Studi Literatur 27](#_Toc518291980)

[3.2 Perancangan Sistem 28](#_Toc518291981)

[3.3 Evaluasi 32](#_Toc518291982)

[4 BAB 4 UJI COBA & ANALISIS HASIL 35](#_Toc518291983)

[4.1 Lingkungan Uji Coba 35](#_Toc518291984)

[4.2 Persiapan Data 35](#_Toc518291985)

[4.3 Hasil Uji Coba 36](#_Toc518291986)

[4.3.1. Skenario 1 36](#_Toc518291987)

[4.3.2. Skenario 2 39](#_Toc518291988)

[5 BAB 5 PENUTUP 41](#_Toc518291989)

[5.1 Kesimpulan 41](#_Toc518291990)

[5.2 Saran 41](#_Toc518291991)

[DAFTAR PUSTAKA 43](#_Toc518291992)

# DAFTAR GAMBAR

[Gambar 3.1 Alur Metodologi Penelitian 26](#_Toc518291941)

[Gambar 3.2 Alur Proses Metode Usulan 27](#_Toc518291942)

[Gambar 3.3 Contoh Tahap Pra-pemrosesan Teks 28](#_Toc518291943)

[Gambar 3.4 Contoh Tahap Ekstraksi Frasa Topik 29](#_Toc518291944)

[Gambar 3.5 Contoh Tahap Konstruksi Graf 30](#_Toc518291945)

[Gambar 3.6 Contoh Penggabungan Klaster 30](#_Toc518291946)

[Gambar 3.7 Contoh Hasil TopicRank 31](#_Toc518291947)

[Gambar 4.1. Analisis *Silhouette* Dataset Asli, Dataset 1, dan Dataset 2 35](file:///D:\Google%20Drive\Thesis\Buku%20-%20Work%20in%20Progress\Buku_Ver1.docx#_Toc518291948)

[Gambar 4.3. Graf MCS Klaster Asli KA-5 dan KA-7 pada Percobaan 1 38](file:///D:\Google%20Drive\Thesis\Buku%20-%20Work%20in%20Progress\Buku_Ver1.docx#_Toc518291949)

[Gambar 4.2. Graf MCS Klaster Asli KA-4 dan KA-7 pada Percobaan 1 38](file:///D:\Google%20Drive\Thesis\Buku%20-%20Work%20in%20Progress\Buku_Ver1.docx#_Toc518291950)

*[Halaman ini sengaja dikosongkan]*

# DAFTAR TABEL

[Tabel 4.1. Hasil Rata - Rata Koherensi Topik Skenario 1 36](file:///D:\Google%20Drive\Thesis\Buku%20-%20Work%20in%20Progress\Buku_Ver1.docx#_Toc518291951)

[Tabel 4.2. Hasil Penggabungan dan Pelabelan Klaster Percobaan 1 37](file:///D:\Google%20Drive\Thesis\Buku%20-%20Work%20in%20Progress\Buku_Ver1.docx#_Toc518291952)

[Tabel 4.3. Perbandingan Ukuran Graf Terhadap Rata- Rata Koherensi Topik 37](file:///D:\Google%20Drive\Thesis\Buku%20-%20Work%20in%20Progress\Buku_Ver1.docx#_Toc518291953)

*[Halaman ini sengaja dikosongkan]*

# BAB 1 PENDAHULUAN

1. Latar Belakang

Metode klasterisasi dokumen dapat digunakan untuk mengelompokkan artikel ilmiah berdasarkan kemiripan topik, namun klasterisasi hanya menghasilkan kelompok artikel ilimiah tanpa disertai label topik penelitian yang mewakili kelompok artikel ilmiah tersebut. Pelabelan klaster diperlukan untu mencari frasa representasi topik dari kelompok artikel ilmiah hasil klasterisasi.

Pada umumnya pelabelan klaster terbagi ke dalam 2 proses utama, yaitu proses ekstraksi frasa kandidat dan proses penilaian frasa. Pada proses ekstraksi frasa, Liao *et al* [1] mengimplementasi proses ekstraksi frasa berbasis aturan yang dibuat secara otomatis berdasarkan struktur *part-of-speech* dari kumpulan judul artikel. Carmel *et al* [2] menggunakan pengukuran *Jensen-Shannon Divergence* (JSD) untuk mencari kumpulan frasa yang memiliki jarak terjauh dengan klaster lainnya. Lopez *et al* [3] mengembangkan metode ekstraksi kata pada dokumen artikel ilmiah dengan menggunakan korpus eksternal GRISP dan *Wikipedia*. GRISP (*General Research Insight* *in Scientific and Technical Publication*) merupakan basis data terminologi dalam berbagai bidang ilmiah [4]. Kemunculan frasa pada basis data GRISP, membuat frasa tersebut lebih penting daripada frasa – frasa lainnya. El-Kishky *et al*. mengusulkan algoritma *Frequent Phrase Mining* (FPM) [5] yang terinspirasi dari algoritma *Apriori* [6]. Algoritma FPM telah dikembangkan untuk menemukan frasa konseptual yang merepresentasikan topik penelitan pada artikel ilmiah dengan menambahkan metode heuristik [7]. Penggunaan metode heuristik dalam ekstraksi frasa telah terbukti memberikan *recall* yang tinggi pada beberapa penelitian lainnya [8].

Pada proses penilaian frasa untuk pelabelan klaster, pendekatan yang paling banyak digunakan adalah pendekatan statistik. Pada pendekatan statistik, frekuensi kemunculan sebuah kata atau frasa dapat menjadi faktor dalam pemilihan label klaster dokumen. Aalla *et al*. menggunakan pembobotan IDF (*Inverse Document Frequency*) untuk mendapatkan frasa konseptual yang signifikan pada kumpulan artikel ilmiah [7]. Suadaa *et al.* menggunakan pembobotan TF-ICF (*Term Frequency* – *Inverse Cluster Frequency*) untuk memilih label klaster [9]. TF-ICF menghitung frekuensi kata pada dokumen dan frekuensi kebalikan kata pada klaster – klaster dokumen. Liao *et al* menggunakan perhitungan berbasis *markov chain* untuk menilai frasa kandidat label klaster [1].

Metode penilaian frasa berbasis graf juga telah banyak diusulkan. Mihalcea *et al* [10] mengusulkan TextRank, metode penilaian frasa berbasis graf yang terinsiprasi dari algoritma PageRank [11]. CollabRank merupakan metode pengembangan TextRank yang menggunakan model graf yang dibuat dari beberapa dokumen dalam klaster yang sama untuk mengekstrasi frasa kata kunci pada sebuah dokumen [12]. Liu *et al* mengusulkan TopicRank yang menggunakan persebaran topik dalam perhitungan algoritma TextRank untuk meningkatkan cakupan topik yang direpresentasikan frasa kunci [13], dan telah dikembangkan oleh Sterckx *et al* untuk mengoptimalkan efisiensi waktu ekstraksi [14].

Label kelompok dokumen yang baik harus dapat dibedakan dengan label kelompok dokumen lainnya [15]. Namun terkadang proses klasterisasi menghasilkan hasil yang kurang optimal (*suboptimal*) sehingga terdapat beberapa kelompok dokumen yang memiliki kemiripan kontekstual [16]. Hal tersebut menyebabkan label kelompok dokumen yang dihasilkan sulit dibedakan. Hasil klasterisasi yang *suboptimal* disebabkan oleh beberapa faktor seperti pemilihan jumlah klaster, adanya dokumen yang bersifat derau (*outlier*), dan pemilihan fitur teks [17]. Oleh karena itu, diperlukan proses identifikasi dan penggabungan kelompok – kelompok artikel ilmiah yang memiliki kemiripan sebelum pelabelan klaster dilakukan.

