Jurnal Pseudocode, Volume V Nomor 2, September 2018, ISSN 2355-5920
www.ejournal.unib.ac.id/index.php/pseudocode

DOCUMENT CLUSTERING DENGAN *LATENT DIRICHLET ALLOCATION* DAN *WARD HIERARICHAL CLUSTERING*

Guntur Budi Herwanto[1]

[1]Ilmu Komputer /Departemen Ilmu Komputer dan Elektronika,
Fakultas Matematik dan Ilmu Pengetahuan Alam, Universitas Gadjah Mada
Sekretariat Ilmu Komputer UGM Sekip Unit III FMIPA Gedung Selatan, Yogyakarta 55281
(+62 (274) 546194; +62 (274) 556194)

[1]gunturbudi@mail.ugm.ac.id

Abstrak: Saat ini konten informasi dalam bentuk berita dihasilkan dengan jumlah besar dari berbagai sumber setiap harinya. Banyaknya konten yang dihasilkan ini menuntut organisasi konten yang baik agar pencarian informasi yang diinginkan dapat dilakukan dengan mudah. Organisasi dan manajemen informasi yang efisien terhadap konten informasi elektronik ini menginspirasi penelitian mengenai document clustering. Pada penelitian ini dirancang sebuah metode document *clustering* dengan melakukan kombinasi pemodelan topik *latent dirichlet allocation* (LDA) dengan *ward hierarchical clustering*. LDA digunakan sebagai representasi vektor dokumen yang berupa distribusi topik. Representasi ini bertujuan untuk mengurangi dimensi vektor yang pada umumnya terlalu panjang jika menggunakan tf-idf. *Ward Hierarchical Clustering* yang memiliki kompleksitas tinggi dapat terbantu prosesnya dengan representasi dari LDA. Selain itu dihasilkan *silhouette coefficient* yang baik yaitu 0.7. Dalam peneltian ini juga ditemukan bahwa penentuan jumlah topik dalam kaitannya dengan *document clustering* dapat dilakukan dengan mempertimbangkan *silhouette coefficient* pada hasil *clustering*. Performa silhouette coefficient pada representasi pemodelan topik lebih baik dibandingkan dengan representasi dengan tf-idf.
Kata Kunci: *document clustering*, *latent dirichlet allocation*, *ward hierarchical clustering*, pemodelan topik

Abstract: Currently online content in the form of text generated in an enormous amount every day. Organizing this content according to their topics can help readers find information relevant to them. Efficient organization and management of information on electronic information content inspire research on document clustering. In this study, a document clustering method is designed by combining Topic Modeling Latent Dirichlet Allocation (LDA) with Ward Hierarchical Clustering. LDA is used as a representation of document vectors in the form of topic distribution. This representation aims to reduce the dimensions of vectors that are too long in commonly used method tf-idf. Such dimension will be helpful for Ward Hierarchical Clustering that has a high complexity. This representation also shows the promising result on the silhouette coefficient, which is 0.7. In this study, we also found that the determination of the number of topics in relation to document clustering can be done by considering silhouette coefficient on the clustering results.

Keywords: document clustering, latent dirichlet allocation, ward hierarchical clustering, topic modelling

I. PENDAHULUAN

Saat ini konten infomasi dihasilkan dengan jumlah yang cukup banyak setiap harinya. New York Times sendiri menerbitkan 300 artikel setiap harinya [1]. Banyaknya konten yang dihasilkan ini menuntut organisasi konten yang baik agar pencarian terhadap sebuah informasi yang diinginkan dapat dilakukan dengan mudah. Organisasi dan manajemen informasi yang efisien terhadap konten informasi elektronik ini menginspirasi penelitian mengenai *document clustering* [2]. *Document clustering* adalah proses membagi sekumpulan teks dokumen ke dalam kelompok-kelompok kecil berdasarkan kesamaan antar dokumen [3] . *Document clustering* bertujuan

untuk membantu manusia dalam pencarian dan pemahaman pada sekumpulan informasi [3].

Beberapa algoritme dalam *document clustering* seperti K-Means, Hierarchical Agglomerative Clustering, Singular Value Decomposition telah diteliti sebelumnya [4]. Seluruh algoritme tersebut dapat membentuk beberapa cluster dari kumpulan dokumen tanpa adanya bantuan manusia (*unsupervised*). Namun, permasalahan yang muncul dalam algoritme *unsuperivsed* tersebut adalah tidak dapat dihasilkannya topik utama dari cluster [5].

