Terakreditasi SINTA Peringkat 2
Surat Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti No. 10/E/KPT/2019
masa berlaku mulai Vol. 1 No. 1 tahun 2017 s.d Vol. 7 No. 3 tahun 2024

Terbit online pada laman web jurnal: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 5 No. 3 (2021) 421 - 428 ISSN Media Elektronik: 2580-0760

Pemodelan Topik dengan *LDA* untuk Temu Kembali Informasi dalam Rekomendasi Tugas Akhir

Diana Purwitasari[1], Aida Muflichah[2], Novrindah Alvi Hasanah[3], Agus Zainal Arifin[4]
[1, 2, 3, 4] Teknik Informatika, Fakultas Teknologi Elektro dan Informatika Cerdas, ITS Surabaya
[1]diana@if.its.ac.id, [2]aida.muflichah@gmail.com, [3]alvinovrindah@gmail.com, [4]agusza@if.its.ac.id

Abstract

Undergraduate thesis as the final project, or in Indonesian called as Tugas Akhir, for each undergraduate student is a pre-requisite before student graduation and the successfulness in finishing the project becomes as one of learning outcomes among others. Determining the topic of the final project according to the ability of students is an important thing. One strategy to decide the topic is reading some literatures but it takes up more time. There is a need for a recommendation system to help students in determining the topic according to their abilities or subject understanding which is based on their academic transcripts. This study focused on a system for final project topic recommendations based on evaluating competencies in previous academic transcripts of graduated students. Collected data of previous final projects, namely titles and abstracts weighted by term occurences of TF-IDF (term frequency–inverse document frequency) and grouped by using K-Means Clustering. From each cluster result, we prepared candidates for recommended topics using Latent Dirichlet Allocation (LDA) with Gibbs Sampling that focusing on the word distribution of each topic in the cluster. Some evaluations were performed to evaluate the optimal cluster number, topic number and then made more thorough exploration on the recommendation results. Our experiments showed that the proposed system could recommend final project topic ideas based on student competence represented in their academic transcripts.

Keywords: topic extraction, recommendation system, latent dirichlet allocation (LDA)–gibbs sampling, k-means clustering

Abstrak

Tugas akhir merupakan prasayarat kelulusan mahasiswa dalam memenuhi target capaian pembelajaran lulusan perguruan tinggi. Penentuan topik tugas akhir sebaiknya disesuaikan dengan kemampuan mahasiswa. Salah satu strategi mendapatkan topik tugas akhir adalah dengan membaca banyak literature namun hal tersebut akan membutuhkan banyak waktu. Sehingga perlu adanya suatu sistem yang membantu mahasiswa dalam menentukan topik tugas akhir dengan cepat sesuai kemampuan pada transkrip akademik yang merepresentasikan pemahaman mata kuliah. Penelitian ini mengusulkan sistem rekomendasi topik tugas akhir berdasarkan kompetensi dalam transkrip akademik. Data tugas akhir yaitu judul dan abstrak dibobotkan dengan kemunculan kata (TF-IDF, *term frequency–inverse document frequency*) dan dikelompokkan menggunakan K-Means Clustering. Kemudian pada setiap klaster akan diekstraksi topik menggunakan Latent Dirichlet Allocation (LDA) dengan Gibbs Sampling karena perlu memperhitungkan distribusi kata setiap topik. Berbagai skenario uji coba menunjukkan bahwa sistem yang diusulkan dapat merekomendasikan ide topik tugas akhir berdasarkan kompetensi akademik dan dapat membantu mahasiswa mempercepat masa pendidikan sarjananya.

Kata kunci: ekstrasi topik, sistem rekomendasi, latent dirichlet allocation (LDA)–gibbs sampling, algoritma k-means

1. Pendahuluan

Tugas akhir merupakan prasayarat kelulusan untuk memenuhi target capaian pembelajaran program sarjana di perguruan tinggi [1] agar seorang mahasiswa mampu mengaplikasikan pengetahuan, keterampilan, dan ide pada suatu masalah dalam bidang keahlian tertentu secara sistematis dan logis [2]. Penentuan topik tugas akhir yang sesuai menjadi suatu hal yang penting seperti dengan melakukan kaji pustaka dari tugas akhir yang sudah ada. Namun hal tersebut membutuhkan banyak waktu. Kemampuan mahasiswa dalam memahami setiap mata kuliah yang telah ditempuh [3][4] juga menjadi faktor penentuan topik tugas akhir. Sehingga perlu adanya sistem rekomendasi topik tugas akhir yang ditentukan dengan beberapa indikator misalnya data tugas akhir terdahulu (teks judul dan abstrak) serta data transkrip akademik mahasiswa karena nilai transkrip

Diterima Redaksi: 29-03-2021 | Selesai Revisi: 03-05-2021 | Diterbitkan Online: 20-06-2021

berhubungan erat dengan kemampuan mahasiswa dalam memahami mata kuliah. Sistem rekomendasi topik tugas akhir tersebut dapat memberikan saran topik yang lebih mewakili kemampuan seorang mahasiswa.