Penggabungan klaster telah diusulkan pada beberapa penelitian untuk mengatasi berbagai macam permasalahan. Krauza et al. menggunakan penggabungan klaster untuk mengatasi hasil klaster yang tumpang tindih pada metode klasterisasi Fuzzy Gustafson-Kessel (FGK) [18]. Morsier et al. melakukan penggabungan klaster pada klaster yang tumpang tindih berdasarkan persebaran outlier antara klaster [19]. Czarnowski et al. mengusulkan penggabungan klaster berbasis konsensus dalam pemilihan data untuk proses pelatihan model klasifikasi [20].

Metode penggabungan klaster yang telah diusulkan pada umumnya memanfaatkan pendekatan *Vector Space Model* (VSM). Pada pendekatan VSM, jarak klaster, sebagai kriteria penggabungan klaster, dihitung dalam ruang fitur menggunakan perhitungan jarak seperti *euclidean distance*. Pada penggabungan artikel ilmiah pendekatan VSM menganggap bahwa setiap kata bersifat independen terhadap kata lainnya [21]. Sedangkan, sebuah artikel ilmiah tersusun atas kata – kata yang saling berhubungan. Hubungan antar kata pada sebuah artikel ilmiah dapat menggambarkan unsur semantik yang terdapat pada artikel ilmiah.

Pendekatan graf merupakan salah satu alternatif pendekatan VSM yang telah dijelaskan sebelumnya. Pendekatan graf dapat memetakan kata berdasarkan konteksnya dengan merepresentasikan kata sebagai *node* yang saling berkaitan [22]. Integrasi visualisasi pada graf dan metode statistik dapat membantu pencarian fitur penting pada data [23], tidak terkecuali fitur semantik teks. Jin *et al* [24] mengusulkan perhitungan similaritas teks berbasis graf menggunakan MCS (*Maximum Common Subgraph*). Perhitungan MCS mencari nilai similaritas antar teks dengan menghitung jumlah *node* dan *edge* yang sama antara graf representasi teks. Implementasi MCS pada proses penggabungan klaster dapat menjadi solusi alternatif penggunaan metode perhitungan berbasis VSM.

Pada penelitian ini diusulkan metode pelabelan klaster dengan proses penggabungan klaster menggunakan pendekatan graf. Pendekatan graf dilakukan karena dapat memetakan hubungan antar kata yang merepresentasikan informasi semantik teks. Pertama, metode klasterisasi k-means dilakukan pada kumpulan artikel ilmiah. Lalu, pada setiap klaster, ekstraksi frasa dilakukan untuk mendapatkan kata - kata yang merepresentasikan topik klaster. Setelah itu, penggabungan klaster dilakukan dengan pendekatan graf menggunakan metode *Maximum Common Subgraph* (MCS). Proses pelabelan klaster dilakukan pada kumpulan klaster yang telah digabung dengan menggunakan metode TopicRank [14]. Pelabelan klaster TopicRank digunakan karena metode ekstraksi frasa yang mengimplementasi klasterisasi topik terbukti menghasilkan hasil yang baik menggunakan data abstrak artikel ilmiah [8].

1. Perumusan Masalah

Rumusan masalah yang diangkat dalam penelitian ini adalah sebagai berikut.

1. Bagaimana cara merepresentasikan klaster artikel ilmiah dalam bentuk graf berdasarkan topik?
2. Bagaimana cara mengidentifikasi beberapa klaster artikel ilmiah yang memiliki kemiripan topik?
3. Bagaimana cara mengevaluasi hasil label klaster artikel ilmiah?
4. Tujuan

Tujuan yang akan dicapai dalam pembuatan tesis ini adalah pelabelan klaster artikel ilmiah menggunakan TopicRank dan *Frequent Phrase Mining* dengan pendekatan *Maximum Common Subgraph*.

1. Manfaat

Manfaat dari penelitian ini adalah memudahkan pengguna dalam mendapatkan topik penelitian pada kumpulan artikel ilmiah.

1. Kontribusi Penelitian

Penelitian ini memiliki kontribusi pada metode pelabelan klaster menggunakan FPM (*Frequent Phrase Mining*) dan TopicRank, dengan penggabungan klaster menggunakan MCS (*Maximum Common Subgraph*).

1. Batasan Masalah

Batasan masalah pada penelitian ini adalah:

1. Data dokumen yang digunakan adalah abstrak artikel ilmiah berbahasa Inggris dari basis data *Aminer.*

# BAB 2 KAJIAN PUSTAKA

Pada bab ini akan dijelaskan tentang pustaka yang terkait dengan landasan penelitian. Pustaka yang terkait dianalisa dan dirangkum dalam bentuk studi komparasi.

1. Klasterisasi Dokumen

Klasterisasi dokumen merupakan sebuah proses yang bertujuan untuk mengelompokkan dokumen. Pada penelitian ini akan digunakan dua metode klasterisasi dokumen yaitu K-Means++ dan Klasterisasi Hirarkikal. Kedua pendekatan klasterisasi dokumen tersebut dijelaskan pada subbab berikut ini.

* 1. K-Means++

K-Means++ merupakan metode pengembangan dari metode klasterisasi yang paling umum digunakan K-Means untuk mengelompokan dokumen [16]. Metode klasterisasi ini memiliki kekurangan dimana jumlah klaster perlu diketahui terlebih dahulu [25]. Pada umumnya K-Means menggunakan pendekatan VSM (*Vector Space Model*), dimana dokumen dimodelkan dalam vektor yang memiliki kata sebagai fitur. Setelah kumpulan dokumen telah melewati tahap pra-pemrosesan teks, seluruh kata pada kumpulan dokumen diekstrak untuk dijadikan fitur dokumen, pendekatan ini disebut juga *“bag-of-words model”*. Vektor dokumen dibentuk dengan menggunakan pembobotan Tf-Idf (*Term Frequency – Inverse Document Frequency*) pada fitur kata.

Pada pendekatan VSM, similaritas antar vektor dokumen dapat dihitung dengan menggunakan *cosine similarity*. Jika terdapat 2 vektor dokumen dan , maka nilai *cosine similarity* antara 2 vektor tersebut dihitung dengan menggunakan rumus (1), dimana merupakan nilai skalar dari .

( 1 )

Masukkan dari K-Means adalah kumpulan dokumen dan parameter jumlah klaster . Setelah vektor dokumen terbentuk, algoritma K-Means dilakukan seperti berikut :

1. Pilih dokumen secara acak sebagai *centroid* awal klaster .
2. Hitung similaritas dokumen dengan tiap *centroid* pada menggunakan rumus (1).
3. Dokumen akan menjadi anggota klaster yang memiliki nilai similaritas *centroid* tertinggi.
4. Ulangi langkah 2-3 untuk setiap dokumen dalam .
5. Hitung ulang *centroid* untuk setiap .
6. Ulangi langkah 2-5, sampai tidak ada dokumen yang berpindah klaster.
   1. Klasterisasi Hirarkikal

Klasterisasi hirarkikal dokumen teks merupakan metode pengelompokkan dokumen yang bekerja dengan membangun sebuah hirarki kelompok dokumen atau klaster. Hirarki klaster dapat disebut juga sebagai dendogram. Klasterisasi hirarkikal telah digunakan untuk pembentukan taksonomi konsep pada suatu teks [26]. Kelebihan klasterisasi hirarkikal dibandingkan dengan metode klasterisasi lain adalah tingkat hirarki yang dapat ditentukan sesuai kebutuhan [27].