Sebuah dokumen pada dasarnya memiliki beberapa topik yang terkandung di dalamnya. Algoritme pemodelan topik *Latent Dirichlet Allocation* (LDA) dapat melihat intuisi ini [6]. Pemodelan topik dan *document clustering* merupakan dua hal yang sangat dekat hubungannya, dan dapat dilakukan integrasi satu sama lain [7]. Penggabungan ini telah dilakukan sebelumnya dalam beberapa penelitian [7], [8]. LDA dan *Self Organizing Map* (SOM) terbukti dapat menampilkan cluster sekaligus visualisasi yang intuituif [8].

Document clustering merupakan teknik *text* mining yang digunakan untuk mengelompokan dokumen yang memiliki kemiripan ke dalam sebuah kluster tunggal [9] [10]. *Document clustering* adalah salah satu tugas yang paling sering digunakan pada bidang *data mining*, *information retrieval*, *knowledge discovery* dari data, pengenalan pola dan sebagainya, karena kemampuannya dalam merangkum koleksi data yang besar [11] [10]. Tujuan dari *document clustering* ini adalah untuk menemukan kelompok dokumen yang saling berhubungan dalam koleksi dokumen yang besar [11], selain itu digunakan untuk membantu manusia dalam proses pencarian dan pemahaman informasi [10].

Pada penelitian yang dilakukan oleh [10] disebutkan bahwa proses *document clustering* dapat dilakukan dengan cara tradisional dan semantik. Proses *document clustering* secara tradisional dilakukan dengan pendekatan model Bag of Words (BoW) untuk mencari kata kunci yang sering muncul dalam sebuah dokumen. Namun pendekatan ini memiliki kelemahan yaitu mengabaikan hubungan semantik antar kata-kata, sehingga tidak mampu menghasilkan kluster yang sesuai dari dokumen dan juga kadang-kadang tidak mampu membedakan antar 2 kluster yang berbeda. Selain itu pula, proses tradisional tidak mampu menyelesaikan permasalahan *high dimensionality*, yaitu kata-kata dalam jumlah besar yang digunakan dalam konstruksi *feature space.*

Sedangkan proses semantik memiliki kelebihan dalam proses pembagian kluster, karena dalam proses semantik, hubungan antar kata dipertimbangkan. Dokumen yang memiliki kesesuaian dalam semantik akan di kumpulkan dalam 1 *cluster* yang sama, sedangkan jika ditemukan semantik yang berbeda akan dipisahkan ke dalam kluster yang lain. Proses semantik ini berhasil menyelesaikan permasalahan yang ada dalam proses tradisional, yaitu *high dimensionality*, pelabelan kluster dan polisemi.

Penelitian pada bidang *document clustering* telah banyak dilakukan dengan berbagai bidang dengan berbagai macam pendekatan, salah satunya adalah *k-means clustering*. Algoritme ini merupakan algoritme yang paling popular dan telah diterapkan di berbagai bidang seperti *information retrieval,* kesehatan dan manajemen data [13].

Penelitian [12] melakukan pendekatan *Singular Value Decomposition* (SVD) dan *Principal-Component Analysis* (PCA) untuk menyelesaikan sebaran data yang acak dari *document clustering*. Kemudian menerapkan *Support Vector Clustering* (SVC) dan Silhouette untuk mendapatkan banyaknya kluster yang dibentuk agar didapatkan hasil yang optimal. Hasil penelitian ini menunjukan bahwa gabungan SVD dan PCA berhasil diterapkan dalam menyelesaikan permasalahan sebaran data yang acak.

Dalam penelitian [10] melakukan *survey* terhadap pendekatan yang dilakukan dalam *document clustering.* Dalam penelitiannya disebutkan bahwa salah satu pendekatan (algoritme) yang sering digunakan adalah *Hierarchical Agglomerative Clustering* (HAC), algoritme ini sangat mudah diimplementasikan, tidak memerlukan penentuan jumlah *clustering* dan kualitas dari *cluster* yang terbentuk sangat memuaskan.

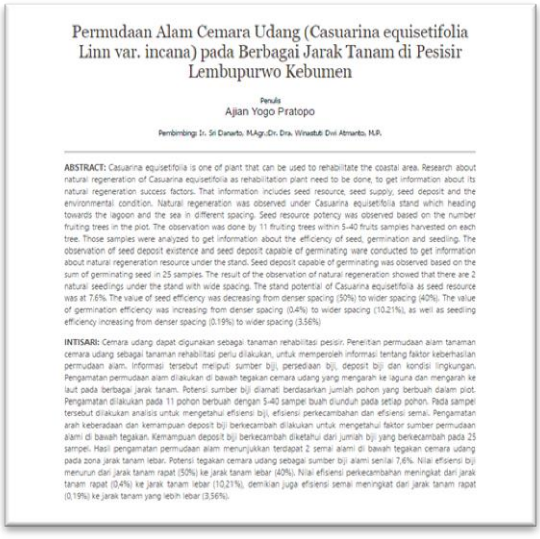
Document clustering secara esensial berasosiasi dengan pemodelan topik [13]. Pemodelan topik yang baik untuk *document clustering* dapat mengurangi *noise* dari perhitungan kemiripan dan dapat mengidentifikasi struktur pengelompokan korpus lebih efektif. Pemodelan topik ini dapat menemukan makna laten yang terdapat di dalam korpus dokumen dan mendapatkan informasi semantik yang berguna untuk mengidentifikasi *document clustering. Latent Dirichlet Allocation* (LDA) [14] adalah metode yang paling sering digunakan saat ini dalam pemodelan topik [13].

