CLUSTERING TOPIK PENELITIAN BERBASIS UNSUPERVISED LEARNING UNTUK REKOMENDASI KOLEKSI PUSTAKA DI PERPUSTAKAAN ITS

Dini Adni Navastara[1)], Eva Mursidah[2)], Yeni Anita Gonti[3)], Davi Wahyuni[4)], Petrus Damianus Sammy Wiyadi[5)], dan Wahyu Suadi[6)]

[1, 2, 5, 6)] Departemen Informatika, Fakultas Teknologi Informasi dan Komunikasi, Institut Teknologi Sepuluh Nopember (ITS)
[3, 4)] Perpustakaan, Institut Teknologi Sepuluh Nopember (ITS)
Kampus ITS, Sukolilo, Surabaya
e-mail: dini_navastara@if.its.ac.id[1)], eva.mursidah@gmail.com[2)], yagyag7@gmail.com[3)], davi.wahyuni@gmail.com[4)], petrus.dsw@gmail.com[5)], wsuadi@if.its.ac.id[6)]

ABSTRAK

Sistem rekomendasi adalah sistem yang dapat memberikan saran bagi pengguna terhadap item/produk yang bermanfaat. Perpustakaan ITS yang merupakan salah satu penyedia jasa informasi di ITS memerlukan sistem rekomendasi untuk koleksi pustaka yang akan dibeli. Hal ini dikarenakan pustakawan mengalami kesulitan dalam melakukan proses seleksi judul-judul bahan pustaka yang masih dilakukan secara manual. Sehingga proses pengadaan bahan pustaka tidak berjalan maksimal dan bahan pustaka, khususnya buku, yang dibeli kebanyakan tidak sesuai dengan kebutuhan pengguna. Oleh karena itu, pada penelitian ini diusulkan clustering topik penelitian berbasis unsupervised learning untuk rekomendasi koleksi pustaka di Perpustakaan ITS. Penelitian ini terdiri dari beberapa tahapan proses yaitu: text preprocessing, proses ekstraksi fitur, proses clustering, dan tahap rekomendasi. Text preprocessing dilakukan untuk memperbaiki kualitas data teks, sehingga dapat menghasilkan klaster yang relevan dan akurat. Langkah-langkah pada tahap text preprocessing adalah case folding, tokenizing, filtering, dan stemming. Kemudian, dilakukan proses ekstraksi fitur yaitu dengan teknik pembobotan menggunakan Term Frequency dan Inverse Document Frequency (TF-IDF). Fitur-fitur yang dihasilkan pada tahap ekstraksi fitur dilakukan proses clustering menggunakan metode unsupervised learning yaitu K-Means clustering untuk menghasilkan klaster topik penelitian. Tahap terakhir adalah proses rekomendasi yang dilakukan berdasarkan hasil klaster topik penelitian. Dataset yang digunakan berasal dari tugas akhir dan tesis mahasiswa. Pengukuran evaluasi pada penelitian ini menggunakan nilai silhouette coefficient. Berdasarkan hasil pengujian, rata-rata nilai silhouette coefficient terbaik yaitu sebesar 0.7119 dengan nilai K=3 pada K-Means clustering dan menggunakan metode Principal Component Analysis (PCA).

***Kata Kunci**: clustering, koleksi pustaka, sistem rekomendasi, topik penelitian, unsupervised learning.*

ABSTRACT

Recommendation system is a system providing suggestion to user about the useful item/product. ITS Library that is one of the information service provider in ITS, needs the recommendation system for library collections. This is because the librarians have difficulty in selecting the library collections manually. Therefore, the collections procurement process is not working well and the library collections, especially books, which are purchased mostly are not according to user needs. This research proposes the clustering of research topic based on unsupervised learning for recommendation of library collections in ITS Library. This research consists of four main processes: text preprocessing, feature extraction, clustering, and recommendation process. Text preprocessing is applied to enhance text quality, therefore it can obtain relevant and accurate cluster. The steps in text preprocessing are case folding, tokenizing, filtering, and stemming. Then, feature extraction is implemented by using Term Frequency dan Inverse Document Frequency (TF-IDF). The features are clustered by using unsupervised learning (K-Means clustering) to produce the research topic cluster. And the last step is the recommendation process based on research topic cluster. Datasets that are used in this research are from the theses of undergraduate and graduate student. In this research, silhouette coefficient is used to evaluate clustering performance. Based on the experimental results, the best mean of silhouette coefficient is 0.7119 with K=3 in K-Means clustering and using Principal Component Analysis (PCA).

***Keywords:** clustering, recommendation system, library collection, research topic, unsupervised learning.*

I. PENDAHULUAN

SISTEM rekomendasi merupakan sebuah sistem yang memberikan saran kepada pengguna tentang item/produk yang bermanfaat [1]. Saran tersebut berhubungan dengan proses pengambilan keputusan yang akan dilakukan oleh pengguna. Oleh karena itu, sistem rekomendasi membutuhkan model rekomendasi yang tepat agar produk

yang direkomendasikan sesuai dengan keinginan pengguna dan memudahkan pengguna untuk mengambil keputusan dalam menentukan produk yang digunakan [2].

Sistem rekomendasi dapat digunakan di unit-unit bidang pendidikan, salah satunya di unit Perpustakaan ITS. Perpustakaan ITS memerlukan sistem rekomendasi untuk pengadaan koleksi bahan pustaka yang akan dibeli. Berbagai koleksi fisik yang dikelola meliputi buku teks, buku tugas akhir, buku tesis, jurnal, majalah, serta prosiding seminar nasional maupun internasional. Setiap tahun, perpustakaan ITS senantiasa mengalokasikan dana untuk pengadaan buku baik koleksi buku wajib, buku penunjang maupun buku pengembangan.