Pada umumnya, pembangunan dendogram menggunakan pendekatan *bottom-up* yang dilakukan secara iteratif. Pada iterasi awal setiap dokumen memiliki klaster tersendiri, lalu klaster yang memiliki similaritas tertinggi akan digabung pada setiap iterasi selanjutnya. Iterasi akan berhenti, jika nilai similaritas tertinggi kurang dari nilai batas yang telah ditentukan. Similaritas antara dua buah dokumen diukur dengan menggunakan MCS (*Maximum Common Subgraph*) yang dijelaskan pada subbab 2.3. Setelah dua klaster digabung, nilai similaritas baru antara gabungan klaster tersebut dengan klaster lain ditentukan dengan rata – rata dari selisih nilai similaritas antara tiap anggota klaster (*average linked*).

Masukkan metode klasterisasi hirarkikal adalah matriks similaritas klaster yang dihitung dengan menggunakan MCS dan nilai batas similaritas minimum . Alur proses klasterisasi hirarkikal dokumen dijelaskan sebagai berikut :

1. Cari pasangan klaster dan yang memiliki nilai kedekatan tertingi, .
2. Jika nilai kedekatan tertinggi klaster dan kurang dari , maka proses dihentikan.
3. Gabung klaster dan menjadi klaster.
4. Perbaharui nilai kedekatan klaster dengan klaster lain pada matriks .
5. Ulangi langkah 1-4.
6. Pra-pemrosesan Teks

Tahap pra-pemrosesan teks bertujuan untuk membersihkan teks, sehingga teks hanya mengandung kata – kata yang diperlukan. Pada penelitian ini, pra-pemrosesan teks terdiri dari konversi ke huruf kecil, penghilangan tanda baca, dan penghilangan *stopwords*.

1. Perhitungan Similaritas Teks menggunakan Maximum Common Subgraph

Metode MCS (*Maximum Common Subgraph*) merupakan metode untuk mencari kesamaan sub-struktur yang optimal sama antara 2 graf. MCS dapat digunakan untuk menghitung similaritas antar teks berbasis graf [24]. Masukkan dari MCS adalah graf representasi teks dan . Pembuatan graf representasi teks telah dijelaskan lebih detil pada subbab 2.5. Keluaran dari MCS adalah sub-graf optimal . Nilai similaritas graf representasi teks dapat dihitung dari hasil sub-graf. Alur metode MCS dapat digambarkan sebagai berikut :

1. Cari *node* yang sama antara dan , dan tambahkan ke .
2. Ambil 2 *node* yang berbeda pada . Jika kedua *node* bersebelahan pada dan , maka *edge* yang menghubungkan kedua *node* tersebut ditambahkan pada . Bobot *edge* terkecil antara bobot *edge* di dan , untuk menjadi bobot *edge* pada .
3. Ulangi langkah 2, sampai tidak ada lagi *edge* yang bisa ditambahkan ke .

Setelah sub-graf optimal didapatkan, similaritas antar graf representasi teks dihitung dengan menggunakan rumus (2), dimana merupakan jumlah *node* pada , merupakan nilai maksimal total *node* pada dan , merupakan jumlah *edge* pada , dan merupakan nilai maksimal total *edge* pada dan . Koefisien merupakan nilai antara 0 dan 1 yang mewakili tingkat kepentingan *node* terhadap *edge* pada sub-graf.

( 2 )

1. Ekstraksi Frasa menggunakan *Frequent Phrase Mining* (FPM)

Pada umumnya dokumen memiliki 2 jenis frasa, yaitu frasa konseptual dan non-konseptual. Frasa yang merepresentasikan topik penelitian merupakan frasa konseptual. El-Kishky et al. mengusulkan penggunaan FPM (*Frequent Phrase Mining*)untuk mengekstraksi frasa dari sebuah korpus [5]. FPM terinspirasi salah satu metode penggalian *association rule* yaitu *frequent itemset mining*.

Metode FPMmemiliki masukkan berupa kumpulan dokumen dan sebuah nilai batas minimum frekuensi . Jika ukuran frasa yang akan diidentifikasi adalah , maka pada iterasi awal nilai . Ukuran akan bertambah seiring dengan bertambahnya iterasi. Metode FPM menghasilkan kumpulan kandidat frasa konseptual dan frekuensi kemunculannya. Alur metode FPM adalah sebagai berikut :

1. Ulangi jika
   1. Ulangi untuk setiap dokumen
      1. Identifikasi seluruh ()-gram pada dokumen , dan simpan indeks aktif ()-gram pada . Indeks aktif merupakan indeks awal sebuah ()-gram pada dokumen . Sebuah ()-gram tersusun atas kata , ()-gram , dimana merupakan indeks aktif ()-gram tersebut.
      2. Hitung frekuensi kemunculan setiap ()-gram pada dokumen .
      3. ()-gram yang memiliki frekuensi kemunculan kurang dari ­­, indeks aktifnya akan dikeluarkan dari .
      4. Ulangi untuk setiap , jika
         1. Jika maka sebuah frasa dibentuk oleh .
         2. Tambahkan penghitung frekuensi frasa , .
         3. Masukkan frasa *P* ke dalam *n*-gram, dan indeks awal frasa *P* ke .
      5. Hilangkan dokumen dari , jika
2. Representasi Teks dalam Graf

Cara paling umum untuk merepresentasikan teks adalah dengan pendekatan VSM (*Vector Space Model*). Pendekatan VSM pada umumnya menggunakan frekuensi kata sebagai fitur teks. Namun, cara tersebut tidak mempertimbangkan informasi semantik dan struktur dari sebuah teks. Model graf dapat merepresentasikan teks secara matematis dengan tetap menjaga informasi semantik dan struktur teks [21].

Sumber teks dapat berupa satu atau banyak dokumen. Setelah melalui tahap pra-pemrosesan teks, kata – kata pada teks tersebut akan menjadi *node*. Relasi kemunculan bersama (*co-occurrence*) antar kata digunakan untuk membentuk *edge* pada graf. Relasi kemunculan bersama telah banyak digunakan untuk menggambarkan hubungan kontekstual antar kata pada teks [28] [24] [22]. *Edge* berjenis *undirected* dan bentang kata pada relasi kemunculan bersama (*co-occurence windows*) sebanyak 2 kata, digunakan pada pembentukan graf berdasarkan hasil pengujian yang telah dilakukan Mihalcea *et al* [10].

Graf representasi teks dapat dinotasikan sebagai , dimana merupakan kumpulan *vertex/node*, merupakan kumpulan *edge* dan merupakan bobot untuk setiap *edge*. Jika merepresentasikan *node* untuk kata , dan merepresentasikan *node* untuk kata , maka sebuah *edge* akan dibentuk jika terdapat kemunculan bersama dan pada maksimum bentang 2 kata.

Bobot *edge* , dinotasikan sebagai , merupakan frekuensi kemunculan bersama yang telah dinormalisasi antara dan pada maksimum bentang 2 kata. Jika merupakan frekuensi kata pada teks, dan merupakan jumlah kata pada teks, maka bobot kata dapat dihitung menggunakan rumus (3). Bobot *edge* dapat dihitung menggunakan rumus (4), dimana frekuensi kemunculan bersama dan pada maksimum bentang 2 kata.