Beberapa penelitian telah dilakukan untuk rekomendasi topik tugas akhir dengan pengelompokkan K-Means Clustering, misal penelitian yang menunjukkan bahwa nilai mata kuliah wajib berpengaruh dengan penentuan topik tugas akhir mahasiswa [1]. Penelitian lain dengan K-Means untuk rekomendasi topik skripsi menunjukkan bahwa mahasiswa memilih rekomendasi berbeda untuk bidang keahlian sesuai dengan nilai *cluster* paling tinggi [4], namun nilai *k* terlalu sedikit akan memberikan hasil *cluster* yang kurang bagus [5]. Penelitian tersebut juga menunjukkan bahwa ketika memasukkan nilai *k* terlalu besar akan mempengaruhi pengelompokan data judul yang seharusnya berada pada satu *cluster* menjadi terpisah antar *cluster*. Penelitian lain memberikan rekomendasi tanpa pengelompokkan dan menerapkan temu kembali informasi [6]. Berbeda dengan penelitian sebelumnya, pada penelitian ini terjadi pengelompokkan dan temu kembali tugas akhir sebagai rekomendasi dengan mempertimbangkan mata kuliah pada suatu kelompok bidang keahlian mahasiswa. Kombinasi beberapa pendekatan tersebut dilakukan untuk menjawab permasalahan dengan menggabungkan kelebihan dari penelitian terdahulu. Penelitian ini juga menggunakan K-Means untuk pengelompokkan teks tugas akhir (judul+abstrak) yang umumnya memakai *cosine similarity* seperti penelitian lain serupa namun bertujuan rekomendasi dosen pembimbing tugas akhir [7] atau proses menghitung kemiripan hasil klasterisasi dokumen [8].

Pada sistem rekomendasi yang diusulkan dilakukan ekstraksi topik dari hasil pengelompokkan dengan Latent Dirichlet Allocation (LDA)-Gibbs Sampling untuk mempercepat proses pencarian kandidat topik. Penelitian terdahulu dengan LDA-Gibbs Sampling untuk ekstraksi topik sebagai fitur dari teks bahasa Indonesia [2], segmentasi teks media sosial dengan tambahan informasi waktu [9], analisis topik yang tren pada situs *online* untuk mengetahui keluhan masyarakat [10]. Kemudian dengan hasil topik yang sudah dikenali maka dihitung *cosine similarity* berdasarkan topik tersebut seperti penelitian sebelumnya [11]. Usulan sistem akan dievaluasi dengan menghitung akurasi dari kecocokan antara data tugas akhir sesungguhnya dengan topik yang direkomendasikan oleh sistem berdasarkan input transkrip akademik dari pengguna atau mahasiswa.

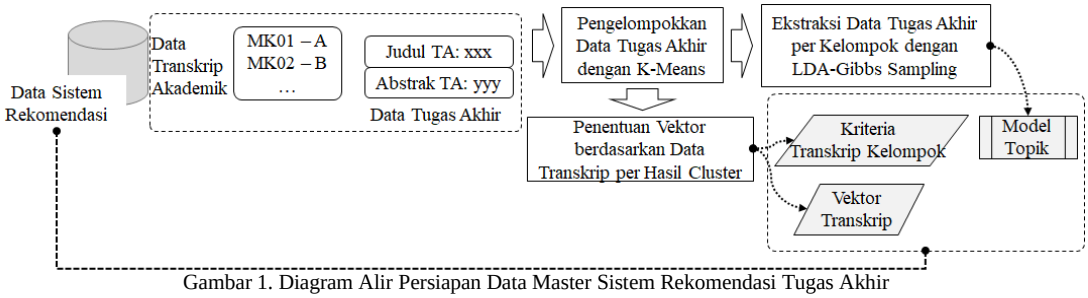
2. Metode Penelitian

Sistem rekomendasi topik tugas akhir (TA) dibangun berdasarkan kemiripan transkrip akademik masukan dari pengguna yaitu mahasiswa yang sedang mencari ide pengerjaan topik. Transkrip mahasiswa tersebut akan dibandingkan dengan transkrip akademik mahasiswa terdahulu. Dikarenakan banyaknya kemungkinan data mahasiswa dengan transkrip akademik serupa maka teknik temu kembali dilakukan untuk memberikan kandidat judul tugas akhir yang sesuai agar dapat menginspirasi pengguna.

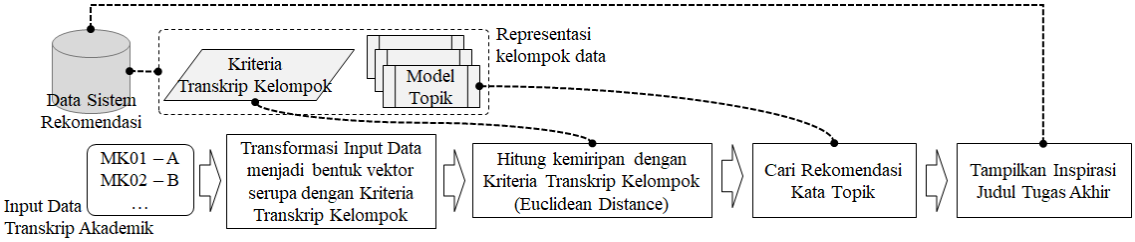
Keseluruhan proses sistem rekomendasi topik tugas akhir ditunjukkan pada Gambar 1 untuk persiapan data master dan Gambar 2 untuk proses rekomendasi yang memanfaatkan data master.

2.1. Persiapan Data

Dataset yang digunakan termasuk data nilai-nilai mata kuliah (MK) dalam transkrip akademik mahasiswa serta data tugas akhir yaitu judul dan abstrak. Terdapat banyak



Gambar 1. Diagram Alir Persiapan Data Master Sistem Rekomendasi Tugas Akhir



Gambar 2. Diagram Alir Proses Pemberian Rekomendasi Tugas Akhir

DOI: https://doi.org/10.29207/resti.v5i3.3049

kemungkinan data transkrip yang mirip dan tersedia dalam dataset dengan data pengguna, sehingga untuk mengurangi kandidat judul tugas akhir maka dilakukan pengklasteran/ pengelompokkan. Tahap pertama yaitu pengelompokan teks judul dan abstrak dengan algoritma K-Means yang diawali praproses menggunakan *library* SASTRAWI (pypi.org/project/Sastrawi) terdiri dari tokenisasi atau pemisahan teks menjadi kata token, *stemming* (ubah kata berimbuhan ke bentuk dasar), dan *stopword removal* (penghapusan kata-kata kurang/tidak bermakna). Kata atau token tersebut masuk dalam daftar kata (indeks) yang digunakan sebagai acuan saat mempersiapkan suatu data judul-abstrak menjadi vektor dengan memperhitungkan bobot kemunculan kata TF-IDF (*term frequency–inverse document frequency*). Jika suatu kata sering muncul maka memiliki nilai bobot TF-IDF yang lebih besar. Ukuran representasi vektor tersebut yang berupa vektor kolom dan bukan vektor baris sama dengan jumlah kata indeks, sedemikian hingga pengurangan kata melalui proses *stemming* dan *stopword removal* akan mengurangi dimensi vektor.

