Klasterisasi Dokumen Tugas Akhir Menggunakan *K-Means Clustering* sebagai Analisa Penerapan Sistem Temu Kembali

Very Kurnia Bakti[1,#], Jatmiko Indriyatno[2]

Abstrak— Pencarian dokumen Tugas Akhir (TA) mahasiswa di Politeknik Harapan Bersama saat ini masih menampilkan hasil pencarian berurut berdasarkan peringkat kecocokan dokumen atau biasa disebut *document ranking*. Cara demikian menyebabkan penemuan data dokumen tidak terkelompok pada masing-masing tema Tugas Akhir secara akurat. Algoritma *Clustering* dapat digunakan dalam pengkategorian atau pengelompokan dokumen. Salah satu penggunaan algoritma c*lustering* yaitu dengan menerapkan metode *K-means* yang merupakan algoritma sederhana dan dikembangkan oleh Mac Queen pada tahun 1967. Dari hasil penelitian yang telah dilakukan, pengklasteran dokumen abstrak tugas akhir berbahasa Indonesia dengan menerapkan Algoritma *K-Means* menunjukan klaster yang dihasilkan cukup baik, sehingga dapat dijadikan rekomendasi bahwa metode *K-Means* klastering cukup baik jika diterapkan dalam penerapan sistem temu kembali, dengan indikator jarak antar klaster yang dihasilkan sangat dekat yaitu sebesar 0,001 ketika dihitung dengan metode *Davies Bouldin Index.*

Kata kunci— Tugas Akhir, *k-means, clustering*

Abstract— Document searching of Final Project in Polytechnic Harapan Bersama today still displays search results ranked by sequential document matches or commonly called document ranking. Thereby this way causes document data discovery is not clustered on each theme of the final project accurately. Clustering algorithms can be used in categorising or grouping of documents. One of clustering algorithm usage is by applying the method of K means, a simple algorithm developed by Mac Queen in 1967. From the research that has been done, the document final projects' abstract clustering in Indonesian language by applying the K Means algorithm shows generated a good enough clusters, so it can be recommendation that K-Means clustering method is good enough if applied in retrieval application system, with indicators of distance between clusters produced are very close, that is 0.001 when calculated by the method of Davies Bouldin Index.

Keywords— Final Project, k-means, clustering

I. PENDAHULUAN

Banyaknya jumlah data dokumen tugas akhir dari berbagai program studi di Politeknik Harapan Bersama dapat memberi kontribusi besar dalam sulitnya proses pencarian suatu dokumen. Pencarian dokumen yang ada saat ini hanya menampilkan hasil pencarian berurut berdasarkan peringkat kecocokan berurut. Hal tersebut menyebabkan penemuan data dokumen tidak secara akurat terkelompok pada masing-masing tema tugas akhir.

Dengan adanya pengelompokan dokumen, maka tidak harus membuka halaman terlalu banyak, karena dokumen hasil pencarian telah dikelompokkan berdasarkan kategori yang dapat menggambarkan isi dari suatu dokumen, hal tersebut tentunya dapat mempermudah dalam menemukan beberapa dokumen yang diinginkan, oleh karenanya sebelum proses tersebut dilakukan maka, proses tersebut perlu dianalisis sebelum benar-benar diaplikasikan ke dalam program yang sifatnya aplikatif. Oleh karena itu diperlukan metode pengelompokan (clustering) yang nantinya dapat dipastikan keberhasilan pengelompokan suatu dokumen tugas akhir dengan baik.

Metode *Clustering* dapat digunakan dalam pengkategorian atau pengelompokan dokumen. Caranya adalah dengan mengelompokan dokumen-dokumen ke dalam *clusters* berdasarkan kedekatan atau kemiripan (*similarity*) antar dokumen [1],[5]. Sehingga dokumen yang berhubungan dengan suatu tema tertentu secara otomatis ditempatkan pada *cluster* yang sama. Saat ini ada beberapa algoritma *clustering* diantaranya *K-Means Clustering* dan *hierarchical*[1]. Pengelompokan dengan Metode *K-means* adalah algoritma sederhana yang dikembangkan oleh Mac Queen pada tahun 1967. Algoritma tersebut terkenal dengan kemampuannya untuk mengklaster data yang besar dan dapat menangani data yang menyimpang terlalu jauh dari data yang lainnya dalam suatu rangkaian data (*data outliers*). Dengan metode *K-means*, penerapan pengklasteran dilakukan dengan cara memisahkan data kedalam beberapa kelompok *(k)* yang berbeda, artinya sebelum dilakukan klasterisasi maka perlu menentukan jumlah *k* yang diinginkan. Selain itu *k-means* dapat membentuk titik pusat *cluster (centroid)* yang dapat menentukan setiap klaster dari titik pusat klasternya [1],[3].