( 3 )

( 4 )

1. TopicRank

TopicRank [14] merupakan metode ekstraksi kata atau frasa yang terinspirasi oleh algoritma penentuan peringkat PageRank [11]. TopicRank menerima masukkan teks yang telah dimodelkan dalam bentuk graf. Teks dapat bersumber dari satu atau banyak dokumen. Graf masukkan tersebut memiliki kata sebagai *node*, dan *edge* berupa relasi antar kata. Pada penelitian ini, *edge* pada graf masukkan berupa kemunculan bersama antar kata (*word co-occurrence*), dan jenis graf merupakan *undirected graph*. Konstruksi graf sebagai masukkan metode TopicRank dijelaskan lebih detil pada subbab 2.6.

Pada dasarnya metode TopicRank menentukan tingkat kepentingan sebuah node kata berdasarkan informasi global pada struktur graf dan persebaran topik dokumen. Metode ini menjalankan PageRank sebanyak jumlah topik yang ada. Untuk setiap perhitungan skor setiap *node*, pembobotan *node* dilakukan dengan menghitung *cosine similarity* antara vektor probabilitas kata terhadap topik dan vektor probabilitas topik terhadap dokumen. Bobot *node* tersebut digunakan untuk mengganti probabilitas *random walk* pada algoritma PageRank. Vektor probabilitas *node* kata dinotasikan dengan , dimana  merupakan probabilitas kata pada topik . Vektor probabilitas topik pada dokumen dinotasikan dengan , dimana merupakan probabiltas topik pada dokumen *d*. Bobot *node* dinotasikan dengan dihitung dengan rumus (5).

( 5 )

Jika sebuah graf dilambangkan sebagai , dimana  adalah kumpulan *vertex/node*, adalah kumpulan *edge*, adalah kumpulan bobot *node*, dan adalah kumpulan bobot *edge*. Notasi *N* merepresentasikan jumlah *node* yang terdapat pada graf. Untuk setiap *node* , notasi merupakan kumpulan *node* yang mengarah ke *node* , dan notasi merupakan kumpulan *node* tujuan *node* . Relasi antara *node*  dan memiliki bobot *edge* yang dilambangkan dengan . Bobot *node* dinotasikan dengan . Maka, skor *node* pada topik *z*, dinotasikan dengan , akan diperbaharui setiap iterasi dengan rumus (6).

( 6 )

Pada rumus (6), merupakan koofisien *damping*, yang memiliki nilai antara 0 dan 1. Koefisien *damping* mewakili kemungkinan loncatan dari sebuah *node* ke *node* acak. Dalam konteks pe nelusuran web, koefisien *damping* menggambarkan probabilitas, sebesar , pengguna untuk memilih *link* yang tersedia pada halaman tersebut, dan probabilitas, sebesar , pengguna untuk pergi ke halaman web yang benar – benar acak. Implementasi koefisien *damping* dapat disebut juga *“Random Surfer Model”*.

Pasca-pemrosesan akan dilakukan setelah konvergensi tercapai, yaitu jika nilai bobot *node* sudah tidak banyak mengalami perubahan. Pada pasca-pemrosesan, sejumlah *n* kata yang memiliki skor bobot *node* terbesar akan dipilih. Setiap *n* kata akan diperiksa kumpulan *in-degree node* dan *out-degree node* milik *node* kata tersebut, untuk mencari kata lain yang terletak bersebelahan pada dokumen asal. Jika ditemukan, maka kata tersebut akan digabung menjadi sebuah frasa. Keluaran tahap pasca-pemrosesan adalah kumpulan kata dan frasa yang dianggap merepresentasikan korpus sumber.

1. Koherensi Topik

Koherensi topik mengukur interpretabilitas topik dari set label frasa yang dihasilkan pelabelan klaster artikel ilmiah. Suatu set label frasa koheren secara topikal jika antara satu label dengan label lainnya memiliki keterkaitan kontekstual. Pada penelitian ini set label frasa yang dihasilkan oleh pelabelan klaster untuk merepresentasikan topik bahasan klaster suatu artikel ilmiah dievaluasi menggunakan koherensi topik. Roder *et al* telah mengusulkan pengukuran koherensi topik yang memiliki performa lebih baik dibandingkan usulan metode koherensi topik sebelumnya dalam hal korelasinya dengan penilaian manusia [29].

Misal proses pelabelan klaster menghasilkan sebanyak *n* label frasa yang merepresentasikan topik bahasan klaster artikel ilmiah ke-*i* yang dinotasikan sebagai . Set kata yang merupakan kumpulan kata unik yang menyusun frasa – frasa pada label di direpresentasikan dalam notasi . Setiap kata direpresentasikan ke dalam vektor ruang kata yang memiliki dimensi sebesar jumlah kata *m*. Elemen vektor ke-*j* yang dinotasikan dengan dihitung dengan (9) yang merupakan perhitungan *Normalized Pointwise Mutual Information* (NPMI). Notasi pada () merupakan nilai bias yang ditentukan sendiri. merupakan probabilitas kemunculan kata pada rentang jendela kata (*sliding window*) yang dinotasikan dengan . merupakan probabilitas kemunculan bersama kata dan pada . Probabilitas kemunculan kata dihitung berdasarkan korpus *Wikipedia* yang terdiri dari artikel *Wikipedia* dari tahun 2009. Nilai dan yang digunakan adalah 110 dan 1 sesuai dengan penelitian pada [29].

(7)

(8)

(9)

Jika kumpulan vektor konteks kata , maka kombinasi pasangan vektor konteks kata disimpan dalam . Koherensi topik label klaster dihitung dengan (10) dimana merupakan cosine similarity dari pasangan vektor konteks kata *S∈.*

(10)

# BAB 3 METODOLOGI PENELITIAN

Bab ini akan memaparkan tentang metodologi penelitian yang digunakan pada penelitian ini, yang terdiri dari (1) studi literatur, (2) desain dan implementasi, (3) pengujian, dan (4) dokumentasi dan pembuatan laporan. Ilustrasi alur metodologi penelitian dapat dilihat pada Gambar 3.1.

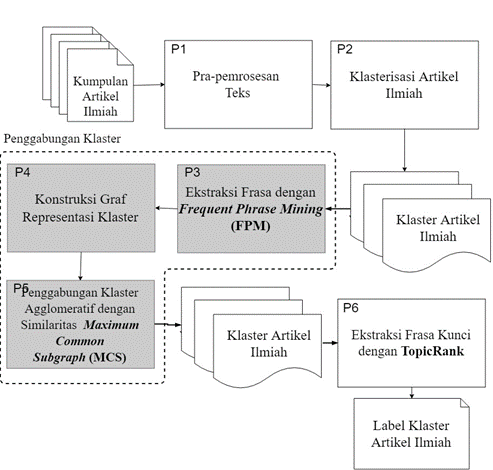
C:\Users\adhin\AppData\Local\Microsoft\Windows\INetCache\Content.Word\Metodologi Penelitian.jpg

Gambar 3.1 Alur Metodologi Penelitian

1. Studi Literatur

Tahap studi literatur bertujuan untuk mengumpulkan referensi - referensi yang dapat menunjang penelitian. Sumber referensi dapat berupa jurnal ilmiah atau buku teks. Referensi yang dikumpulkan berhubungan dengan metode pemrosesan teks yang dapat dipakai untuk pelabelan klaster khususnya pada dokumen artikel ilmiah. Referensi tersebut digunakan untuk merumuskan permasalahan yang menjadi landasan dilakukannya penelitian ini dan solusi yang akan diusulkan. Berdasarkan studi literatur yang telah dilakukan, informasi yang berkaitan dengan penilitian yang dilakukan ini, seperti berikut :