2.2. Ekstrasksi Topik pada Kelompok Judul Tugas Akhir

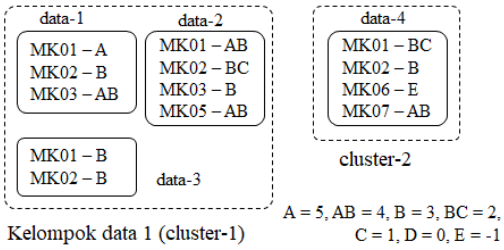
Representasi vektor judul-abstrak dengan pembobotan TF-IDF dikelompokkan menggunakan K-Means. Hasil proses tersebut adalah kelompok atau *cluster* dengan judul-judul yang serupa karena memiliki kata-kata yang sering muncul. Penentuan inisial *centroid* dan jumlah kelompok yang diharapkan dapat mempengaruhi hasil rekomendasi, sehingga uji coba dengan variasi skenario telah dilakukan.

Pada setiap kelompok judul tugas akhir dilakukan ekstraksi topik dengan LDA-Gibbs Sampling. Inisiasi matriks topik-kata secara random dilakukan sebagai representasi pemetaan topik dan kata yang memiliki kemungkinan berasosiasi ke topik tersebut. Iterasi perhitungan probabilitas topik untuk setiap kata pada teks judul-abstrak dilakukan berdasarkan parameter dari Gibbs Sampling antara lain nilai yang menunjukkan distribusi topik pada satu data (teks judul-abstrak) serta nilai distribusi kata dalam satu topik. Setelah semua iterasi selesai maka dilakukan perhitungan matriks indikasi kepentingan kata untuk tiap topik dan matriks indikasi kepentingan topik untuk tiap data. Kesemua proses tersebut diimplementasikan dengan bantuan *library* GENSIM (radimrehurek.com/gensim).

Hasil proses ekstraksi topik adalah daftar topik sedemikian hingga satu topik dapat dianalogikan sebagai suatu kelompok kata. Oleh karena itu proses ekstrasi topik dengan pemodelan topik LDA atau tepatnya LDA-Gibbs Sampling adalah setara dengan pendekatan *unsupervised* (pengelompokkan/ pengklasteran).

2.3. Rekomendasi Topik Tugas Akhir

Satu kelompok berisi data tugas akhir (teks judul-abstrak) dapat memiliki banyak topik hasil ekstraksi.



Gambar 3. Contoh Representasi Kriteria Transkrip Kelompok

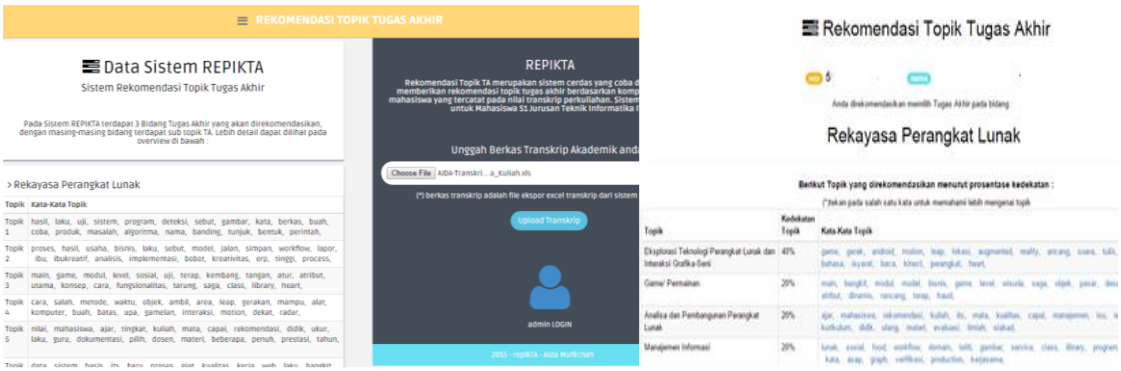
Gambar 4. Contoh Data Tugas Akhir untuk Persiapan Data Master

Satu judul tugas akhir akan berasosiasi dengan nilai-nilai mata kuliah pada suatu transkrip mahasiswa, sehingga pada satu kelompok dengan banyak transkrip akan ditentukan kriteria transkrip kelompok seperti yang ditunjukkan pada Gambar 3. Pada contoh tersebut terdapat dua kelompok yaitu cluster-1 dan cluster-2. Vektor representasi untuk Kriteria Transkrip Kelompok akan menjadi vektor kolom dengan tujuh dimensi karena terdapat mata kuliah (MK) dari kode MK01 … MK07. Contoh data yang menjadi input proses persiapan data master dari sistem rekomendasi (Gambar 1) ditunjukkan pada Gambar 4. Penentuan nilai per dimensi ditentukan dengan melihat kemunculan nilai terbanyak di suatu MK (nilai modus). Misal nilai MK02 di cluster-1 pada Kriteria Transkrip Kelompok akan diset 3 karena terdapat dua data yang memiliki nilai B. Pada kondisi terburuk jika tidak ada nilai yang dominan, maka diambil nilai tengah. Misal nilai MK01 di cluster-1 akan diset 4. Sebagai catatan, di satu kelompok hasil pengklasteran akan memiliki satu vektor Kriteria Transkrip Kelompok.