Selama ini pustakawan mengalami kesulitan dalam memperoleh informasi judul bahan pustaka, khususnya buku, sehingga dana yang disediakan tidak terserap secara maksimal yang mengakibatkan usulan daftar buku baru untuk perpustakaan dan Unit Layanan Pengadaan (ULP) tidak mendapatkan *feedback* yang baik. Cara yang dilakukan untuk memperoleh informasi buku baru hanya berupa kegiatan survei minat pemakai melalui *form* usulan pengadaan buku, yang tersedia di website perpustakaan ITS. Selain itu, pustakawan juga merujuk pada katalog silabus departemen sehingga dengan cara tersebut judul buku baru yang dihasilkan tidak beragam dan tidak sesuai dengan kebutuhan pengguna, dalam hal ini adalah civitas akademika.

Salah satu upaya untuk menyerap dana alokasi tersebut adalah mencari informasi buku baru sebagai bahan pustaka yang sesuai dengan kebutuhan pengguna berbasis teknologi informasi agar memudahkan pustakawan. Berdasarkan data pengadaan buku di perpustakaan ITS, koleksi buku lebih didominasi oleh buku pengembangan yang mendukung referensi publikasi ilmiah. Publikasi ilmiah yang dilakukan oleh para dosen mayoritas merupakan luaran dari penelitian dosen. Oleh karena itu, pada penelitian ini diusulkan *clustering* topik penelitian berbasis *unsupervised learning* sebagai rekomendasi pengadaan bahan pustaka di Perpustakaan ITS. Salah satu metode *unsupervised learning* yang sangat populer adalah K-Means Clustering. Beberapa penelitian sebelumnya terkait *text mining* telah menerapkan metode K-Means clustering, diantaranya Deka dkk melakukan proses klasterisasi judul buku menggunakan K-Means clustering [3]. Peneliti lain yaitu Ferlyna mengkombinasikan metode K-Means Clustering dengan Genetic Algorithm untuk mengelompokkan pengguna pada Badan Perpustakaan dan Kearsipan [4].

Penelitian ini menerapkan konsep *text mining* yang terdiri dari beberapa tahapan proses yaitu: *text preprocessing*, proses ekstraksi fitur menggunakan teknik pembobotan *Term Frequency* dan *Inverse Document Frequency* (TF-IDF), proses clustering menggunakan metode *unsupervised learning*, dan tahap rekomendasi buku berdasarkan hasil klaster topik penelitian. Diharapkan hasil penelitian ini dapat memudahkan pustakawan untuk mengambil keputusan dalam pengadaan buku yang efisien, bermanfaat dan sesuai dengan kebutuhan pengguna serta meningkatkan jumlah peminjaman di perpustakaan ITS.

II. METODE PENELITIAN

Penelitian ini terdiri dari 4 tahapan utama, yaitu: *text preprocessing*, ekstraksi fitur, clustering, dan rekomendasi. Diagram alir sistem ditunjukkan pada Gambar 1.

A. Text Preprocessing

Text preprocessing merupakan proses pengolahan data teks untuk mendapatkan representasi terstruktur dari data tekstual mentah yang tidak terstruktur. Penggalian informasi tidak dapat dilakukan pada sumber daya teks mentahan dengan format teks yang masih bercampur antara karakter, kata, dan kalimat [5, 9]. Sehingga *text preprocessing* merupakan proses fundamental atau dasar yang harus dilakukan untuk mendapatkan informasi inti dari dokumen teks. Informasi inti inilah yang akan diteruskan dan diolah ke tahapan proses selanjutnya.

Sumber data dokumen teks awal dikatakan mentahan karena pasti berisi hal format angka, tanggal, kata-kata umum yang tidak bermakna, serta kata-kata inti sudah berubah menjadi kata lain karena preposisi dan imbuhan-imbuhan yang ada. Dan hal-hal tersebut menutupi informasi inti dari suatu dokumen teks [6].

Tahap *text preprocessing* pada penelitian ini terdiri dari:

- Case Folding
  Case folding merupakan tahapan yang mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf ‘a’ sampai dengan ‘z’ yang diterima. Sehingga yang tampak berubah hanya format huruf, karakter selain huruf tidak akan terlihat perubahan.
- Tokenizing
  Tokenizing adalah proses penghilangan tanda baca pada kalimat yang ada dalam dokumen sehingga menghasilkan kata-kata yang berdiri sendiri-sendiri, secara istilah disebut token. Hasil dari proses ini tentu dokumen teks akan bersih dari tanda baca, angka, atau karakter-karakter lain yang tidak bermakna, dan antar token akan terpisah dengan spasi [12]. Contoh Tokenization ditunjukkan pada Gambar 2.