Penelitian yang menggabungkan antara *document clustering* dengan pemodelan topik adalah penelitian yang dilakukan oleh [7][8][13]. Dari penelitian tersebut, dapat diketahui bahwa intergrasi antara pemodelan topik dengan *document clustering* dapat membantu menyimpulkan cluster dengan topik yang lebih koheren dan dapat membedakan topik ke kelompok tertentu dan kelompok independen.

Dalam penelitian ini akan mengkombinasikan antara pemodelan topik dengan *document clustering* yang bertujuan untuk mengetahui performa dari penggabungan algoritme pemodelan topik, yaitu LDA dengan *Ward Hierarchical Clustering*.

II. METODOLOGI

Sumber data pada penelitian ini adalah dokumen abstrak penelitian yang ada pada http://etd.repository.ugm.ac.id/. Data tersebut diakuisisi dengan metode *web crawling* dan *web scraping*. Bagian halaman yang diutamakan pada proses *scraping* ini adalah abstrak/intisari penelitian dalam bahasa indonesia dan bahasa inggris. Contoh dari tampilan website etd dapat dilihat pada Gambar 1.



Permudaan Alam Cemara Udang (Casuarina equisetifolia Linn var. incana) pada Berbagai Jarak Tanam di Pesisir Lembupurwo Kebumen

Penulis
Ajian Yogo Pratopo

Pembimbing: Ir. Sri Danarto, M.Agr.,Dr. Dra. Winastuti Dwi Atmanto, M.P.

ABSTRACT: Casuarina equisetifolia is one of plant that can be used to rehabilitate the coastal area. Research about natural regeneration of Casuarina equisetifolia as rehabilitation plant need to be done, to get information about its natural regeneration success factors. That information includes seed resource, seed supply, seed deposit and the environmental condition. Natural regeneration was observed under Casuarina equisetifolia stand which heading towards the lagoon and the sea in different spacing. Seed resource potency was observed based on the number fruiting trees in the plot. The observation was done by 11 fruiting trees within 5-40 fruits samples harvested on each tree. Those samples were analyzed to get information about the efficiency of seed, germination and seedling. The observation of seed deposit existence and seed deposit capable of germinating ware conducted to get information about natural regeneration resource under the stand. Seed deposit capable of germinating was observed based on the sum of germinating seed in 25 samples. The result of the observation of natural regeneration showed that there are 2 natural seedlings under the stand with wide spacing. The stand potential of Casuarina equisetifolia as seed resource was at 7.6%. The value of seed efficiency was decreasing from denser spacing (50%) to wider spacing (40%). The value of germination efficiency was increasing from denser spacing (0.4%) to wider spacing (10.21%), as well as seedling efficiency increasing from denser spacing (0.19%) to wider spacing (3.56%).

INTISARI: Cemara udang dapat digunakan sebagai tanaman rehabilitasi pesisir. Penelitian permudaan alam tanaman cemara udang sebagai tanaman rehabilitasi perlu dilakukan, untuk memperoleh informasi tentang faktor keberhasilan permudaan alam. Informasi tersebut meliputi sumber biji, persediaan biji, deposit biji dan kondisi lingkungan. Pengamatan permudaan alam dilakukan di bawah tegakan cemara udang yang mengarah ke laguna dan mengarah ke laut pada berbagai jarak tanam. Potensi sumber biji diamati berdasarkan jumlah pohon yang berbuah dalam plot. Pengamatan dilakukan pada 11 pohon berbuah dengan 5-40 sampel buah diunduh pada setiap pohon. Pada sampel tersebut dilakukan analisis untuk mengetahui efisiensi biji, efisiensi perkecambahan dan efisiensi semai. Pengamatan arah keberadaan dan kemampuan deposit biji berkecambah dilakukan untuk mengetahui faktor sumber permudaan alami di bawah tegakan. Kemampuan deposit biji berkecambah diketahui dari jumlah biji yang berkecambah pada 25 sampel. Hasil pengamatan permudaan alam menunjukkan terdapat 2 semai alami di bawah tegakan cemara udang pada zona jarak tanam lebar. Potensi tegakan cemara udang sebagai sumber biji alami senilai 7,6%. Nilai efisiensi biji menurun dari jarak tanam rapat (50%) ke jarak tanam lebar (40%). Nilai efisiensi perkecambahan meningkat dari jarak tanam rapat (0,4%) ke jarak tanam lebar (10,21%), demikian juga efisiensi semai meningkat dari jarak tanam rapat (0,19%) ke jarak tanam yang lebih lebar (3,56%).