Urutan proses rekomendasi dilakukan sebagai berikut sesuai dengan Gambar 2:

a. Pengguna memasukkan data transkrip akademik berbentuk CSV (comma-separated values) terdiri dua kolom yaitu kolom kode MK dan kolom nilai dalam huruf. Pada sistem rekomendasi yang

DOI: https://doi.org/10.29207/resti.v5i3.3049



Gambar 5. Contoh Usulan Sistem Rekomendasi Topik Tugas Akhir (kiri: laman input, kanan: tampilan hasil)

diusulkan sudah tersimpan data katalog kode MK dan nama MK sebagai pemetaan.

b. Transformasi input data transkrip akademik ke bentuk vektor berdimensi sama dengan Kriteria Transkrip Kelompok.
c. Menggunakan *Euclidean Distance* untuk hitung kemiripan vektor input dengan semua vektor Kriteria Transkrip Kelompok dari representasi hasil proses K-Means, pilih kelompok dengan jarak terdekat.
d. Pada kelompok terpilih, ambil sejumlah data tugas akhir berdasarkan jarak terdekat antara setiap vektor transkrip akademik dengan vektor Kriteria Transkrip Kelompok (memakai *Euclidean Distance*).
e. Di setiap data tugas akhir tersebut, ambil topik dengan nilai probabilitas tertinggi berdasarkan kemungkinan teks judul dan atau abstrak dapat dipetakan ke model topik hasil ekstraksi.
f. Tampilkan hasil rekomendasi ke pengguna. Sistem dapat menampilkan beberapa informasi sebagai berikut: (i) judul tugas akhir dari mahasiswa terdahulu yang memiliki nilai-nilai mata kuliah mirip dengan pengguna; (ii) kata-kata dalam topik yang memiliki kemungkinan terbesar berasosiasi dengan topik tersebut; (iii) daftar judul tugas akhir terdahulu yang memiliki kata-kata di poin (ii).

3. Hasil dan Pembahasan

3.1 Definisi Data Sistem Rekomendasi Tugas Akhir

Uji coba dilakukan di mahasiswa program studi sarjana dengan jumlah mahasiswa per tahun angkatan akademik 150-200 orang. Program studi tersebut mengadakan wisuda satu tahun sebanyak dua kali. Data penelitian ini diambil dari mahasiswa yang lulus dalam tiga periode wisuda sejumlah 240 data latih dan 80 data uji.

Program studi tersebut memiliki tiga bidang ilmu dasar yaitu Rekayasa Perangkat Lunak (RPL), Komputer Cerdas dan Visi (KCV), dan Komputasi Berbasis Jaringan (KBJ) meski terdapat lebih dari tiga laboratorium yang biasanya digunakan oleh mahasiswa menyelesaikan pengerjaan tugas akhirnya. Pada saat uji coba dengan data latih, nilai K dari algoritma K-Means diset beberapa variasi. Namun analisis menunjukkan bahwa tiga kelompok memberikan hasil yang sesuai.

Antar muka sistem rekomendasi pada penelitian ini diimplementasikan dengan bahasa pemrograman PHP *framework* Laravel (Gambar 5). Implementasi dilakukan pada sistem operasi Windows dengan basis data MySQL serta uji coba bersama mahasiswa menggunakan *browser* Google Chrome. Selain itu perlu dilakukan setting konfigurasi MySQL serta web server untuk memastikan keberhasilan operasi pengolahan teks panjang, misal: *max_allowed_packet, post_max_size memory_limit, wait_timeout,* dan lain-lain.

3.2. Uji Coba K-Means Clustering

Hasil pengujian pada Tabel 1 menunjukkan variasi jumlah kelompok, sedangkan Tabel 2 menunjukkan variasi data teks yang digunakan pada suatu skenario pengelompokkan dengan nilai *k*=3. Kemudian hasil pada Tabel 3 menunjukkan pelabelan manual yang diberikan dan sudah dicocokkan dengan data label sesungguhnya dari setiap data tugas akhir.

Hasil pengujian pada Tabel 1 berfokus pada pemilihan *centroid* serta penentuan nilai k *cluster* yang dibentuk. Untuk pendekatan manual dilakukan pemilihan data tugas akhir secara random dengan menanyakan kepada sejumlah mahasiswa yang memilih berdasarkan tingkat ketertarikan mereka terhadap suatu subyek bahasan. Evaluasi dilakukan berdasarkan nilai *Cluster Variance* dari hasil pengelompokan untuk menunjukkan *cluster* yang ideal. Kriteria kelompok tersebut yaitu nilai *Vw* (*Variance Within*) minimum, karena merepresentasikan *internal homogeneity* (kerekatan data dalam tiap *cluster*) dan nilai *Vb* (*Variance Between*) maksimum karena menyatakan *external homogeneity* (jarak antar *cluster*). Hasil suatu skenario pengelompokkan yang baik akan memiliki nilai *Cluster Variance* (*Vw/Vb*) yang kecil.