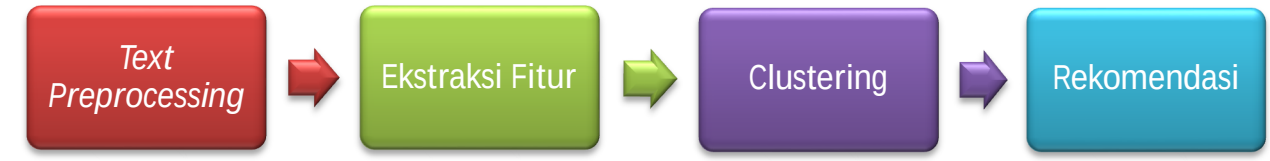
- Filtering
  Banyaknya kemunculan kata-kata dalam dokumen dapat diartikan sebagai kata yang mewakili dokumen, namun hal ini akan menjadi rusak dan mengaburkan inti dokumen dengan kata-kata penghubung yang juga akan sering muncul, seperti ‘and’, ‘then’, ‘but’, ‘after’, dan lain-lain. Kata-kata inilah yang akan menjadi hambatan untuk mengambil intisari dokumen, dan harus dihilangkan untuk dapat menyaring kata-kata bermakna yang mewakili dokumen. Hasil dari proses tokenizing akan disaring dan dihilangkan dari kata-kata tak bermakna yang masuk dalam daftar kata *stopword* [12]. Kumpulan dari *stopword* ini disebut *stop list*. Penggunaan *stop list* ini dapat mengurangi jumlah term yang disimpan dalam kamus dan dapat membuat waktu *indexing* menjadi efisien. Hasil dari tahapan ini tentu daftar kata yang mewakili intisari dokumen. Contoh *stop list* ditunjukkan pada Gambar 3.
- Stemming
  Stemming adalah proses yang dilakukan untuk mengambil bentuk dasar dari suatu kata yang telah melalui proses *filtering* [9]. Algoritma *stemming* untuk bahasa yang satu berbeda dengan algoritma stemming untuk bahasa lainnya. Sebagai contoh bahasa Inggris memiliki morfologi yang berbeda dengan bahasa Indonesia sehingga algoritma *stemming* untuk kedua bahasa tersebut juga berbeda. Proses *stemming* pada teks berbahasa Indonesia lebih rumit/kompleks karena terdapat variasi imbuhan yang harus dibuang untuk mendapatkan *root word* (kata dasar) dari sebuah kata. Pada umumnya kata dasar pada bahasa Indonesia terdiri dari kombinasi misalnya “berjalan”, “menjalani”, “perjalanan” sama-sama memiliki kata dasar “jalan”.

B. Ekstraksi Fitur

Proses penggalian informasi dokumen teks atau lebih khususnya disebut teknologi sistem temu kembali dokumen teks, memiliki 2 tahapan awal yang penting. Selain *text preprocessing* yang telah dijabarkan di atas, proses selanjutnya yang harus dilakukan adalah representasi teks. Tahap representasi teks ini biasa dikenal dengan tahapan pembobotan teks. Telah banyak penelitian-penelitian yang mengusulkan metode-metode baru untuk pembobotan teks, namun metode pembobotan yang sampai saat ini masih sering digunakan (*popular*) dan hasilnya masih dianggap yang terbaik karena efisien, sederhana, dan akurat adalah Metode *Term Frequency – Inverse Document Frequency*, atau lebih dikenal dengan metode pembobotan Algoritma TF-IDF [8].

Algoritma TF-IDF mempertimbangkan seringnya kemunculan *term* (kata) dalam dokumen dan rasio panjang dokumen tersebut di dalam *corpus* (sekumpulan dokumen teks) [10]. Bobot dari perhitungan TF-IDF inilah yang menggambarkan seberapa pentingnya *term* (kata) dalam sebuah dokumen dan *corpus*.

Frekuensi kemunculan (*term frequency*) merupakan petunjuk sejauh mana *term* tersebut mewakili isi dokumen atau secara formula diartikan sebagai ukuran seringnya kemunculan sebuah *term* dalam sebuah dokumen dan juga dalam seluruh dokumen di dalam *corpus*. Semakin besar kemunculan suatu *term* dalam dokumen akan memberikan nilai



Gambar 1. Diagram Alir Sistem Rekomendasi

Input: Friends, Romans, Countrymen, lend me your ears;
Output: Friends | Romans | Countrymen | lend | me | your | ears

Gambar 2. Contoh proses tokenization [10]

a	an	and	are	as	at	be	by	for	from
has	he	in	is	it	its	of	on	that	the
to	was	were	will	with					

Gambar 1. Contoh stop list dari Reuters-RCV1

kesesuaian yang semakin besar. *Term frequency* ini dihitung menggunakan Persamaan (1) dengan $tf_i(d_j)$ adalah notasi frekuensi kemunculan *term* ke-*i* dalam dokumen ke-*j*.

Faktor kebalikan yang diperhatikan juga dalam pemberian bobot TF-IDF adalah kejarangmunculan *term* (*Inverse Document Frequency*) dalam koleksi atau secara formula diartikan sebagai logaritma dari rasio jumlah seluruh dokumen dalam *corpus* dengan jumlah dokumen yang memiliki *term* yang dimaksud seperti yang dituliskan secara matematis pada Persamaan (2), di mana idf_i adalah frekuensi kemunculan *term* ke-i dalam seluruh dokumen atau satu *corpus*. *Term* yang muncul pada sedikit dokumen harus dipandang sebagai *term* yang lebih penting (*uncommon terms*) daripada *term* yang muncul pada banyak dokumen (Karmayasa & Mahendra, 2012).

Bobot TF-IDF sendiri menggabungkan kedua faktor penting di atas, dan secara formula nilai TF-IDF didapatkan dengan mengalikan nilai TF dengan nilai IDF, ditunjukkan pada Persamaan (3) di mana $(tf\text{-}idf)_{ij}$ adalah nilai bobot *term* ke-*i* dalam dokumen ke-*j*.

$$tf_i(d_j) = \frac{freq_i(d_j)}{\sum_{i=1}^{k} freq_i(d_j)} \quad (1)$$

$$idf_i = \log \frac{|D|}{|\{d: t_i \in d\}|} \quad (2)$$

$$(tf - idf)_{ij} = tf_i(d_j) \,.\, idf_i \quad (3)$$

dimana,

$freq_i(d_j)$ adalah frekuensi *term* ke- i dalam dokumen ke- j.