Gambar 1. Contoh Halaman Web ETD

Dokumen tersebut kemudian ditransformasikan menjadi corpus dengan bentuk *bag-of-words*. Corpus tersebut menjadi masukan untuk pemodelan topik LDA. Tahapan transofrmasi

tersebut meliputi tokenisasi, eliminasi stopword, pembuatan kamus, membentuk matriks dokumen, dan pada akhirnya menjadi corpus dengan bentuk *bag-of-words*. Representasi vektor *bag-of-words* tersebut dapat dilihat pada (1).

$$D_i = (t_1, t_2, \dots, t_n) \quad (1)$$

Dimana dokumen ke *i* (D_i) memiliki distribusi jumah terms (t_n). Vektor bag-of-wods tersebut akan menjadi *corpus* masukan untuk proses LDA. Proses LDA akan menghasilkan model topik yang merupakan distribusi kata sebanyak *k*-topik. Representasi model yang dihasilkan LDA dapat dilihat pada (2).

$$MT = (T_1, T_2, \dots, T_k) \quad (2)$$

Dimana model topik (MT) merupakan distribusi dari topik sebanyak *k*. Nilai *k* merupakan parameter yang dapat ditentukan sendiri oleh pengguna. Representasi dari salah satu topik dapat dilihat pada (3).

$$T_i = (tm_1, tm_2, \dots., tm_n) \quad (3)$$

Dimana topik ke-i (T_i) merupakan distribusi probabilitas dari kata ke (tm_i). Muncul dalam topik tersebut. Model tersebut kemudian diinferensi ke dokumen untuk dihasilkan vektor dokumen topik. Vektor tersebut merupakan distribusi probabilitas topik dalam sebuah dokumen. Representasi vektor tersebut dapat dilihat pada (4).

$$DT_i = (P_{u_d}^{T_0}, P_{u_d}^{T_1}, \dots, P_{u_d}^{T_k}) \quad (4)$$

Vektor dokumen topik (DT_i) berisi probabilitas kemiripan sebuah dokumen tersebut terhadap topik ke-i ($P_{u_d}^{T_k}$). Vektor ini menjadi masukan untuk algoritme *ward hierarchical clustering*. Dalam penelitian ini digunakan *agglomerative clustering* yang artinya pada setiap tahap *clustering* dokumen yang sudah dalam representasi dokument topik (DT_i) akan dibandingkan kemiripannya, dan yang mempunyai kemiripan dengan jarak yang minimum akan digabungkan menjadi sebuah kelompok, sampai seluruh dokumen tergabung dalam satu kelompok. Kriteria kemiripan pada metode Ward ditentukan dari nilai optimal dari fungsi objektif yang dijalankan pada setiap pasangan calon *cluster* yang akan digabungkan [15]. Fungsi objektif ini dapat berupa apapun yang diinginkan dari tujuan pembuatan *cluster*. Sebagai contoh, ward menggunakan fungsi objektif berupa *error sum of squares* yang juga akan digunakan dalam penelitian ini. Fungsi *error sum of squares* dapat dilihat pada persamaan (5) berikut:

$$ESS = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \quad (5)$$

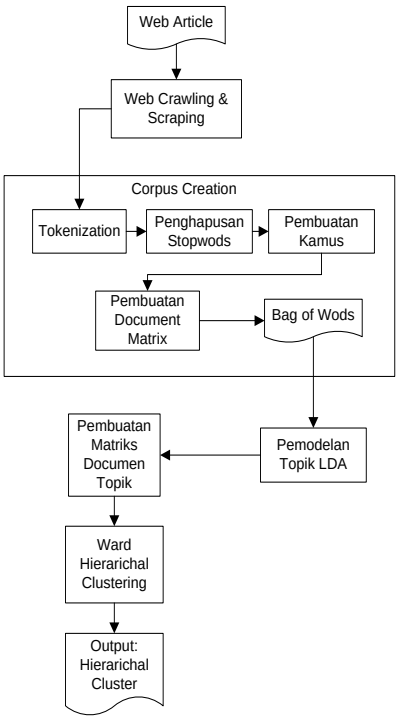
Dalam peneltian ini, *x* merupakan vektor dokumen topik, sebanyak *n* dokumen dalam corpus. Keluaran akhir dari metode ini adalah *cluster* yang berbentuk pohon dendogram. Alur keseluruhan sistem yang akan dikembangkan dalam penelitian ini dapat dilihat pada Gambar 2.