Pada Tabel 1 terlihat bahwa set nilai *k* semakin besar memberikan hasil kelompok yang cenderung lebih ideal dengan tingginya nilai *Cluster Variance*. Akan tetapi waktu yang diperlukan untuk eksekusi akan lebih lama. Hasil skenario-1 sampai skenario-5 menunjukkan bahwa

DOI: https://doi.org/10.29207/resti.v5i3.3049

Tabel 1. Hasil Percobaan Penentuan *Centroid* dan Nilai *k* Terbaik

No	Centroid	*k-cluster*	Vw	Vb	*Cluster Variance* (Vw/Vb)	Keterangan
1.	Mean	3	0,054	0,093	0,578	Disebut skenario-1
2.	Mean	6	0,054	0,093	0,574	30% lebih lama dibanding
3.	Mean	8	0,053	0,090	0,593	waktu skenario-1
4.	Mean	10	0,053	0,081	0,657	2x lebih lama dibanding
5.	Mean	16	0,053	0,079	0,670	waktu skenario-1
6.	Manual	3	0,054	0,127	0,424	Cenderung sama dengan
7.	Manual	6	0,053	0,108	0,494	waktu skenario-1
8.	Manual	8	0,053	0,107	0,493	

Tabel 2. Hasil Percobaan Penentuan Konten Teks Dokumen (*centroid manual*, *k*=3)

No	Teks	Vw	Vb	*Cluster Variance* (Vw/Vb)	Keterangan
1.	Judul + Abstrak	0,054	0,127	0,424	Disebut skenario-1
2.	Judul	0,313	0,742	0,433	80% lebih cepat dibanding waktu skenario-1
3.	Abstrak	0,053	0,123	0,430	30% lebih cepat dibanding waktu skenario-1

Tabel 3. Data Hasil Ekstraksi Topik Masing-Masing Kelompok Bidang Ilmu Tugas Akhir

Klaster	Bidang Ilmu	Jumlah Data TA	Hasil Ekstraksi Topik (label diberikan manual)	Kata-Kata yang Mewakili
1	Rekayasa Perangkat Lunak (RPL)	104	1. Pembangunan Aplikasi Penyelesaian Permasalahan 2. Manajemen Informasi 3. Analisa dan Pembangunan Perangkat Lunak 4. Eksplorasi Teknologi Perangkat Lunak dan Interaksi Grafika-Seni 5. Game/Permainan	1. ibukreatif, informatika, usaha, workflow, apps, ziarah, modul facebook, obat 2. sosial, food, workflow, domain, merchant 3. mahasiswa, kualitas, manajemen, iso, iec, prestasi, prasyarat, kurikulum, didik, evaluasi 4. game, gerak, android, motion, leap, lokasi, augmented, reality, suara, tulis, gamelan 5. bangkit, model, bisnis, game, level, wisuda, saga, pasar, desain, card, atribut, frau
2	Komputer Cerdas dan Visi (KCV)	69	1. Pengolahan Citra 2. Data Mining 3. Optimasi Problem dan Analisa Data	1. warna, segmentasi, daun, gigi, smartphone, fuzzy, dipstick, kamera, urinalysis, tani 2. klasifikasi, k-means, fuzzy, sel, cluster, neural, darah, modifikasi, clustering, leukemia 3. batik, kain, optimasi, impresi, cortical, dokter, radiograph, bone, optimal, motif, metode
3	Komputasi Berbasis Jaringan (KBJ)	67	1. Sistem Monitoring 2. Teknologi Terapan 3. Jaringan Multimedia 4. Keamanan Sistem	1. kendaraan, android, server, lokasi, workover, adaptif, sql, enkripsi, qr, citra, parallel 2. sensor, air, suhu, mikrokontroler, arduino, robot, adaptif, kualitas, mobile, awan, wi-fi 3. pantau, kereta, streaming, api, jaringan, cctv, sensor, rate, terap, mesh, getar, picture 4. serang, aman, gempa, steganografi, kompresi, honeypot, injection, tweet, radar, email, retas

tiga kelompok hasil memberikan kondisi yang cukup ideal baik dari segi waktu eksekusi maupun tingkat kerekatan data. Meskipun uji coba serupa dilakukan dengan skenario-6 sampai skenario-8, terlihat nilai *k*=3 adalah kondisi yang sebaiknya dipilih. Hal tersebut tidak bertentangan dengan bidang ilmu dari tugas akhir seperti yang dijelaskan pada sub bahasan 3.1 merepresentasikan kelompok RPL, KCV, dan KBJ. Penentuan *centroid* dengan pendekatan *mean* (secara random) dilakukan dengan menghitung nilai maksimum dan minimum TF-IDF tiap kata pada vektor dokumen. Kemudian dari *range* angka nilai maksimum dan minimum tersebut dibagi ke dalam *k* sub-range nilai TF-IDF. *Centroid* setiap *cluster* diambil dari nilai tengah tiap sub-*range* secara beruntun (Gambar 6).

Kemudian hasil Tabel 2 dilakukan dengan eksekusi K-Means untuk nilai *k*=3 dan pemilihan *centroid* secara manual yang memanfaatkan variasi data teks. Indikator keberhasilan kondisi ideal kelompok memberikan nilai yang serupa. Hal yang membedakan adalah waktu eksekusi, sehingga semakin panjang teks yaitu judul dan

Gambar 6. Ilustrasi penentuan nilai *centroid* dengan *k*=3

Tabel 4. Hasil Uji Coba Ekstraksi Topik

No	RPL	KCV	KBJ	Rata-rata kemiripan
1	9	4	5	0,36
2	5	4	4	0,39
3	5	5	5	0,40
4	6	3	4	0,41
5	5	3	4	0,44

DOI: https://doi.org/10.29207/resti.v5i3.3049

Tabel 5. Hasil Percobaan Penentuan Bobot Judul+Abstrak pada Konten Teks Dokumen (centroid manual, k=3)