1. Frasa lebih deskriptif dibandingkan kata dalam melabeli klaster dokumen.
2. Hasil dari proses klasterisasi mempengaruhi proses pelabelan klaster.
3. Pendekatan Vector Space Model pada proses klasterisasi tidak mempertimbangkan hubungan antar kata pada teks.
4. Graf dapat memberikan informasi yang lebih mendalam mengenai hubungan antar kata di dalam sebuah dokumen.
5. Perancangan Sistem

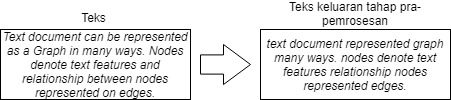


Gambar 3.2 Alur Proses Metode Usulan

Alur proses metode pelabelan klaster usulan terdiri dari beberapa tahap, seperti pada Gambar 3.2. Pertama, tahap pra-pemrosesan teks dilakukan pada kumpulan artikel ilmiah sebelum proses klasterisasi dokumen dilakukan. Kedua, pada setiap klaster artikel ilmiah, ekstraksi frasa dilakukan untuk mendapatkan frasa – frasa yang merepresentasikan topik. Penggabungan klaster dilakukan pada setiap klaster melalui 2 tahap, yaitu konstruksi graf representasi klasters dan penggabungan klaster menggunakan pengukuran similaritas dengan *Maximum Common Subgraph* (MCS). Pada tahap konstruksi graf representasi klaster, untuk setiap klaster artikel ilmiah, kata – kata yang menyusun frasa – frasa dari keluaran tahap ekstraksi frasa digunakan untuk menjadi *node*, dan relasi kemunculan bersama (*co-occurrence*) antar kata digunakan untuk menghitung bobot *edge*. Perhitungan similaritas dilakukan menggunakan metode MCS pada graf – graf klaster yang telah dibuat. Klaster – klaster yang memiliki nilai similaritas MCS yang lebih dari nilai batas akan digabungkan pada tahap penggabungan klaster. Pelabelan klaster dilakukan untuk setiap klaster hasil proses penggabungan klaster. Konstruksi graf representasi klaster akan dilakukan kembali menggunakan klaster hasil penggabungan untuk menjadi masukkan metode ekstraksi frasa kunci *TopicRank*. Keluaran dari proses ekstraksi frasa kunci adalah *n* label frasa terbaik untuk merepresentasikan klaster artikel ilmiah. Tahap – tahap proses metode usulan akan dijelaskan pada subbab berikut.

1. Pra-pemrosesan Teks

Tahap pra-premosesan teks bertujuan untuk mempersiapkan data teks untuk dapat diproses pada tahap selanjutnya. Tahap pra-pemrosesan teks terdiri dari konversi ke huruf kecil, penghilangan tanda baca, dan penghilangan *stopwords*. Contoh dari tahap pra-pemrosesan teks dapat dilihat pada Gambar 3.3.



Gambar 3.3 Contoh Tahap Pra-pemrosesan Teks

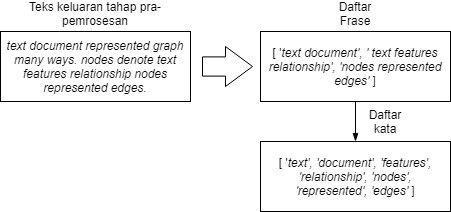
1. Klasterisasi Artikel Ilmiah

Tahap klasterisasi artikel ilmiah bertujuan untuk membagi kumpulan dokumen artikel ilmiah ke beberapa klaster. Pada penelitian ini, metode K-Means akan digunakan untuk klasterisasi dokumen artikel ilmiah. Pengukuran similaritas antar dokumen akan menggunakan *Cosine Similarity*. Penjelasan metode klasterisasi K-Means dengan *Cosine Similarity* dijelaskan pada subbab 2.1.1.

1. Ekstraksi Frasa Topik dengan *Frequent Phrase Mining* (FPM)

Tahap ekstraksi frasa topik bertujuan untuk mengurangi jumlah kata pada setiap klaster artikel ilmiah dengan mengidentifikasi kata – kata yang berpotensi untuk menyusun frasa topik. Tahap ini memiliki masukkan berupa seluruh teks yang terdapat pada klaster artikel ilmiah dan memprosesnya menjadi daftar kata – kata pembentuk frasa yang merepresentasikan topik.

Identifikasi frasa topik dilakukan menggunakan metode FPM. Metode FPM terinspirasi dari algoritma *apriori* pada *frequent item set mining*. Metode FPM bekerja dengan mengidentifikasi *n-gram* dimulai dari nilai *n* terkecil yaitu sampai *n* maksimal yang masih memenuhi nilai batas yang telah ditentukan. Alur proses metode FPM telah dijelaskan pada subbab 2.4. Contoh hasil dari tahap ekstraksi frasa dapat dilihat pada Gambar 3.4.



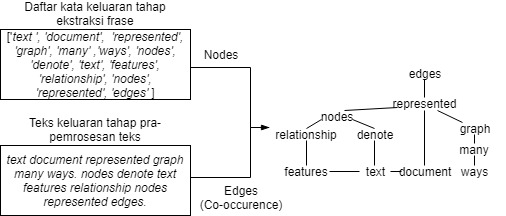
Gambar 3.4 Contoh Tahap Ekstraksi Frasa Topik

1. Penggabungan Klaster

Proses penggabungan klaster bertujuan untuk menggabungkan klaster yang memiliki kemiripan topik. Klaster – klaster yang memiliki similaritas tinggi akan digabungkan dengan pendekatan graf. Pertama, setiap klaster akan direpresentasikan dalam bentuk graf. Lalu, similaritas antar graf representasi klaster dihitung menggunakan *Maximum Common Subgraph* (MCS). Tahap – tahap pada proses penggabungan klaster dijelaskan pada subbab berikut.

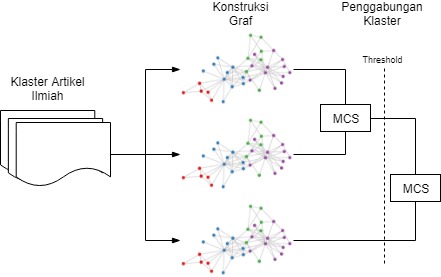
* 1. *Konstruksi Graf Representasi Klaster*

Tujuan tahap konstruksi graf representasi klaster adalah untuk membuat graf yang dapat memetakan hubungan antar kata di dalam sebuah klaster artikel ilmiah. Contoh tahap konstruksi graf dapat dilihat pada Gambar 3.5. Masukkan pada tahap konstruksi graf adalah daftar kata keluaran tahap ekstraksi frasa pada sebuah klaster artikel ilmiah. Seluruh kata pada klaster akan menjadi *node* dan relasi *co-occurrence* antar kata akan menjadi *edge*. Detil konstruksi graf representasi klaster dijelaskan pada subbab 2.5.