$\sum_{i=1}^{k} freq_i(d_j)$ adalah jumlah *term* pada dokumen ke- j.

$|D|$ adalah jumlah dokumen dalam *corpus*.

$|\{d: t_i \in d\}|$ adalah dokumen yang mengandung *term* ke- i.

Fitur – fitur yang dihasilkan menggunakan metode TF-IDF, kemudian dilakukan proses reduksi dimensi menggunakan metode *Principle Component Analysis* (PCA). PCA diimplementasikan untuk mendapatkan fitur-fitur yang memiliki informasi nilai yang tinggi yang kemudian akan dilakukan proses clustering. Metode PCA dipilih agar fitur-fitur yang digunakan dalam tahap proses clustering merupakan fitur-fitur yang relevan sehingga waktu yang dibutuhkan menjadi lebih cepat dan efisien

PCA merupakan sebuah transformasi linear yang digunakan untuk mereduksi dimensi. Metode ini dipakai untuk mendapatkan komponen-komponen yang saling tidak berkorelasi, dari variable-variabel yang saling berkorelasi, melalui suatu transformasi orthogonal yang melibatkan perhitungan *eigenvalue* dan *eigenvector*.

Step–step pada *principle component analysis* meliputi

1. Menghitung *m*, berdasarkan persamaan (4)

$$m = \frac{1}{nTr} \sum_{i=1}^{nTr} Pi \quad (4)$$

dimana *nTr* merupakan jumlah data sampel *training* dan *Pi* merupakan data sampel *training* pada dimensi *d*.

2. Menghitung *covariance* matrik, berdasarkan persamaan (5)

$$\sum = \sum_{i=1}^{nTr} (Pi - m)(Pi - m)^t \quad (5)$$

3. Menemukan *eigenvalue* dan *eigenvector*, berdasarkan persamaan (6)

$$\sum x = \lambda x \quad (6)$$

4. Men-generate sebuah matrik A *d x k* yang kolomnya terdiri dari *k eigenvector* yang mempunyai *eigenvalue* terbesar, berdasarkan persamaan (7)

$$A = [e_1, e_2, \dots, e_k] \quad (7)$$

5. Tampilkan *original* data dengan memproyeksikan data ke dalam *subspace* k – dimensional, berdasarkan persamaan (8)

$$p' = A^t(p - m) \quad (8)$$

dimana *p'* = [*p1' … pk'*] merupakan PCA *feature vector*

C. Clustering

Pada penelitian ini, metode yang digunakan pada tahap Clustering adalah metode *K-Means clustering* yang merupakan salah satu metode clustering non hirarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih cluster [7, 11]. Metode ini mempartisi data ke dalam cluster sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu cluster yang sama dan data yang mempunyai karateristik yang berbeda di kelompokan ke dalam cluster yang lain. Secara umum algoritma dasar dari K-Means Clustering adalah sebagai berikut:

1. Menentukan jumlah cluster
2. Menentukan nilai centroid
   Dalam menentukan nilai centroid untuk awal iterasi, nilai awal centroid dilakukan secara acak. Sedangkan jika menentukan nilai centroid yang merupakan tahap dari iterasi, maka digunakan rumus pada persamaan (9) sebagai berikut:

$$v_{ij} = \frac{1}{N_i}\sum_{k=0}^{N_i} x_{kj} \quad (9)$$

   dimana :
   v_{ij} adalah centroid/rata-rata cluster ke-*i* untuk variable ke-*j*
   N_i adalah jumlah data yang menjadi anggota cluster ke-*i*
   i,k adalah indeks dari cluster *j* adalah indeks dari variabel *x*
   kj adalah nilai data ke-*k* yang ada di dalam cluster tersebut untuk variable ke-*j*
3. Menghitung jarak antara titik centroid dengan titik tiap objek. Untuk menghitung jarak tersebut dapat menggunakan Euclidean Distance pada persamaan (10), yaitu

$$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2} \quad (10)$$

   dimana :
   De adalah Euclidean Distance
   i adalah banyaknya objek,
   (*x,y*) merupakan koordinat object dan
   (*s,t*) merupakan koordinat centroid.
4. Pengelompokan objek
   Untuk menentukan anggota cluster adalah dengan memperhitungkan jarak minimum objek. Nilai yang diperoleh dalam keanggotaan data pada distance matriks adalah 0 atau 1, dimana nilai 1 untuk data yang dialokasikan ke cluster dan nilai 0 untuk data yang dialokasikan ke cluster yang lain.

Kembali ke tahap 2, lakukan perulangan hingga nilai centroid yang dihasilkan tetap dan anggota cluster tidak berpindah ke cluster lain.

D. Rekomendasi

Setelah proses clustering, dilakukan proses rekomendasi buku berdasarkan hasil cluster. Hasil cluster berupa kata-kata yang penting yang menjadi topik penelitian dari masing-masing grup riset. Kata-kata tersebut dijadikan sebagai kata kunci untuk pencarian buku/jurnal/prosiding pada mesin pencari *google books* sehingga dapat menjadi rekomendasi koleksi pustaka. Informasi yang ditampilkan dari hasil pencarian berupa judul, *cover*, *author*, dan harga buku/jurnal/prosiding.