No	Judul + Abstrak	*Vw*	*Vb*	*Cluster Variance*	RPL	KCV	KBJ	Rata-rata
1.	10% + 90%	0,058	0,143	0,405	87 %	77 %	82 %	82 %
2.	20% + 80%	0,067	0,173	0,387	89 %	79 %	82 %	83 %
3.	30% + 70%	0,081	0,210	0,387	89 %	77 %	83 %	83 %
4.	40% + 60%	0,100	0,251	0,399	87 %	76 %	78 %	80 %
5.	50% + 50%	0,124	0,309	0,401	86 %	76 %	77 %	80 %
6.	60% + 40%	0,154	0,373	0,408	86 %	76 %	74 %	79 %
7.	70% + 30%	0,184	0,441	0,417	84 %	76 %	74 %	78 %
8.	75% + 25%	0,204	0,482	0,424	84 %	76 %	71 %	77 %
9.	80% + 20%	0,224	0,524	0,427	84 %	76 %	69 %	76 %
10.	90% + 10%	0,267	0,625	0,427	86 %	76 %	68 %	77 %

abstrak maka semakin lama proses pengelompokkan akan berakhir. Berdasarkan indikator hasil di Tabel 2 maka disarankan kebutuhan data tugas akhir berbentuk teks untuk persiapan data master sistem rekomendasi (Gambar 1) cukup menggunakan judul yang umumnya memiliki panjang 15-20 kata. Sebagai catatan, hasil pada Tabel 1 dan Tabel 2 adalah nilai rata-rata dari lima kali eksekusi suatu skenario.

3.3 Uji Coba LDA-Gibbs Sampling

Analisis manual pada hasil *k*=3 dari Tabel 2 dilakukan dan memberikan label kelompok menurut sebagian besar label teks tugas akhir terlihat pada Tabel 3. Label topik pada kolom-3 di Tabel 3 juga diberikan secara manual dengan mengamati kata-kata yang memiliki nilai probabilitas tinggi dalam suatu topik. Namun jumlah topik yang diekstraksi ditentukan dari hasil uji coba pada Tabel 4. Jumlah data tugas akhir (TA, yaitu teks judul) pada kolom-2 di Tabel 3 menunjukkan suatu kewajaran bahwa jumlah topik di kelompok berlabel RPL akan lebih banyak. Ekstraksi topik termasuk pendekatan *unsupervised* (misal pengklasteran) dan asumsi yang sering digunakan adalah setidaknya satu kelompok beranggotakan ± 10% data. Oleh karena itu pada skenario-1 di Tabel 4 diset terdapat 9 topik yang akan diekstraksi dari kelompok data TA berlabel RPL. Kemudian berbagai kombinasi jumlah topik di masing-masing kelompok teks TA dicoba untuk proses evaluasi yang menggunakan rata-rata kemiripan antar teks judul menggunakan rumus jarak *cosine similarity*. Salah satu kombinasi adalah memberikan jumlah topik yang sama pada tiap kelompok (skenario-3, Tabel 4) atau menambahkan jumlah topik secara bertahap. Kombinasi jumlah topik dengan rata-rata kemiripan tertinggi antar teks judul dalam suatu topik ada pada skenario-5. Kemudian analisis data dari skenario-5 dilakukan untuk pelabelan manual dengan hasil ditunjukkan pada kolom-3 di Tabel 3.

Untuk memperjelas penentuan topik hasil klaster atau bidang ilmu RPL, KCV dan KBJ maka dilakukan evaluasi terkait bobot kata pada judul dan abstrak (Tabel 5) sebagai observasi lanjutan dari Tabel 3 menggunakan *centroid* manual dan nilai *k*=3. Ujicoba pembobotan dengan variasi teks judul dan abstrak menunjukkan hasil pengelompokkan yang maksimal jika dilihat dari nilai *Cluster Variance*. Tabel 5 kolom 1 berisi bobot pada vektor TF-IDF teks judul dan abstrak dengan konstanta perbandingan yang jika dijumlahkan sama dengan nilai 1. Evaluasi variasi pembobotan di Tabel 5 menunjukkan hasil pengelompokan optimal dari indikator *cluster variance* terjadi di skenario-9 dan skenario-10. Akan tetapi hasil klaster divalidasi dengan label pada setiap data tugas akhir. Label tersebut adalah RPL, KCV dan KBJ. Sebagai contoh pada skenario-9 dengan hasil klaster dianggap baik (*cluster variance* = 0.427) namun validasi data pada label KCV dan KBJ memiliki akurasi yang jauh berbeda dibanding RPL. Sehingga dengan mempertimbangkan indikator *cluster variance* serta akurasi label bidang keilmuan (RPL, KCV, KBJ) maka sistem rekomendasi topik tugas akhir akan diuji dengan pengelompokan dokumen optimal *centroid* manual, *k*=3, dan pembobotan judul : abstrak adalah 20% : 80%.

3.4 Hasil Rekomendasi Topik Tugas Akhir

Uji coba pada 80 data tugas akhir dilakukan dengan mencocokkan bidang ilmu pilihan sistem rekomendasi topik dengan bidang ilmu sesungguhnya. Untuk itu juga disiapkan 80 data transkrip yang akan menjadi input sistem rekomendasi. Tabel 6 menunjukkan contoh hasil pengujian dengan kolom d(0), d(1), d(2) berisi nilai *cosine similarity distance* dengan tiap kelompok bidang keilmuan. Pada data uji 1, jarak terdekat ada di kolom d(1) sehingga mahasiswa tersebut direkomendasikan untuk mengambil topik di bidang RPL (kolom Rekom) dan memang judul tugas akhir sesungguhnya ada di RPL (kolom Fakta).