II. METODE PENELITIAN

A. Pengumpulan Data

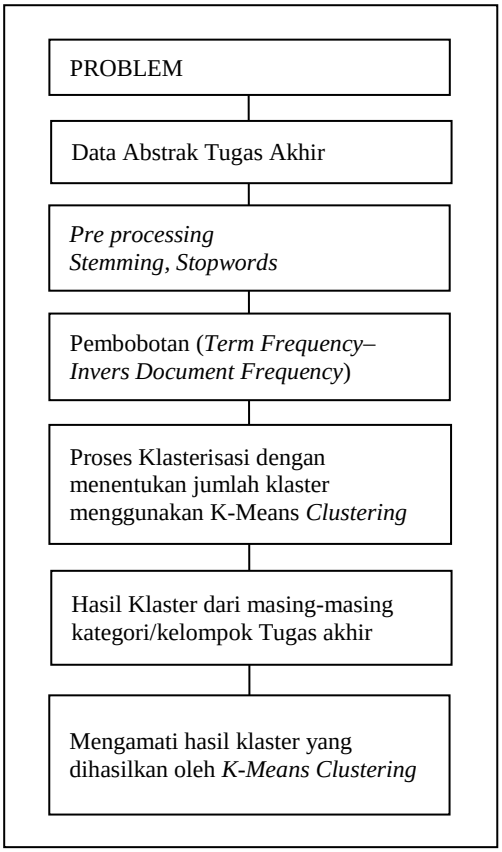
Bahan yang digunakan dalam penelitian ini berupa data-data laporan tugas akhir berbentuk soft copy dari empat program studi Diploma tiga (D3) yaitu D3 Teknik Komputer, D3 Farmasi, D3 Akuntansi dan D3 Kebidanan yang datanya ada di perpustakaan Politeknik Harapan Bersama. Data laporan Tugas akhir tersebut diambil secara acak dari tahun 2013 -2015 dengan masing-masing program studi sebanyak 50 Laporan Tugas Akhir, sehingga jumlah keselurah laporan tugas akhir adalah 200 laporan. Dengan rincian seperti pada Tabel 1.

Artikel diterima 10 Desember 2016; direvisi 13 Januari 2017; disetujui 14 Januari 2017; dipublikasikan Februari 2017

[1,2]Program Studi Teknik Komputer, Politeknik Harapan Bersama, Jl.Mataram No. 9, Kota Tegal, Jawa Tengah 52147, Indonesia
[#] E-mail: verykurniabakti@poltektegal.ac.id

TABEL I
PENGELOMPOKAN MANUAL TUGAS AKHIR

Pengelompokan Manual	
Nama Program Studi	**Jumlah Tugas Akhir**
D3 Teknik Komputer	50
D3 Kebidanan	50
D3 Akutansi	50
D3 Farmasi	50



Gbr. 1 Prosedur penelitian

B. Alat Penelitian

Alat yang digunakan pada saat penelitian adalah perangkat keras dan perangkat lunak komputer. Perangkat keras yang dibutuhkan berupa komputer/laptop Sedangkan perangkat lunak yang dibutuhkan adalah: (1) Sistem Operasi *Windows* yang digunakan sebagai sistem untuk menjalankan aplikasi pemrograman, (2) Aplikasi *Rapidminer* 5.3 untuk menganalisa.

C. Prosedur Penelitian

Penelitian dilakukan dengan langkah awal untuk memperoleh data berupa 200 Laporan berbentuk *softcopy* dengan format doc dan docx. dari data yang ada kemudian dilakukan pemilahan data-data yang dibutuhkan saja, yaitu abstrak dari masing-masing laporan. Data abstrak tersebut kemudian dijadikan kedalam format txt, hal ini dilakukan bertujuan untuk memangkas waktu dalam melakukan proses analisis pada aplikasi *rapidminer*. Prosedur penelitian ini dijelaskan pada Gbr. 1.

TABEL III
HASIL KATA PROSES *CASE FOLDING*

alpinia	bangsa	coefficient
alri	bangun	coklat
alternatif	bangunan	coli
alternative	bani	coliform
altilis	bank	collected
alun	banteng	collection
alur	bantu	coloni
amalia	bantuan	coming
amaliyah	banyak	commerce

III. HASIL DAN PEMBAHASAN

A. Ekstraksi Dokumen

Dalam tahap ini data abstrak dari masing-masing tugas akhir yang berbentuk dokumen berformat txt diperlukan *pre-processing*, yaitu menyamakan huruf besar dan huruf kecil (*case folding*), tokenisasi (*Tokenizing*), mendapatkan kata dasar (*stopwords*) dan pembobotan.