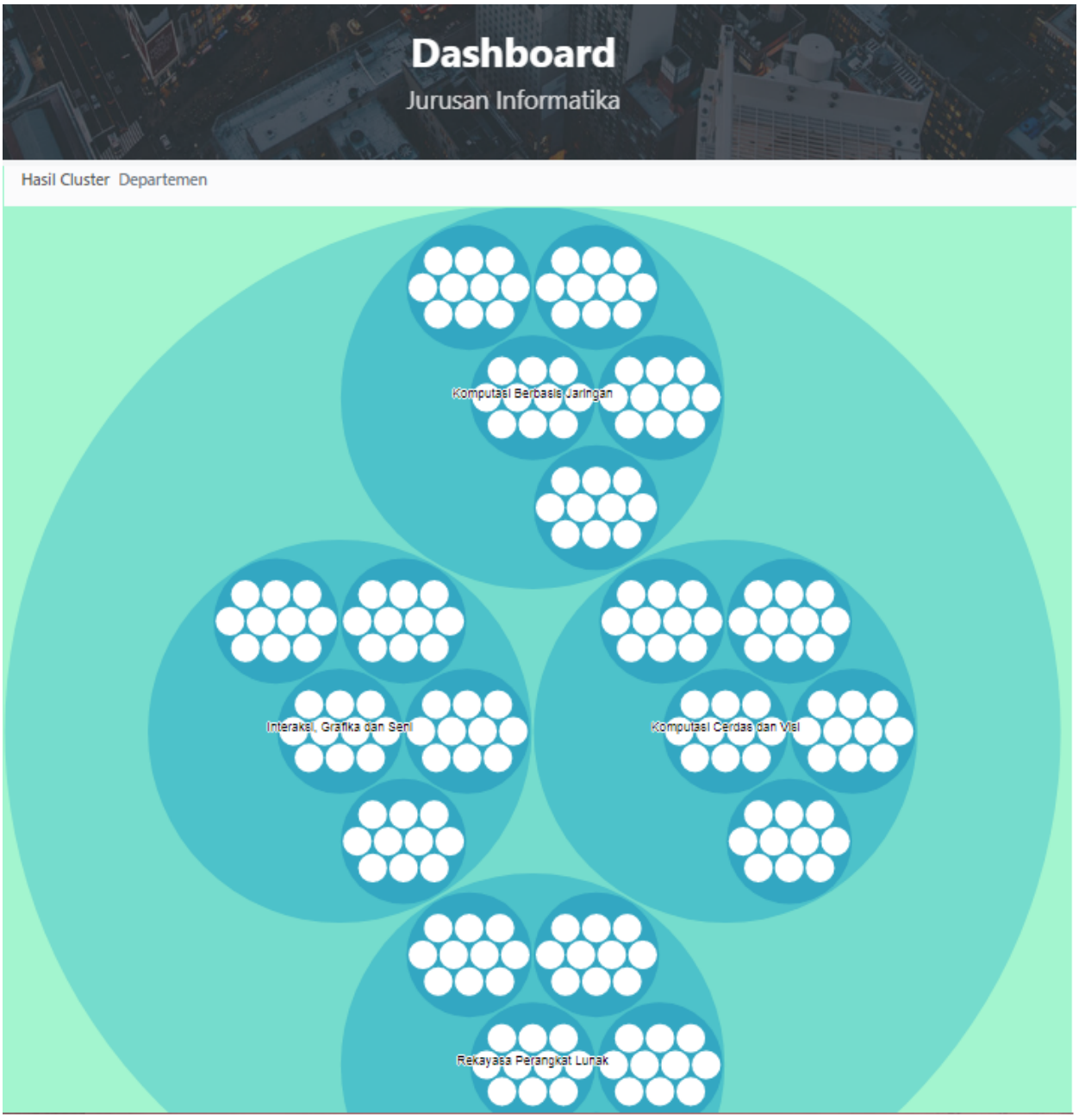
```
Deteksi Kendaraan Bermotor dari Citra CCTV
Menggunakan Gray Level Co-Occurrence Matrix dan
Random Forest
```

(a)

```
deteksi kendara motor citra cctv gray level matrix
random forest
```

(b)

Gambar 4. Contoh Text Preprocessing
(a) Teks asli. (b) Teks hasil preprocessing.



Gambar 5. Tampilan Hasil Clustering

III. HASIL DAN PEMBAHASAN

Data pengujian pada penelitian ini berupa judul dan abstrak tugas akhir dan tesis mahasiswa yang diperoleh dari beberapa departemen di ITS selama 3 tahun terakhir (2015-2017), yaitu departemen Informatika, departemen Sistem Informasi, departemen Teknik Industri, departemen Perencanaan Wilayah Kota, dll. Data tersebut dilakukan pelabelan berdasarkan grup riset/laboratorium yang nantinya akan digunakan sebagai *ground truth* untuk menguji kebenaran hasil klaster.

A. Pengujian Modul Text Preprocessing

Pengguna akan diminta untuk mengunggah data abstrak/judul tugas akhir mahasiswa ke sistem. Kemudian data tersebut dilakukan *text preprocessing.* Tahap *text preprocessing* ini terdiri dari 4 proses, yaitu: *case folding* (mengubah tulisan pada teks menjadi huruf kecil), *tokenizing* (menghilangkan tanda baca, angka, kata-kata yang tidak bermakna sehingga sehingga menghasilkan kata-kata yang berdiri sendiri-sendiri), *stemming* (mengambil bentuk dasar dari suatu kata), dan menghapus *stopword*. Contoh hasil *text preprocessing* ditunjukkan pada Gambar 4(b), dimana teks asli ditunjukkan pada Gambar 4(a).