Uji coba juga dilakukan dengan mempertimbangkan nilai mata kuliah (MK) berdasarkan kategori MK-wajib atau MK-pilihan. Kurikulum dari data mahasiswa yang digunakan telah menetapkan bahwa MK-pilihan mulai bisa diambil dari Semester-5. Uji coba dari indikator akurasi label rekom dan fakta menunjukkan bahwa rata-rata kemiripan atau similaritas rekomendasi topik lebih besar jika hanya memperhitungkan MK-pilihan saja tanpa mengikutsertakan MK-wajib. Nilai kemiripan tersebut dihitung dari jarak antara vektor berisi 20 kata bobot tertinggi pada suatu data uji dengan vektor berisi 20 kata bobot tertinggi pada kelompok topik yang direkomendasikan. Nilai kemiripan tersebut ditunjukkan pada kolom Similaritas Kata Inti di Tabel 6. Data uji di Tabel 6 dituliskan secara urut berdasarkan kolom tersebut dan menghasilkan rata-rata kemiripan 0,44.

DOI: https://doi.org/10.29207/resti.v5i3.3049

Tabel 6. Hasil Percobaan Penentuan Centroid dan Nilai *k* Terbaik

No	d(0) KCV	d(1) KBJ	d(2) RPL	Rekom	Fakta	Judul Tugas Akhir Sesungguhnya dari Mahasiswa	Similaritas Kata Inti
1.	10,12	13,60	11,00	KBJ	KBJ	Sistem Pendeteksi Serangan Adaptif dengan Menggunakan Algoritma Genetik	0,93
2.	7,64	12,25	10,42	KBJ	KBJ	IRITS: Rancang Bangun Sistem Irigasi Stadion Sepak Bola ITS dengan Mikrokontroler Arduino	0,78
3.	8,27	11,80	11,02	RPL	RPL	Rancang Bangun Modul Pengenalan Suara Menggunakan Teknologi Kinect	0,76
4.	11,66	6,44	12,63	RPL	RPL	Rancang Bangun Layanan Pemesanan Fasilitas Umum di ITS, Studi Kasus Asrama	0,76
…	…	…	…	…	…		
8.	7,65	10,32	12,36	RPL	RPL	Implementasi Picture Streaming pada Jaringan Mesh Berbasis Fisheye State Routing menggunakan Raspberry Pi untuk Pemantauan Jalan Raya	0,71
9.	8,13	11,42	12,25	RPL	RPL	Rancang Bangun Aplikasi `Icare` Media Pembelajaran Siswa Autis di Sekolah Dasar	0,71
…	…	…	…	…	…		
78.	12,51	8,94	12,37	KCV	KCV	Rancang Bangun Aplikasi Buku Dongeng - Kumpulan Cerita Rakyat Interaktif Berbasis iOS	0,13
79.	12,33	12,78	7,47	KBJ	KBJ	Rancang Bangun Web Service untuk Implementasi Aturan Main dan Manajemen Transaksi dalam Game Sosial Food Merchant Saga pada Perangkat Android	0,13
80.	8,52	11,61	12,95	RPL	RPL	Game Edukasi Simulasi Haji Menggunakan RenPy pada Perangkat Android untuk Simulasi Perjalanan Ibadah Haji	0,12
						Rata-rata similaritas	0,44

Tabel 7. Contoh Rekomendasi Dokumen Uji Teratas dari 80 data uji

No	Tugas Akhir	Rekomendasi Bidang	Rekomendasi Topik	Similaritas Kata Inti	Nilai Survei
1	Sistem Pendeteksi Serangan Adaptif dengan Menggunakan Algoritma Genetik	KBJ	Keamanan Sistem/ Teknologi Terapan	0,93	0,94
2	IRITS: Rancang Bangun Sistem Irigasi Stadion Sepak Bola ITS dengan Mikrokontroler Arduino	KBJ	Teknologi Terapan	0,78	0,97
3	Rancang Bangun Modul Pengenalan Suara Menggunakan Teknologi Kinect	RPL	Eksplorasi Teknologi Perangkat Lunak dan Interaksi Grafik-Seni	0,76	0,95
4	Rancang Bangun Layanan Pemesanan Fasilitas Umum di ITS, Studi Kasus Asrama	RPL	Pembangunan Aplikasi Penyelesaian Permasalahan	0,76	0,76
5	Sistem Pendeteksi Dan Pencegah Peretasan Terhadap Aplikasi Berbasis Web Dengan Teknik Web Application Firewall (WAF)	KBJ	Keamanan Sistem	0,74	0,97

Gambar 7. Contoh Laman Web Penilaian Survei Hasil Rekomendasi Data Uji

Meskipun nilai kemiripan kurang dari 0,5, namun kolom Rekom dan Fakta menunjukkan akurasi kecocokan 100%.

Meski tidak terlihat sepenuhnya, namun Tabel 6 memperlihatkan 10 dokumen teratas dengan tingkat ketepatan rekomendasi paling tinggi dan 3 dokumen terbawah. Hasil perhitungan 80 data uji menunjukkan

DOI: https://doi.org/10.29207/resti.v5i3.3049

lebih dari 29 data uji menghasilkan similaritas diatas 0,50 dari rentang nilai 0 hingga 1.

Sistem rekomendasi mengeluarkan daftar topik dari 5 dokumen terdepan, namun vektor topik dibentuk dari topik dengan kemunculan terbanyak atas 5 dokumen terdekat. Uji coba jumlah rekomendasi yang diberikan telah dilakukan dengan membandingkan hasil kemiripan atau Similaritas Kata Inti untuk 1, 5, dan 10. Namun nilai kemiripan tertinggi yaitu 0,44 dihasilkan jika diset 5 rekomendasi. Sehingga Tabel 7 menunjukkan sebagian atau 5 data uji dari Tabel 6 dengan rekomendasi bidang keilmuan dan rekomendasi topik yang diberikan beserta nilai Similaritas Kata Inti dan nilai kepuasan dalam survei. Contoh survei dilakukan dengan laman web pada Gambar 7. Pelaksanaan survei dilakukan dengan 10 data uji yang memiliki similaritas tertinggi antara nilai kesepakatan pengguna mahasiswa. Responden diminta memberikan status yang dikonversi ke nilai yaitu sangat mendekati (1,0), mendekati (0,7), tidak mendekati (0,4) dan sangat tidak mendekati (0,1). Hasil survey pada Tabel 7 untuk semua data responden menunjukkan nilai kesepakatan diatas 0,7 yang diperkuat dengan rata-rata kesepakatan kesesuaian rekomendasi dan label bidang sebesar 92%.