Pengujian dilakukan dengan menggunakan metrics *silhouette analysis*. *Silhouette analysis* dapat digunakan untuk mempelajari jarak yang dihasilkan antar cluster. *Silhouette plot* dapat menampilkan nilai kedekatan setiap titik dalam satu cluster dengan titik yang ada di cluster tetangganya. Nilai yang di miliki antara [-1, 1].

Jika nilai dari *silhouette coefficients* mendekati +1 menujukan bahwa sampel jauh dari cluster tetangga. Jika nilai 0 maka hal tersebut menunjukan bahwa sampel berada pada atau sangat dekat dengan batas antar dua cluster. Jika nilai yang diperoleh negatif maka itu menunjukan bahwa sampel tersebut kemungkinan berada dalam cluster yang salah.

Silhouette coefficient dihitung menggunakan rata-rata dari jarak intra-cluster (a) dan rata-rata dari jarak nearest-cluster (b) untuk setiap sampelnya. *Silhouette coefficient* untuk sebuah sampel dapat dilihat pada (6).

$$Silhouette = \frac{b - a}{\max(a, b)} \quad (6)$$



Gambar 2. Alur Sistem

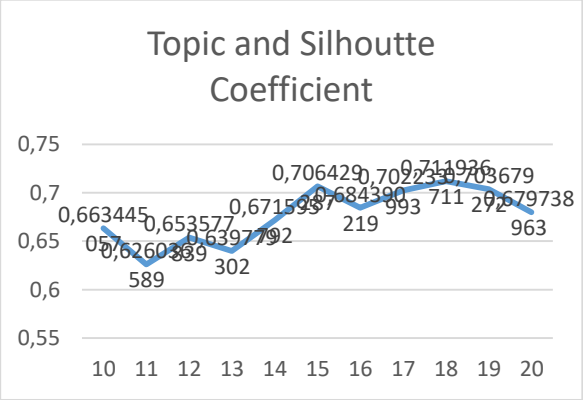
III. HASIL DAN PEMBAHASAN

Hasil dan pembahasan akan dibagi menjadi dua bagian. Bagian pertama membahas mengenai keluaran dari tiap tahap pengujian. Sedangkan bagian kedua membahas mengenai perbandingan dengan metode tf-idf.

A. Hasil penggabungan LDA dengan *Ward Hierarchical Clustering*

Data dalam penelitian tugas akhir yang berasal dari etd UGM (http://etd.repository.ugm.ac.id/). Tujuan utama ekstraksi data adalah untuk mengambil abstrak peneletian, namun metadata seperti, kata kunci, program studi, serta tahun penelitian juga disertakan. Total penelitian yang bisa diambil adalah 91.191 penelitian. Dalam penelitian kali ini, peneliti berasumsi bahwa dokumen-dokumen tersebut telah ter-*cluster* ke dalam program studi masing-masing, sehingga proses *clustering* dokumen akan lebih bermakna jika dilakukan pada masing-masing program studi. Pada pengujian kali ini, peneliti menguji dengan dokumen pada program studi "S2 Ilmu Komputer UGM" dengan total 980 dokumen. Asumsi selanjutnya adalah, dari dokumen-dokumen tersebut, memiliki topik yang berbeda-beda dan setiap dokumennya memiliki lebih dari satu topik.

Pemodelan topik dengan menggunakan *Latent Dirichlet Allocati*on (LDA) dapat menangkap asumsi tersebut. Namun tidak ada cara yang pasti dalam menentukan jumlah topik LDA. Dalam penelitian kali ini, peneliti mencoba untuk memaksimalkan *silhouette coefficient* untuk menentukan jumlah topik. Pengujian dilakukan dengan merubah jumlah topik dari 10 sampai dengan 20. Pada tiap jumlah topik dilakukan clustering dengan metode *hierarchical clustering*. Hasil cluster yang terbentuk kemudian dilihat hasil silhouette *coefficient*nya. Dari pengujian yang telah dilakukan *silhouette coefficient* terbaik didapatkan ketika jumlah topik 18. Hasil selengkapnya ditunjukkan pada Gambar 3.