Gambar 3.5 Contoh Tahap Konstruksi Graf

* 1. *Penggabungan Klaster dengan Pengukuran similaritas Maximum Common Subgraph*

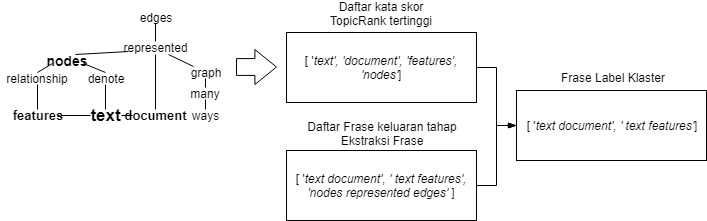


Gambar 3.6 Contoh Penggabungan Klaster

Tahap ini bertujuan untuk menggabungkan klaster – klaster yang memiliki kemiripan topik. Masukkan dari tahap ini adalah graf – graf representasi klaster. Matriks similaritas antar klaster dibuat dengan pengukuran similaritas MCS (*Maximum Common Subgraph*) syang dijelaskan pada subbab 2.3. Klaster - klaster yang memiliki nilai similaritas tinggi digabung secara iteratif, seperti ilustrasi pada Gambar 3.6, secara hirarkikal yang dijelaskan pada subbab 2.1.2. Penggabungan klaster berhenti jika nilai similaritas antar klaster tertinggi kurang dari nilai batas similaritas yang ditentukan.

1. Pelabelan Klaster

Pada tahap ini graf representasi klaster dibuat dari klaster hasil penggabungan sebagai masukkan pada tahap ekstraksi frasa kunci menggunakan *TopicRank*. *TopicRank* memberikan skor pada *node* secara iteratif berdasarkan struktur graf dan distribusi topik yang didapatkan dengan metode *Latent Dirchlet Allocation* (LDA). Kata – kata dengan nilai tertinggi akan dicocokan dengan daftar frasa klaster artikel ilmiah. Jika frasa tersusun atas kata – kata yang memiliki nilai *TopicRank* tertinggi, maka frasa tersebut dapat merepresentasikan label klaster artikel ilmiah. Detil metode *TopicRank* dijelaskan pada subbab 2.6. Contoh hasil penilaian *TopicRank* dapat dilihat pada Gambar 3.7.



Gambar 3.7 Contoh Hasil TopicRank

1. Evaluasi
2. Dataset

Data artikel ilmiah yang digunakan untuk tahap pengujian adalah dataset *Citation* dari basis data publikasi ilmiah *AMiner*. Seterusnya dataset tersebut disebut sebagai dataset asli. Dataset asli tersusun atas artikel ilmiah dari jurnal *IEEE Transaction of Computers* beserta artikel ilmiah kutipannya dan berjumlah ±12.000 artikel ilmiah. Dataset asli akan dibagi menjadi dua dataset yaitu dataset 1 dan dataset 2 berdasarkan tingkat keberagamannya.

1. Skenario Evaluasi

Tahapan selanjutnya setelah implementasi usulan metode dilakukan adalah tahap pengujian. Tahap pengujian bertujuan untuk mengevaluasi performa usulan metode dan kesesuaian terhadap tujuan penelitian. Performa usulan metode akan diukur berdasarkan kualitas label frasa klaster menggunakan koherensi topik. Pada tahap pengujian dilakukan beberapa skenario pengujian yaitu :

* + - 1. Membandingkan kualitas label frasa dengan dan tanpa implementasi usulan metode pada dataset 1 dan dataset 2 dengan nilai parameter FPM yang berbeda.
      2. Membandingkan persentase ukuran graf MCS pada dataset 1 dan dataset 2 terhadap kualitas label frasa.

*[Halaman ini sengaja dikosongkan]*

# BAB 4 UJI COBA & ANALISIS HASIL

1. Lingkungan Uji Coba

Implementasi dan uji coba pada penelitian ini dilakukan pada perangkat keras dengan spesifikasi seperti berikut :

* 1. Sistem Operasi Windows 10 64-bit
  2. *Processor* Intel i5-4200U CPU @ 1.60GHz (4 CPUs)
  3. Kapasistas RAM 8GB
  4. Kapasitas *harddisk* 500GB

Perangkat lunak pendukung dalam implementasi usulan metode pada penelitian ini adalah seperti berikut :

* + - * 1. Python 2.7 dan Python 3.6
        2. Basis Data *MySQL* 5.7

1. Persiapan Data

Dataset asli merupakan kumpulan artikel ilmiah yang bersumber dari jurnal *IEEE Transaction of Computers* beserta artikel ilmiah kutipannya yang memiliki total jumlah ±12.000 artikel ilmiah. Observasi awal dilakukan pada dataset asli dengan melakukan analisis *silhouette* untuk menentukan jumlah klaster optimal. Hasil observasi awal pada dataset asli dapat dilihat pada Gambar 4.1 yang menunjukkan dataset asli tidak memiliki jumlah klaster optimal. Hal tersebut disebabkan oleh banyaknya data – data artikel ilmiah yang tidak memiliki kemiripan oleh klaster manapun (ditandai dengan nilai *silhouette* yang rendah) atau biasa disebut data derau (*outlier*). Pada dataset asli dengan jumlah klaster 11 terdapat 51.16% data artikel ilmiah yang memiliki nilai silhouette kurang dari 0.1. Oleh karena itu dataset asli dibersihkan dari data derau dengan 2 cara :

* + - 1. Memilih 50 data artikel ilmiah dengan nilai *silhouette* tertinggi pada tiap klaster dari hasil klasterisasi dataset asli dengan jumlah klaster 11. Hasil pemilihan data tersebut berjumlah 550 data artikel ilmiah dan seterusnya disebut sebagai **dataset 1**.
      2. Memilih data artikel ilmiah dengan nilai *silhouette* lebih dari 0.1 dari hasil klasterisasi dataset asli dengan jumlah klaster 17. Hasil pemilihan data tersebut berjumlah ±5500 artikel ilmiah dan seterusnya disebut sebagai **dataset 2**.

Hasil analisis *silhouette* pada dataset 1 dan 2 pada Gambar 4.1 menunjukkan bahwa dataset 1 memiliki rata – rata nilai *silhouette* yang lebih tinggi dibandingkan dataset 2, sehingga dataset 1 bersifat homogen dan dataset 2 bersifat heterogen. Gambar 4.1 juga menunjukkan bahwa jumlah klaster optimal pada dataset 1 dan dataset 2 adalah 11 dan 14 berturut – turut.

Gambar 4.1. Analisis *Silhouette* Dataset Asli, Dataset 1, dan Dataset 2

1. Hasil Uji Coba
   1. Skenario 1

Skenario 1 bertujuan untuk membandingkan hasil pelabelan klaster pada dataset 1 dan dataset 2 dalam hal koherensi topik, dengan menggunakan parameter *support* metode *Frequent Phrase Mining* (FPM) yang bebeda. Untuk setiap dataset, koherensi topik set label klaster hasil penggabungan (klaster gabungan) dibandingkan dengan koherensi topik set label klaster sebelum dilakukan penggabungan (klaster asli).