B. Case Folding

Dari dokumen tugas akhir yang ada selanjutnya masuk kedalam proses *case folding* dimana proses ini melakukan penyamaan antar kata dengan cara mengubah huruf besar menjadi seluruhnya huruf kecil. Sehingga seluruh kata yang diproses semuanya menjadi huruf kecil. Dari hasil pemrosesan tersebut dapat diambil contoh hasil proses *case folding* seperti pada Tabel 2.

C. Tokenizing

Proses selanjutnya setelah melalui tahap tokenisasi (*tokenization*). Tokenisasi dilakukan terhadap kata atau frasa di dalam dokumen. Namun, kata-kata yang tidak memberikan perbedaan atau kata yang merupakan ekspressi verbal dari suatu pengertian (*term*) seperti kata “ini, itu, saya, kamu”, serta tanda baca dihilangkan atau dianggap bukan *term.* Hal ini bertujuan untuk mendapatkan hanya kata-kata tertentu saja yang nantinya didapat dan berkontribusi sebagai ciri dari masing-masing jenis judul tugas akhir. Selain itu masih dalam proses *tokenizing* dilakukan pula *filter token* yaitu dengan memberikan batasan jumlah karakter dari masing-masing kata minimal 3 karakter dan maksimal 25 karakter dari setiap kata. Hal ini dilakukan untuk menyortir atau menghilangkan kata salah pengetikan karena terlalu pendek atau terlalu panjang dalam tiap kata.

D. Stopwords

Proses tahap berikutnya yang dilakukan dalam penelitian ini adalah mendapatkan kata dasar (*stopwords*), dikarenakan ditiap dokumen tugas akhir banyak terdapat kata yang memiliki banyak imbuhan. Hal tersebut akan mempengaruhi hasil klaster nantinya. Dalam proses *stopwords* inilah nantinya tiap kata-yang memiliki kata dasar yang sama akan dihilangkan imbuhannya sehingga didapat hanya kata dasar saja. Langkah menerapkan *stopwords* ini dilakukan dengan cara manual yaitu dengan membuat kamus *stopwords* sendiri sebanyak 8093 kata yang tentunya masih sedikit jika dibandingkan dengan jumlah kata pada bahasa Indonesia. Hal ini dikarenakan pada *rapidminer* tidak terdapat kamus kata *stopwords* berbahasa Indonesia, sehingga dengan mengikuti

prosedur pada *rapidminer* pembuatan *stopwords* dibuat tersendiri.

Dalam proses pencarian kata dasar ini, mengingat kata yang terdapat dalam dokumen tugas akhir tidak hanya kata berbahasa Indonesia saja, melainkan juga terdapat kata dengan bahasa Inggris, maka dalam penelitian ini dilakukan pemrosesan sebanyak dua kali *stopwords* yaitu dengan *stopwords* bahasa Indonesia dan *stopwords* bahasa Inggris. Dengan harapan hasil kata dasar yang dihasilkan memiliki kontribusi dalam menentukan hasil klaster nantinya.

E. Pembobotan Term Frequency-Inverse Document Frequency (TF-IDF)

Tahap pembobotan ini menggunakan Metode TF-IDF dimana terdapat integrasi antar *term frequency* (TF), dan inverse *document frequency* (IDF) dengan rumusan berikut:

$$w(t,d) = tf(t,d) * \log 2\left(\frac{N}{nt}\right) \quad (1)$$

Simbol *w(t,d)* merupakan bobot dari *term t* dalam sebuah dokumen tugas akhir *d* sedangkan *tf(t,d)* adalah frekuensi term dalam dokumen tugas akhir *(tf)* dan *N* merupakan ukuran data training yang diterapkan dalam memperoleh hasil hitungan IDF. Adapun *nt* merupakan jumlah dari banyaknya dokumen yang ditraining dan mengandung nilai *t*.

F. Proses Klasterisasi dengan K-Means Clustering

Dari tahapan-tahapan yang telah dilalui dengan proses ekstraksi dokumen, langkah selanjutnya adalah proses mengklaster dokumen. Dalam penelitian ini proses klaster dokumen dilakukan dengan menggunakan algoritma *k-means clustering* dengan pertimbangan beberapa refrensi dari penelitian sebelumnya dengan tema yang sama yaitu, *information retrieval* dengan *text mining*. Pada pengklasteran menggunakan *k-means clustering* dengan dasar algoritmanya adalah sebagai berikut:

1) Langkah pertama adalah dengan mengelompokan atau inisialisasi klaster dalam penelitian ini dibuat empat klaster disesuaikan dengan jumlah dokumen empat program studi, D3-Teknik Komputer, D3-Kebidanan, D3-Akutansi, D3-Farmasi. Masing-masing program studi ada 50 judul tugas akhir.
2) Memasukan semua dokumen ke klaster yang paling cocok dengan berdasarkan pusat klaster. Dengan persamaan yang disajikan pada Tabel 3.