Gambar 3. Nilai *Silhouette Coeficient* berdasarkan jumlah topik

Jumlah topik 18 kami gunakan sebagai parameter untuk proses pemodelan topik untuk

corpus penelitian pada S2 Ilmu Komputer UGM. Setiap dokumen dalam corpus direpresentasikan dengan vektor *bag-of-words* seperti pada persamaan (1). Hasil dari pemodelan topik adalah probabilitas kata terhadap sebuah topik seperti telah dijelaskan pada persamaan (2) dan (3). Berikut adalah hasil pemodelan topik yang dilakukan:

Topik 1 = 0.014*"keputusan" + 0.010*"informasi" + 0.010*"pendukung" + 0.008*"web" + 0.007*"kriteria" + 0.006*"aplikasi" + 0.006*"service" + 0.005*"dosen" + 0.005*"penentuan" + 0.005*"studi"

Topik 2 = 0.014*"keputusan" + 0.009*"informasi" + 0.009*"pendukung" + 0.005*"lokasi" + 0.005*"kasus" + 0.005*"studi" + 0.004*"jaringan" + 0.004*"klasifikasi" + 0.004*"aplikasi" + 0.004*"web"

Topik 3 = 0.008*"dokumen" + 0.006*"protokol" + 0.006*"kunci" + 0.005*"teks" + 0.005*"informasi" + 0.004*"prediksi" + 0.004*"akurasi" + 0.004*"fuzzy" + 0.004*"klasifikasi" + 0.004*"xml"

Topik 4 = 0.012*"web" + 0.011*"aplikasi" + 0.008*"genetika" + 0.008*"informasi" + 0.007*"jaringan" + 0.007*"kasus" + 0.006*"service" + 0.006*"keputusan" + 0.004*"studi" + 0.004*"teknologi"

Topik 5 = 0.015*"citra" + 0.007*"kasus" + 0.006*"aplikasi" + 0.006*"informasi" + 0.005*"pakar" + 0.005*"perangkat" + 0.004*"implementasi" + 0.004*"objek" + 0.004*"penyakit" + 0.004*"permasalahan"

Topik 6 = 0.009*"kasus" + 0.009*"klasifikasi" + 0.008*"akurasi" + 0.007*"informasi" + 0.006*"jaringan" + 0.005*"dokumen" + 0.004*"pencarian" + 0.004*"citra" + 0.004*"system" + 0.004*"bandwidth"

Topik 7 = 0.028*"informasi" + 0.013*"web" + 0.011*"aplikasi" + 0.006*"teknologi" + 0.006*"kasus" + 0.005*"pengguna" + 0.005*"pencarian" + 0.005*"bahasa" + 0.005*"studi" + 0.004*"komputer"

Topik 8 = 0.014*"citra" + 0.007*"pesan" + 0.006*"jaringan" + 0.006*"steganografi" + 0.005*"bahasa" + 0.005*"pengenalan" + 0.004*"kunci" + 0.004*"karakter" + 0.004*"informasi" + 0.004*"akurasi"

Topik 9 = 0.016*"citra" + 0.008*"bahasa" + 0.005*"dokumen" + 0.004*"identifikasi" + 0.004*"algorithm" + 0.004*"xml" + 0.004*"aturan" + 0.004*"kalimat" + 0.004*"aplikasi" + 0.003*"informasi"

Topik 10 = 0.024*"informasi" + 0.007*"web" + 0.006*"studi" + 0.005*"jaringan" + 0.005*"pengguna" + 0.005*"kasus" + 0.005*"objek" + 0.004*"keputusan" + 0.004*"database" + 0.004*"teknologi"

Topik 11 = 0.017*"citra" + 0.010*"fuzzy" + 0.008*"keputusan" + 0.005*"web" + 0.005*"pendukung" + 0.005*"klasifikasi" + 0.004*"aplikasi" + 0.004*"digital" + 0.003*"ciri" + 0.003*"clustering"

Topik 12 = 0.009*"web" + 0.007*"aplikasi" + 0.006*"jaringan" + 0.006*"teknologi" + 0.006*"informasi" + 0.005*"implementasi" + 0.005*"lokasi" + 0.004*"studi" + 0.004*"akurasi" + 0.004*"daerah"

Topik 13 = 0.015*"keputusan" + 0.010*"citra" + 0.009*"pakar" + 0.008*"pendukung" + 0.007*"kriteria" + 0.007*"informasi" + 0.006*"jaringan" + 0.006*"penyakit" + 0.005*"aplikasi" + 0.005*"web"

Topik 14 = 0.010*"web" + 0.007*"pakar" + 0.007*"service" + 0.007*"aplikasi" + 0.006*"jaringan" + 0.005*"informasi" +

0.005*"akurasi" + 0.005*"keputusan" + 0.004*"aturan" + 0.004*"perhitungan"

Topik 15 = 0.007*"fuzzy" + 0.006*"informasi" + 0.005*"fungsi" + 0.005*"linear" + 0.005*"peramalan" + 0.005*"pendekatan" + 0.004*"pakar" + 0.004*"permasalahan" + 0.004*"teknologi" + 0.004*"programming"

Topik 16 = 0.014*"informasi" + 0.011*"web" + 0.008*"penyakit" + 0.007*"klasifikasi" + 0.006*"aplikasi" + 0.006*"jaringan" + 0.006*"kasus" + 0.005*"service" + 0.005*"pencarian" + 0.005*"kunci"