Pada Tabel 4.1 percobaan pertama dilakukan pada dataset 1 dengan nilai parameter *support* metode *Frequent Phrase Mining* (FPM) 3. Hasil percobaan 1 menunjukkan rata – rata koherensi topik pada klaster gabungan lebih rendah daripada rata – rata koherensi topik pada klaster asli. Nilai parameter *support* FPM menentukan seberapa banyak kata yang akan dijadikan sebagai *node*/*vertex* pada graf representasi klaster, semakin banyak kata yang digunakan maka semakin besar ukuran graf. Pada percobaan 2 usulan metode dilakukan pada dataset 1 dengan menggunakan nilai parameter *support* FPM 2 dengan tujuan untuk membentuk graf representasi klaster yang berukuran lebih besar. Hasil percobaan 2 menunjukkan bahwa dengan memperbesar ukuran graf klaster, rata – rata koherensi topik klaster gabungan lebih besar daripada rata – rata koherensi topik klaster asli. Percobaan 3 mengimplementasi usulan metode pada dataset 2 dengan nilai *support* FPM 3. Hasil percobaan 3 menunjukkan rata – rata koherensi topik klaster gabungan lebih besar daripada rata – rata koherensi topik klaster asli. Perlu dicatat bahwa jumlah data pada dataset 2 lebih besar daripada dataset 1 yang menyebabkan ukuran graf klaster pada dataset 2 lebih besar daripada dataset 1 pada penggunaan parameter *support* FPM 3. Hasil skenario 1 menunjukkan bahwa peningkatan rata – rata koherensi topik label klaster terjadi pada dataset 1 dan dataset 2 menggunakan parameter *support* FPM 2 dan 3 berturut – turut.

Tabel 4.1. Hasil Rata - Rata Koherensi Topik Skenario 1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Percobaan** | **Dataset** | ***Support* FPM** | **Klaster Asli** | | **Klaster Hasil Penggabungan** | |
| **Jumlah Klaster** | **Rata - Rata Koherensi Topik** | **Jumlah Klaster** | **Rata - Rata Koherensi Topik** |
| 1 | 1 | 3 | 11 | **0.443** | 8 | 0.426 |
| 2 | 2 | 11 | 0.443 | 6 | **0.45** |
| 3 | 2 | 3 | 14 | 0.449 | 8 | **0.452** |

Hasil penggabungan dan pelabelan klaster pada percobaan 1 dapat dilihat di Tabel 4.2. Pada Tabel 4.2 terlihat bahwa klaster gabungan KG-1 tersusun atas klaster asli KA-4, KA-5, dan KA-7. Penggabungan klaster tersebut menghasilkan set label klaster dengan nilai koherensi topik lebih kecil daripada rata - rata koherensi topik set label klaster penyusunnya. Nilai koherensi topik yang kecil pada klaster gabungan tercermin pada label klaster gabungan ‘*test application time*’ dan ‘*system fault diagnosis*’ yang sulit untuk diiterpretasi sebagai satu topik.

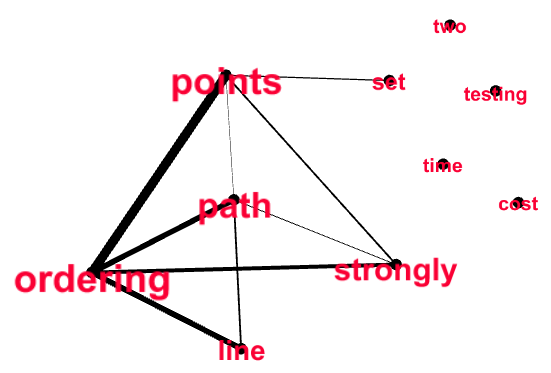
Tabel 4.2. Hasil Penggabungan dan Pelabelan Klaster Percobaan 1

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Klaster Gabungan** | | | **Klaster Asli** | | | | |
| **ID Klaster** | **Label Klaster** | **Koherensi Topik** | **ID Klaster** | **Label Klaster** | **Koherensi Topik** | **Koherensi Topik Gabungan** |
| KG-1 | • test generation algorithm  • test generation time  • new fault model  • test application time  • system fault diagnosis | 0.386 | KA-4 | test pattern generation, test generator circuit, test generation method, test generation time, test generation algorithm | 0.390 | 0.463 |
| KA-5 | polynomial time algorithm, n algorithm, n log n, time algorithm, best such algorithm | 0.569 |
| KA-7 | system fault diagnosis, new fault model, fault injection techniques, transient fault tolerance, rtl fault model | 0.430 |

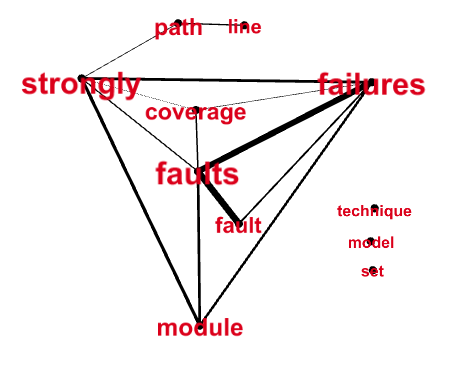
Tabel 4.3. Perbandingan Ukuran Graf Terhadap Rata- Rata Koherensi Topik

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Percobaan** | **Dataset** | **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 | 1123 | 5.29 % (4 *vertex*) | 0.15 % (2 *edge*) | (0 , 11) | **0.443** | 0.426 |
| 2 | 2 | 203 | 7137 | 12.65 % (26 vertex) | 0.60 % (43 *edge*) | (8 , 45) | 0.443 | **0.45** |
| 3 | 2 | 3 | 379 | 34,305 | 18.89 % (72 vertex) | 1.40 % (479 *edge*) | (21 , 211) | 0.449 | **0.452** |

Pada usulan metode, similartias antar klaster direpresentasikan dengan graf *Maximum Common Subgraph* (MCS) antar klaster. Gambar 4.2 memperlihatkan graf MCS antara klaster asli KA-4 dan KA-7. Pada graf MCS tersebut relasi antara kata ‘*fault*’ dan ‘*coverage*’ menunjukkan bahwa klaster asli KA-4 dan KA-7 memiliki kesamaan pada topik ‘*fault coverage*’. Sementara graf MCS antara klaster asli KA-5 dan KA-7 yang ditunjukkan oleh Gambar 4.3 sulit diinterpretasi ke dalam sebuah topik. Hal ini disebabkan oleh ukuran graf MCS yang teridentifikasi tidak memadai untuk menangkap similaritas topik antar klaster.



Gambar 4.3. Graf MCS Klaster Asli KA-5 dan KA-7 pada Percobaan 1



Gambar 4.2. Graf MCS Klaster Asli KA-4 dan KA-7 pada Percobaan 1

* 1. Skenario 2

Pada skenario 1 dengan memperbesar ukuran graf representasi klaster melalui parameter *support* FPM dapat meningkatkan rata – rata koherensi topik klaster gabungan. Skenario 2 mengamati pengaruh perubahan ukuran graf tersebut terhadap rata – rata koherensi topik hasil label frasa. Pada Tabel 4.3 terlihat dengan memperbesar ukuran graf klaster asli melalui parameter *support* FPM, ukuran graf *Maximum Common Subgraph* (MCS) yang dihasilkan juga meningkat. Meningkatnya ukuran graf MCS juga diikuti dengan meningkatnya rata – rata koherensi topik. Graf MCS merepresentasikan kemiripan klaster berdasarkan informasi semantik teks dalam relasi kata yang dimiliki. Sehingga hasil skenario 2 menunjukkan bahwa semakin banyak informasi semantik yang digunakan untuk menentukan kemiripan klaster, semakin baik rata – rata koherensi topik label frasa klaster gabungan yang dihasilkan.

# BAB 5 PENUTUP

Berdasarkan pengujian pada metode pelabelan klaster artikel ilmiah yang telah diusulkan, dapat ditarik beberapa kesimpulan dan saran penelitian yang akan dilakukan selanjutnya.