TABEL IIIII
DOKUMEN YANG PALING COCOK BERDASARKAN PUSAT KLASTER

word	in documents	total	in class (komputer)	in class (akutansi)	in class (kebidanan)	in class (farmasi)
Aaffecting	1	1	0	0	0	0
abad	1	2	0	0	0	0
abortus	1	1	0	0	0	0
Abrasiver	1	1	0	0	0	0

TABEL IVV
DOKUMEN YANG PALING COCOK BERDASARKAN PUSAT KLASTER

Pengelompokan Manual		**Pengelompokan dengan K-Means**	
Nama Program Studi	**Jumlah Tugas Akhir**	**Jumlah Klaster**	**Anggota Klaster**
D3 Teknik Komputer	50	Cluster 0:	46 items
D3 Kebidanan	50	Cluster 1:	66 items
D3 Akutansi	50	Cluster 2:	45 items
D3 Farmasi	50	Cluster 3:	43 items

Hasil Perbandingan antara pengelompokan secara manual dengan pengelompokan menggunakan metode *k-Means* terdapat perbedaan namun, hasil antar klaster yang terbentuk hampir rata. Hasil Pengelompokan tersebut dapat dilihat pada Tabel 4.

IV. KESIMPULAN

Dari hasil penelitian yang telah dilakukan dapat diambil kesimpulan bahwa pengklasteran dokumen abstrak tugas akhir berbahasa Indonesia dengan menerapkan Algoritma *K-Means*, menghasilkan nilai jarak antar klaster yang lebih baik berdasarkan nilai *davies bouldin index* 0,001 Dengan demikian, metode *K-Means clustering* sangat baik digunakan dalam penerapan aplikasi sistem temu kembali tugas akhir dikarenakan hasil dari klaster yang dibentuk *K-Means* sudah dapat mengelompokan TA berdasarkan tema tugas akhir masing-masing program studi.

REFERENSI

[1] Agusta, Y. 2007. K-means-Penerapan, Permasalahan dan Metode Terkait. Jurnal Sistem dan Informatika Vol. 3 (Februari 2007): 47-60.

[2] Arifin, Agus Zainal, and Ari Novan Setiono. "Klasifikasi Dokumen Berita Kejadian Berbahasa Indonesia dengan Algoritma Single Pass *Clustering*." *Prosiding Seminar on Intelligent Technology and its Applications (SITIA), Teknik Elektro, Institut Teknologi Sepuluh Nopember Surabaya.[This page intentionally left blank]*. 2002.

[3] Cui, Xiaohui, Thomas E. Potok, and Paul Palathingal. "Document *clustering* using particle swarm optimization." *Swarm Intelligence Symposium, 2005. SIS 2005. Proceedings 2005 IEEE*. IEEE, 2005.

[4] Gosno, Eric Budiman, Isye Arieshanti, and Rully Soelaiman. "Implementasi KD-Tree K-Means ustering untuk Klasterisasi Dokumen." *Jurnal Teknik ITS* 2.2 (2013): A432-A437.

[5] Haryo Guritno. "Klasterisasi Dokumen Cerpen Dengan Metode K-Means Clustering" thesis Udinus (2015).

[6] Huang, Anna. "Similarity measures for text document clustering." Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand. 2008.

[7] Selim, Shokri Z., and Mohamed A. Ismail. "K-means-type algorithms: a generalized convergence theorem and characterization of local optimality." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 1 (1984): 81-87.

[8] Tala, Fadillah Z. "A study of stemming effects on information retrieval in Bahasa Indonesia." *Institute for Logic, Language and Computation Universeit Van Amsterdam* (2003).

[9] Vidya Ayuningtias, M. Arif Bijaksana, Rimba Widhiana Ciptasari "Pengkategorian hasil Pencarian Dokumen dengan klastering"*tugas akhir, Universitas telkom university*. 2008

[10] Yang, Yiming, et al. "Learning approaches for detecting and tracking news events." *IEEE Intelligent Systems* 4 (1999): 32-43.
[11] Yi, B., Qiao, H., Yang, F., & Xu, C. (2010). An Improved Initialization Center Algorithm for K-Means Clustering. 2010 *International Conference on Computational Intelligence and Software Engineering*, *IEEE* (1), 1–4.