Berbagai skenario evaluasi sistem rekomendasi telah dilakukan untuk menunjukkan performa sistem yang diusulkan. Akan tetapi rekomendasi bidang keilmuan yang didapatkan dari pengelompokkan K-Means dan topik dari *LDA-Gibbs Sampling* juga sebaiknya memperhatikan batas nilai similaritas pada temu kembali rekomendasi. Sehingga akan ada kemungkinan tingkat kepercayaan hasil rekomendasi yang rendah jika nilai similaritas tersebut kurang dari nilai ambang yang ditetapkan. Usulan sistem rekomendasi topik ini dapat menjadi pelengkap pada sistem tugas akhir [12].

4. Kesimpulan

Sistem rekomendasi topik tugas akhir menunjukkan bahwa ekstraksi topik metode *LDA-Gibbs Sampling* dengan pemilihan kata di teks abstraksi dan judul dapat menghasilkan kata inti topik tugas akhir yang lebih sesuai dengan bidang keilmuan tugas akhir. Semakin tinggi nilai similaritas vektor kata inti dari sistem akan memberikan nilai probabilitas yang tinggi dalam topik terpilih. Nilai similaritas vektor kata inti yang rendah dikarenakan tidak banyak data yang bertopik sama dengan data uji. Sehingga semakin banyak data latih untuk persiapan data master pada sistem rekomendasi topik tugas akhir akan menghasilkan rekomendasi yang lebih bervariasi.

Ucapan Terima Kasih

Penelitian pada makalah ini merupakan bagian dari penelitian tentang pengembangan model pada sistem pendukung keputusan untuk rekomendasi pakar peneliti sesuai Kontrak Penelitian antara Institut Teknologi Sepuluh Nopember (ITS) serta Kementerian Riset dan Teknologi/ Badan Riset dan Inovasi Nasional Tahun Anggaran 2021 nomor 3/E1/KP.PTNBH/2021.

Daftar Rujukan

[1] Haviluddin, S. J. Patandianan, G. M. Putra, and H. S. Pakpahan, 2021. "Implementasi Metode K-Means untuk Pengelompokkan Rekomendasi Tugas Akhir," *Inform. Mulawarna J. Ilm. Ilmu Komput.*, 16 (1), pp. 13–18.

[2] P. M. Prihatini, I. K. Suryawan, and I. N. Mandia, 2017. "Feature Extraction for Document Text Using Latent Dirichlet Allocation," *2nd Int. Jt. Conf. Sci. Technol.*, Bali, Indonesia 27-28 September 2017. IOP Publishing Ltd.

[3] L. Farokhah and R. Aditya, 2017. "Implementasi K-Means Klustering untuk Rekomendasi Tema Tugas Akhir pada Stmik Asia Malang," *J. Teknol. dan Manaj. Inform.*, 3 (2), pp. 142–148.

[4] M. R. Muttaqin and M. Defriani, 2020. "Algoritma K-Means untuk Pengelompokan Topik Skripsi Mahasiswa," *Ilk. J. Ilm.*, 12 (2), pp. 121–129.

[5] M. Sholehhudin, M. Fauzi Ali, and S. Adinugroho, 2018. "Implementasi Metode Text Mining dan K-Means Clustering untuk Pengelompokan Dokumen Skripsi (Studi Kasus : Universitas Brawijaya)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, 2 (11), pp. 5518–5524.

[6] V. Kurnia Bakti and J. Indriyatno, 2017. "Klasterisasi Dokumen Tugas Akhir Menggunakan K-Means Clustering, sebagai Analisa Penerapan Sistem Temu Kembali," *KOPERTIP J. Ilm. Manaj. Inform. dan Komput.*, 1 (1), pp. 31–34.

[7] L. Yasni, I. M. I. Subroto, and S. F. C. Haviana, 2018. "Implementasi Cosine Similarity Matching Dalam Penentuan Dosen Pembimbing Tugas Akhir," *Transmisi*, 20 (1), pp. 1–7.

[8] B. K. Triwijoyo and K. Kartarina, 2019. "Analysis of Document Clustering based on Cosine Similarity and K-Main Algorithms," *J. Inf. Syst. Informatics*, 1 (2), pp. 164–177.

[9] Z. Shahbazi and Y.-C. Byun, 2020. "Analysis of Domain-Independent Unsupervised Text Segmentation using LDA Topic Modeling over Social Media Contents," *Int. J. Adv. Sci. Technol.*, 29 (6), pp. 5993–6014.

[10] A. R. Destarani, I. Slamet, and S. Subanti, 2019. "Trend Topic Analysis using Latent Dirichlet Allocation (LDA) (Study Case: Denpasar People's Complaints Online Website)," *J. Ilm. Tek. Elektro Komput. dan Inform.*, 5 (1), pp. 50–58.

[11] H. Shimodaira, 2015. "Similarity and recommender systems," *Japan Similarity Recomm. Syst.*, 20 January 2015, pp. 1–25.

[12] C. Juliane, R. Dzulkarnaen, and W. Susanti, 2019. "Metode McCall's untuk Pengujian Kualitas Sistem Informasi Administrasi Tugas Akhir (SIATA)", *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 3(3), 488 - 495.

DOI: https://doi.org/10.29207/resti.v5i3.3049