1. Kesimpulan
   1. Persentase ukuran graf MCS yang teridentifikasi mempengaruhi koherensi topik label frasa klaster yang dihasilkan.
   2. Jika data yang digunakan tidak banyak, maka nilai *support* FPM perlu disesuaikan untuk dapat mengidentifikasi graf MCS dengan ukuran yang memadai. Sehingga koherensi topik label klaster gabungan dapat lebih baik dari label klaster asli.
2. Saran

Pada metode penggabungan klaster berbasis graf yang diusulkan, pengukuran similaritas klaster yang digunakan tidak mempertimbangkan bobot *node*/*vertex*. Bobot *node*/*vertex* merepresentasikan tingkat kepentingan kata tersebut terhadap klaster. Sehingga pada penelitian selanjutnya metode pengukuran klaster berbasis graf dengan mempertimbangkan bobot *node*/*vertex* perlu dikembangkan.

*[Halaman ini sengaja dikosongkan]*

# DAFTAR PUSTAKA

[1] Z. Li, J. Li, Y. Liao, S. Wen, and J. Tang, “Labeling clusters from both linguistic and statistical perspectives: A hybrid approach,” *Knowledge-Based Syst.*, vol. 76, pp. 219–227, 2015.

[2] D. Carmel, H. Roitman, and N. Zwerdling, “Enhancing cluster labeling using wikipedia,” *Proc. 32nd Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. - SIGIR ’09*, no. January 2016, pp. 139–146, 2009.

[3] P. Lopez and L. Romary, “HUMB : Automatic Key Term Extraction from Scientific Articles in GROBID,” *Proc. 5th Int. Work. Semant. Eval.*, no. July, pp. 248–251, 2010.

[4] P. Lopez and L. Romary, “GRISP : A Massive Multilingual Terminological Database for Scientific and Technical Domains,” *Knowl. Creat. Diffus. Util.*, pp. 2269–2276, 2010.

[5] A. El-Kishky, Y. Song, C. Wang, C. Voss, and J. Han, “Scalable Topical Phrase Mining from Text Corpora,” *Proc. VLDB Endow.*, vol. 8, no. 3, pp. 305–316, 2014.

[6] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules in Large Databases,” *J. Comput. Sci. Technol.*, vol. 15, no. 6, pp. 487–499, 1994.

[7] C. Aalla and V. Pudi, “Mining Research Problems from Scientific Literature,” *2016 IEEE Int. Conf. Data Sci. Adv. Anal.*, pp. 351–360, 2016.

[8] K. S. Hasan and V. Ng, “Automatic Keyphrase Extraction: A Survey of the State of the Art,” *Proc. 52nd Annu. Meet. Assoc. Comput. Linguist. (Volume 1 Long Pap.*, pp. 1262–1273, 2014.

[9] L. H. Suadaa and 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*, vol. 4, no. c, 2016.

[10] R. Mihalcea and P. Tarau, “TextRank: Bringing order into texts,” *Proc. EMNLP*, vol. 85, pp. 404–411, 2004.

[11] L. Page, S. Brin, R. Motwani, and T. Winograd, “The PageRank Citation Ranking: Bringing Order to the Web,” *World Wide Web Internet Web Inf. Syst.*, vol. 54, no. 1999–66, pp. 1–17, 1998.

[12] X. Wan and J. Xiao, “CollabRank: towards a collaborative approach to single-document keyphrase extraction,” *Proc. 22nd Int. Conf. Comput. Linguist. Coling 2008*, no. August, pp. 969–976, 2008.

[13] Z. Liu, W. Huang, Y. Zheng, and M. Sun, “Automatic Keyphrase Extraction via Topic Decomposition,” *Comput. Linguist.*, no. October, pp. 366–376, 2010.

[14] L. Sterckx, T. Demeester, J. Deleu, and C. Develder, “Topical Word Importance for Fast Keyphrase Extraction,” *Proc. 24th Int. Conf. World Wide Web - WWW ’15 Companion*, no. 2, pp. 121–122, 2015.

[15] Q. Mei, X. Shen, and C. Zhai, “Automatic labeling of multinomial topic models,” *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD ’07*, p. 490, 2007.

[16] N. Y. Saiyad, H. B. Prajapati, and V. K. Dabhi, “A Survey of Document Clustering using Semantic Approach,” *Int. Conf. Electr. Electron. Optim. Tech.*, vol. 6, no. 4, pp. 2555–2562, 2016.

[17] L. Xiong, “Survey on text clustering algorithm,” *2011 IEEE 2nd Int. Conf. Softw. Eng. Serv. Sci.*, no. 4, pp. 901–904, 2011.

[18] A. Krauza, “Extension of fuzzy Gustafson-Kessel algorithm based on adaptive cluster merging,” *2015 IEEE MIT Undergrad. Res. Technol. Conf. URTC 2015*, pp. 0–3, 2016.

[19] F. De Morsier, D. Tuia, M. Borgeaud, V. Gass, and J. P. Thiran, “Cluster validity measure and merging system for hierarchical clustering considering outliers,” *Pattern Recognit.*, vol. 48, no. 4, pp. 1474–1485, 2015.

[20] I. Czarnowski, P. Jċdrzejowicz, and I. Member, “Consensus-based Cluster Merging for the Prototype Selection,” in *2013 IEEE International Conference on Cybernetics*, 2013.

[21] S. Sonawane and P. Kulkarni, “Graph based Representation and Analysis of Text Document: A Survey of Techniques,” *Int. J. Comput. Appl.*, vol. 96, no. 19, pp. 1–8, 2014.

[22] F. Role and M. Nadif, “Beyond cluster labeling: Semantic interpretation of clusters’ contents using a graph representation,” *Knowledge-Based Syst.*, vol. 56, pp. 141–155, 2014.

[23] A. Perer and B. Shneiderman, “Integrating Statistics and Visualization: Case Studies of Gaining Clarity during Exploratory Data Analysis,” *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, p. 265, 2008.

[24] C. Jin and Q. Bai, “Text Clustering Algorithm Based on the Graph Structures of Semantic Word Co- occurrence,” *2016 Int. Conf. Inf. Syst. Artif. Intell.*, 2016.

[25] H. Gupta and R. Srivastava, “K-means Based Document Clustering with Automatic ‘ K ’ Selection and Cluster Refinement,” *Int. J. Comput. Sci. Mob. Appl.*, vol. 2, pp. 7–13, 2014.

[26] J. De Knijff, F. Frasincar, and F. Hogenboom, “Data & Knowledge Engineering Domain taxonomy learning from text : The subsumption method versus hierarchical clustering,” *Data Knowl. Eng.*, vol. 83, no. 0, pp. 54–69, 2013.

[27] S. K. Popat and M. Emmanuel, “Review and Comparative Study of Clustering Techniques,” *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 1, pp. 805–812, 2014.

[28] J. Jayabharathy, S. Kanmani, and A. A. Parveen, “Document clustering and topic discovery based on semantic similarity in scientific literature,” *Commun. Softw. Networks (ICCSN), 2011 IEEE 3rd Int. Conf.*, pp. 425–429, 2011.

[29] M. Röder, A. Both, and A. Hinneburg, “Exploring the Space of Topic Coherence Measures,” *Proc. Eighth ACM Int. Conf. Web Search Data Min. - WSDM ’15*, pp. 399–408, 2015.

[30] S. R. El-Beltagy and A. Rafea, “KP-Miner: Participation in SemEval-2,” *Proc. 5th Int. Work. Semant. Eval.*, no. July, pp. 190–193, 2010.

[31] J. H. Lau, K. Grieser, D. Newman, and T. Baldwin, “Automatic Labelling of Topic Models,” *Proc. 49th Annu. Meet. Assoc. Comput. Linguist.*, pp. 1536–1545, 2011.