B. Pengujian Modul Clustering

Pada tahap clustering digunakan metode K-Means Clustering untuk mendapatkan klaster topik-topik penelitian. Pada uji coba ini digunakan data tugas akhir mahasiswa dari departemen Teknik Informatika. Gambar 5 menunjukkan hasil clustering dari data tugas akhir mahasiswa di departemen Teknik Informatika. Pada Gambar 5 terlihat bahwa klaster yang dihasilkan sebanyak 4 klaster, yaitu klaster Komputasi Berbasis Jaringan, klaster Komputasi Cerdas dan Visi, klaster Rekayasa Perangkat Lunak, dan klaster Interaksi, Grafik, dan Seni. Setiap klaster terdapat bulatan-bulatan lebih kecil berwarna putih yang merupakan topik-topik penelitian yang dihasilkan pada klaster tersebut. Ketika bulatan klaster diklik, maka bulatan berwarna putih menjadi lebih besar sehingga terlihat kata-kata yang berupa topik pada bulatan tersebut seperti yang ditunjukkan pada Gambar 6. Terlihat pada Gambar 6 terdapat kata-kata *resource planning* ERP, penggunaan *enterprise*, *enterprise resource system*, dll.

Gambar 6. Tampilan Topik pada Suatu Cluster

Selain itu, pada modul clustering dilakukan pengujian dengan 2 skenario, yaitu pengujian variasi nilai K pada K-Means Clustering dan pengujian berdasarkan penggunaan metode PCA. Evaluasi dilakukan dengan mencari nilai *silhouette coefficient* (SC) pada masing-masing skenario berdasarkan rumus pada persamaan (11). *Silhouette coefficient* merupakan pengukuran yang umum digunakan untuk mengevaluasi performa clustering dengan kondisi ketika cluster sebenarnya tidak diketahui.

$$SC = \frac{b-a}{\max(a,b)} \tag{11}$$

dimana *a* adalah rata-rata jarak ke semua titik di dalam cluster dan *b* adalah rata-rata jarak ke semua titik pada cluster terdekat berikutnya. Nilai *silhouette* terletak pada rentang -1 (terburuk) sampai 1 (terbaik). Semakin tinggi nilai *silhouette*, maka semakin baik hasil clustering.

Tabel I menunjukkan nilai *silhouette* dari hasil pengujian variasi nilai K pada K-Means Clustering dengan menggunakan PCA. Berdasarkan Tabel I, nilai *silhouette* terbaik dihasilkan pada parameter K=3 dengan rata-rata *silhouette* sebesar 0.7119. Hal ini menunjukkan hasil clustering setiap grup riset cukup baik. Nilai *silhouette* terbaik ditunjukkan pada grup riset KCV sebesar 0.8410, yang diikuti oleh grup riset KBJ, IGS, dan RPL dengan nilai *silhouette* berturut-turut sebesar 0.7376, 0.7351, dan 0.5337.

Tabel II menunjukkan nilai *silhouette* dari hasil pengujian berdasarkan penggunaan metode PCA. Pada Tabel II ini dilakukan perbandingan antara penggunaan metode PCA untuk mereduksi dimensi dan tanpa PCA. Berdasarkan Tabel II, penggunaan metode PCA menghasilkan nilai *silhouette* 5 kali lebih tinggi dibandingkan tanpa PCA. Dan tentunya dengan penggunaan PCA ini menyebabkan kompleksitas waktu semakin rendah karena dimensi fitur yang digunakan pada proses clustering berkurang.

C. Pengujian Modul Statistik

Pada tahap ini, pengujian dilakukan dengan cara menampilkan jumlah judul/abstrak tugas akhir mahasiswa pada setiap cluster per tahun (2015, 2016, dan 2017). Hasil data statistik berdasarkan tugas akhir mahasiswa setiap klaster per tahun ditunjukkan pada Gambar 7.

Berdasarkan Gambar 7 dapat dilihat bahwa pada tahun 2015 grup riset Interaksi Grafik dan Seni (IGS) paling banyak diminati dan terus meningkat hingga tahun 2016, terlihat pada jumlah judul tugas akhir paling banyak dibandingkan grup riset Komputasi Berbasis Jaringan (KBJ), Komputasi Cerdas dan Visi (KCV) dan Rekayasa Perangkat Lunak (RPL). Namun, di tahun 2017 IGS mengalami penurunan yang cukup drastis sehingga jumlah judul/abstrak tugas akhir mahasiswa dilampaui oleh grup riset KCV. Grup riset RPL mengalami penurunan peminat dari tahun 2015 sampai 2017, hal ini terlihat dari menurunnya grafik secara statistik. Sedangkan grup riset KBJ mengalami kenaikan peminat dari 2015 sampai 2017.