Topik 17 = 0.009*"informasi" + 0.005*"jaringan" + 0.004*"citra" + 0.004*"studi" + 0.004*"keputusan" + 0.004*"kasus" + 0.004*"penentuan" + 0.004*"program" + 0.004*"server" + 0.004*"web"

Topik 18 = 0.013*"keputusan" + 0.011*"kasus" + 0.008*"pendukung" + 0.008*"jaringan" + 0.006*"aplikasi" + 0.005*"informasi" + 0.004*"kunci" + 0.004*"akurasi" + 0.004*"penyakit" + 0.004*"studi"

Dari model tersebut kita dapat melakukan inferensi ke setiap dokumen yang ada di corpus. Inferensi ini dilakukan untuk mendapatkan distribusi topik dalam setiap dokumen. Berikut adalah contoh representasi vektor dokumen topik seperti pada persamaan (4).

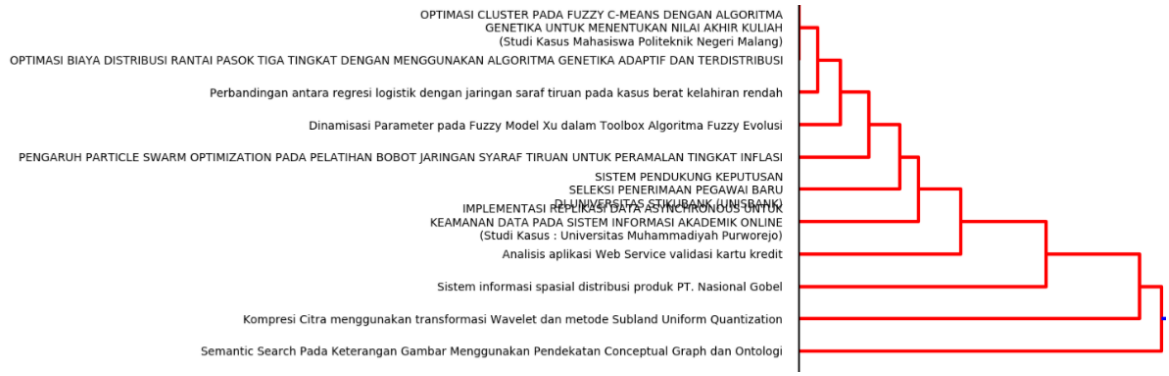
"Sistem Pendukung Keputusan Seleksi Anggota Paduan Suara Dewasa Menggunakan Metode Fuzzy Mamdani (Studi Kasus : Sanggar Bina Vokalia Menteng Palangka Raya)" = [0,0,0,0.123593570057,0,0,0,0,0,0,0,0.0351049853134,0,0,0,0,0.833513281961]

Dari vektor dokumen topik tersebut dapat dilihat bahwa dokumen memiliki probabilitas 0.123 untuk Topik 5, 0.35 untuk Topik ke 13, dan 0.833 untuk Topik 18. Vektor dokumen topik ini akan menggambarkan distribusi topik dalam setiap dokumen. Dengan penggambaran tersebut maka kesamaan antar dokumen dari segi topik dapat dikalkulasikan. Selanjutnya *ward hierarchical clustering* dilakukan untuk menemukan kelompok dokumen yang sama. Gambar 4 menunjukkan potongan dendogram hasil dari *clustering*. Dendogram keseluruhan tidak ditampilkan dalam laporan ini karena bentuknya yang sangat besar berasal dari 980 dokumen.

Dendogram keseluruhan dokumen dapat menjadi dasar untuk menentukan berapa cluster yang kita inginkan dengan mebuat sebuah garis vertikal dimana kita akan memotong dendogram, sehingga didapatkan sejumlah *cluster*. Seperti halnya pada penentuan jumlah topik, tidak ada cara yang baku dalam penentuan berapa *cluster* yang kita inginkan. Dalam penelitian ini kami mengambil 18 *cluster*, dan menghasilkan *silhouette coefficient* 0.712. Hal tersebut menandakan persebaran antar *cluster* sudah cukup baik.

B. Perbandingan Performa dengan Tf-Idf

Hierarchical clustering memiliki kompleksitas yang sangat tinggi dibandingkan dengan *K-means clustering* [16]. Dalam peneletian ini kami berasumsi bahwa pemodelan topik dapat mereduksi dimensi vektor setiap dokumen agar proses *clustering* dapat berjalan lebih ringan. Oleh karena itu kami mencoba membandingkan signifikansi representasi dokumen dengan pemodelan topik dibandingkan dengan metode yang paling sering digunakan yaitu Tf-Idf.