D. Pengujian Modul Rekomendasi

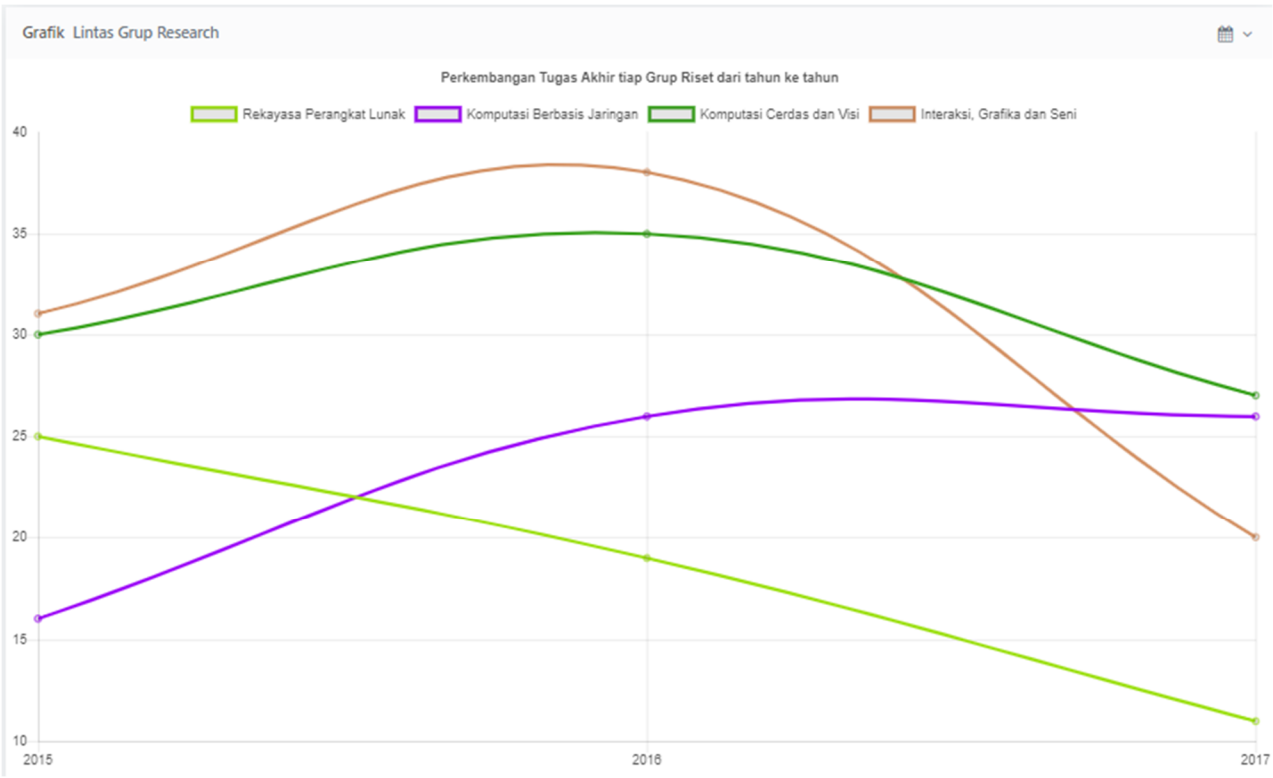
Pada tahap ini, pengujian dilakukan berdasarkan topik-topik penelitian yang dihasilkan dari setiap cluster. Topik-topik tersebut dilakukan *parsing* dan menjadi input pada proses pencarian di *google books*. Modul rekomendasi ini menggunakan API *googlebooks*. Hasil pencarian yang ditampilkan berupa judul buku, penulis buku, *cover* buku, jumlah halaman, dan penerbit seperti terlihat pada Gambar 8.

TABEL I.
NILAI SILHOUETTE UJI COBA VARIASI NILAI K PADA K-MEANS CLUSTERING

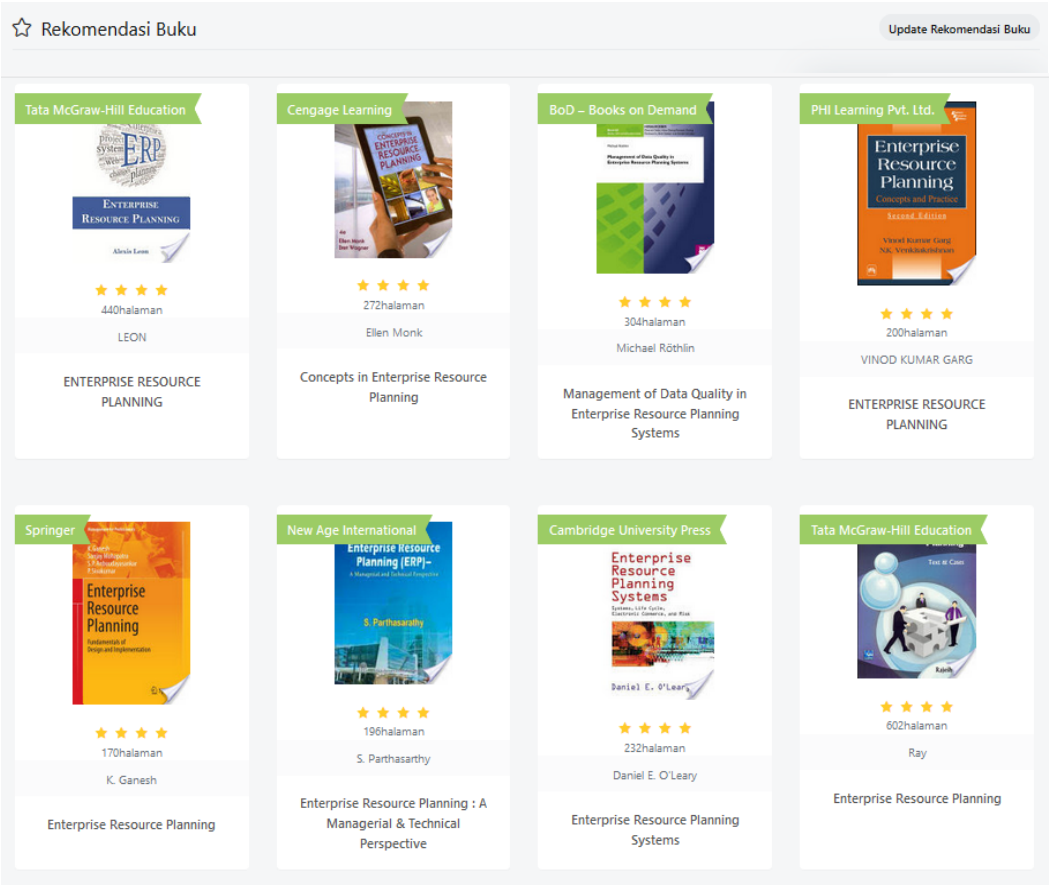
K	IGS	KCV	RPL	KBJ	Rata-Rata
2	0.6458	0.7666	0.4817	0.6454	0.6349
3	**0.7351**	**0.8410**	**0.5337**	**0.7376**	**0.7119**
4	0.5438	0.7375	0.5166	0.6635	0.6154
5	0.5738	0.7438	0.5267	0.6840	0.6321

TABEL II.
NILAI SILHOUETTE UJI COBA PERBANDINGAN BERDASARKAN PENGGUNAAN METODE PCA

Metode	IGS	KCV	RPL	KBJ	Rata-Rata
Dengan PCA	**0.7351**	**0.8410**	**0.5337**	**0.7376**	**0.7119**
Tanpa PCA	0.0673	0.1206	0.1620	0.1398	0.1224



Gambar 7. Tampilan Data Statistik



Gambar 8. Tampilan Hasil Rekomendasi Buku

IV. KESIMPULAN

Clustering topik penelitian berbasis *unsupervised learning* dapat digunakan sebagai rekomendasi untuk pengadaan koleksi pustaka di perpustakaan ITS. Hasil cluster dalam bentuk visual berupa bulatan – bulatan yang berisi kata – kata kunci dari topik penelitian lebih memudahkan untuk mengetahui topik penelitian yang paling banyak diminati di setiap klaster grup riset. Hasil data statistik perkembangan tugas akhir mahasiswa dalam bentuk grafik memudahkan untuk mengetahui tren topik penelitian setiap grup riset beberapa tahun terakhir. Berdasarkan hasil pengujian, rata-rata nilai *silhouette coefficient* terbaik yaitu sebesar 0.7119 dengan nilai K=3 pada K-Means clustering dan menggunakan metode Principal Component Analysis (PCA) untuk mereduksi dimensi. Selain itu, hasil pengujian membuktikan bahwa penggunaan metode Principal Component Analysis (PCA) dapat meningkatkan nilai *silhouette* sebesar 5 kali lipat dibandingkan dengan tanpa menggunakan PCA.

Saran untuk pengembangan penelitian selanjutnya adalah menambah penggunaan dataset sampai 5 tahun terakhir dan memperluas area penggunaan dataset yang diambil dari seluruh departemen di ITS sehingga dapat diketahui tren topik penelitian dari setiap departemen yang memungkinkan adanya kolaborasi penelitian antar departemen. Selain itu, untuk meningkatkan performa hasil clustering dapat dilakukan pengujian menggunakan pendekatan *unsupervised learning* yang lain, seperti *hierarchical clustering analysis*.

UCAPAN TERIMA KASIH

Penelitian ini didanai oleh dana lokal ITS tahun 2018 melalui skema Penelitian Kebijakan dengan nomor kontrak: 1234/PKS/ITS/2018.

DAFTAR PUSTAKA

[1] Ricci, F., Rokach, L. and Shapira, B., 2011. Introduction to recommender systems handbook. In Recommender systems handbook (pp. 1-35). Springer, Boston, MA.
[2] McCarthy, K., Salamó, M., Coyle, L., McGinty, L., Smyth, B. and Nixon, P., 2006. Cats: A synchronous approach to collaborative group recommendation. In Florida Artificial Intelligence Research Society Conference (FLAIRS) (pp. 86-91).
[3] Deka D, Makhfuzi F, Zumrotun N, Novi S,. 2014. Klasterisasi Judul Buku dengan Menggunakan K-Means. Seminar Nasional Aplikasi Teknologi Informasi (SNATI). Yogyakarta.
[4] Ferlyna, K. 2012. Penerapan Metode GA - K Means untuk pengelompokan pengguna pada badan perpustakaan dan kearsipan (BAPERSIP) Provinsi Jawa Timur. Departemen Sistem Informasi ITS. Surabaya.
[5] Larose, D.T. 2005. Discovering Knowledge in Data : An Introduction to Data Mining. John Wiley & Sons, Inc.
[6] Kannan, S. and Gurusamy, V., 2014, October. Preprocessing techniques for text mining. In Conference Paper. India.
[7] Han, J, Kamber, M, Pei, J. 2011. Data Mining Concept and Techniques 3rd Edition. United States : Morgan Kauffman Publisher
[8] Karmayasa, O. & Mahendra, I. B., 2012. Implementasi Vector Space Model dan Beberapa Notasi Metode Term Requency Inverse Document Frequency (TF-IDF) pada Sistem Temu Kembali Infomasi. Jurnal Elektronik Ilmu Komputer Universitas Udayana, 1(1)
[9] Manning, C. D., Raghavan, P. & Schütze, H. 2009. An Introduction to Information Retrieval. Cambridge: Cambridge University Press.
[10] Saadah, M. N., Atmagi, R. W., Rahayu, D. S. & Arifin, A. Z., 2012. Sistem Temu Kembali Dokumen Teks dengan Pembobotan TF-IDF dan LCS. Jurusan Teknik Informatika, Fakultas Teknologi Informasi, Institut Teknologi Sepuluh Nopember.
[11] Tan P.N, Steinbach M, & Kumar, V. 2007. Data mining : Introduction to Data Mining. Graha Ilmu. Pearson Education
[12] Usmaida, A., 2007. Web Mining untuk Pencarian Berdasarkan Kata Kunci dengan Teknik Clustering. Tugas Akhir Jurusan Teknologi Informasi Politeknik Elektronika Negeri Surabaya.