Gambar 4. Dendogram *Document Clustering*

Data dalam percobaan ini sama dengan yang digunakan sebelumnya yaitu dengan menggunakan data penelitian S2 Ilmu Komputer UGM sebanyak 980 dokumen. Dengan menggunakan vektorisasi tf-idf, vektor mempunyai panjang yang bervariasi antar dokumennya. Dari percobaan yang kami lakukan, terdapat satu dokumen yang memiliki panjang hingga 138 elemen. Pada pemodelan topik dengan jumlah topik 18, vektor dokumen mempunyai panjang vektor yang konsisten sesuai jumlah topiknya. Dari beberapa hal yang telah ditunjukkan, dapat dilihat bahwa terdapat signifikansi yang berarti dalam representasi dokumen dengan menggunakan pemodelan topik. Hal ini akan berdampak terhadap waktu komputasi maupun memori yang digunakan untuk proses selanjutnya.

Performa *clustering* dalam kaitannya dengan *silhouette cofficient* dengan menggunakan tf-Idf dan pemodelan topik juga kami ukur. *Silhouette coeffient* yang dihasilkan dengan representasi tf-idf adalah 0.43. Hal ini menunjukkan bahwasannya pemodelan topik dapat merepresentasikan secara lebih *distinctive* antar dokumennya.

IV. KESIMPULAN

Penelitian ini mengkombinasikan teknik *document clustering*, yaitu LDA dengan *Ward Hierarchical Clustering*. Penentuan jumlah topik dalam kaitannya dengan document clustering dapat dilakukan dengan mempertimbangkan *silhouette coefficient* pada hasil *clustering*. Pemodelan topik dapat mereduksi dimensi representasi dokumen untuk mengurangi waktu komputasi serta memori yang digunakan dalam proses *ward hierarchical clustering*. Performa *silhouette coefficient* pada representasi pemodelan topik lebih baik dibandingkan dengan representasi dengan tf-idf.

REFERENSI

[1] A. Spangher, “Building the Next New York Times Recommendation Engine - The New York Times.” New York Times, New York, 2015.

[2] S. S. Desai and J. A. Laxminarayana, “WordNet and Semantic similarity based approach for document clustering,” *2016 Int. Conf. Comput. Syst. Inf. Technol. Sustain. Solut.*, pp. 312–317, 2016.

[3] C. O. Truica, F. Radulescu, and A. Boicea, “Comparing Different Term Weighting Schemas for Topic Modeling,” in *2016 18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, 2016, pp. 307–310.

[4] S. K. Sahu and S. Srivastava, “Review of Web Document Clustering algorithms,” in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 2016, pp. 1153–1155.

[5] X. Sun, “Textual Document Clustering Using Topic Models,” in *2014 10th International Conference on Semantics, Knowledge and Grids*, 2014, pp. 1–4.

[6] D. M. Blei, B. B. Edu, A. Y. Ng, A. S. Edu, M. I. Jordan, and J. B. Edu, “Latent Dirichlet Allocation,” *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

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

[8] J. Millar, G. Peterson, and M. Mendenhall, "Document Clustering and Visualization with Latent Dirichlet Allocation and Self-Organizing Maps.," *FLAIRS Conf.*, pp. 69–74, 2009.

[9] M. Lu, X. J. Zhao, L. Zhang, and F. Z. Li, "Semi-supervised concept factorization for document clustering," *Inf. Sci. (Ny).*, vol. 331, no. 1, pp. 86–98, 2016.

[10] N. Y. Saiyad, H. B. Prajapati, and V. K. Dabhi, "A Survey of Document Clustering," pp. 2555–2562, 2006.

[11] M. T. Hassan, A. Karim, J. B. Kim, and M. Jeon, "CDIM: Document Clustering by Discrimination Information Maximization," *Inf. Sci. (Ny).*, vol. 316, pp. 87–106, 2015.

[12] S. Jun, S. S. Park, and D. S. Jang, "Document clustering method using dimension reduction and support vector clustering to overcome sparseness," *Expert Syst. Appl.*, vol. 41, no. 7, pp. 3204–3212, 2014.

[13] Y. Ma, Y. Wang, and B. Jin, "A three-phase approach to document clustering based on topic significance degree," *Expert Syst. Appl.*, vol. 41, no. 18, pp. 8203–8210, 2014.

[14] D. Blei, L. Carin, and D. Dunson, "Probabilistic topic models," *IEEE Signal Process. Mag.*, vol. 27, no. 6, pp. 55–65, 2010.

[15] J. H. Ward, "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association*, vol. 58, no. 301. pp. 236–244, 1963.

[16] M. Steinbach, G. Karypis, and V. Kumar, "A Comparison of Document Clustering Techniques," *KDD Work. text Min.*, vol. 400, pp. 1–2, 2000.