Klasterisasi Dokumen Penelitian Perguruan Tinggi Menggunakan *K-Means Clustering,* Sebagai Analisa Penerapan Sistem Temu Kembali

Very Kurnia Bakti[1], Arif Rakhman[2]
Email : verykurniabakti@gmail.com, cakrakirana7@gmail.com
[1.2] DIII Teknik Komputer Politeknik Harapan Bersama

Abstrak

Pencarian dokumen yang ada saat ini yaitu menampilkan hasil pencarian berurut berdasarkan peringkat kecocokan (document ranking). Hal tersebut menyebabkan penemuan data dokumen tidak secara akurat terkelompok pada masing-masing tema. *Clustering* dapat digunakan dalam pengkategorian atau pengelompokan dokumen. *Clustering* dengan metode K means adalah algoritma sederhana yang dikembangkan Mac Queen pada tahun 1967. Dari hasil penelitian yang telah dilakukan pengklasteran dokumen abstrak penelitian dosen berbahasa Indonesia dengan menerapkan Algoritma K Means, klaster yang dihasilkan cukup baik, sehingga dapat dijadikan rekomendasi bahwa metode K-Means klastering baik digunakan dalam penerapan sistem temu kembali. Dengan nilai Davies Bouldin Index sebesar -6.186.

Kata Kunci: *dokumen, k-means, clustering*

1. Pendahuluan

Banyaknya jumlah data dokumen penelitiandari berbagai program studi dapat memberi kontribusi besar dalam sulitnya proses pencarian suatu dokumen. Pencarian dokumen yang ada saat ini hanya menampilkan hasil pencarian berurut berdasarkan peringkat kecocokan (document ranking). Hal tersebut menyebabkan penemuan data dokumen tidak secara akurat terkelompok pada masing-masing tema.

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 aplikasi yang sifatnya *executable*. Oleh karena itu diperlukan metode pengelompokan/ *clustering* yang nantinya dapat dipastikan keberhasilan pengelompokan suatu dokumen penelitiandengan baik.

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

2. Metode Penelitian

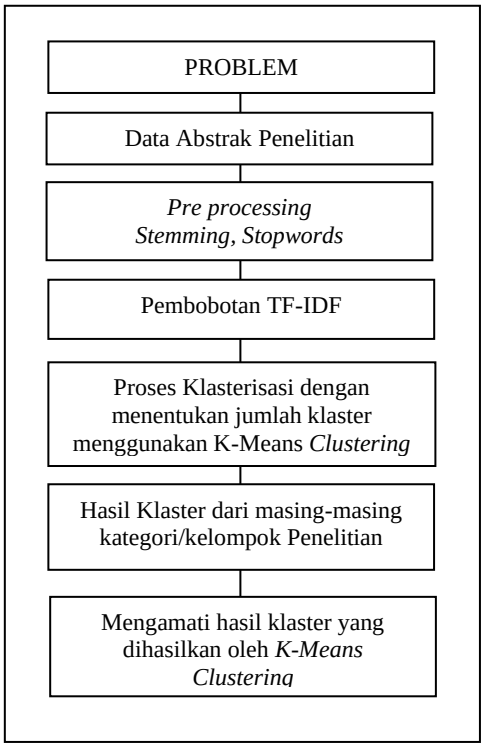
Bahan Penelitian yang digunakan dalam penelitian berupa, data Penelitian Perguruan Tinggiyang diambil pada bagian abstraknya saja. Dari masing-masing 7 program studi di Politeknik Harapan Bersama.

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

Dokumen yang digunakan dalam penelitian ini diambil dari sampel abstrak Penelitian Perguruan Tinggi dari P3M

Politeknik Harapan Bersama dari 7 program studi.



Gambar 1. Prosedur Penelitian

3. Hasil dan Pembahasan

Dalam tahap ini data abstrak dari masing-masing penelitianyang berbentuk dokumen berformat *.txt diperlukan proses *pre-processing,* yaitu *case folding, Tokenizing, stopwords* dan pembobotan.

Dari dokumen penelitianyang ada selanjutnya masuk kedalam proses *case folding* dimana proses ini melakukan penyamaan antar kata dengan cara merubah huruf besar menjadi seluruhnya huruf kecil. Sehingga seluruh kata yang diproses semuanya menjadi huruf kecil. Dari hasil pemrosesan tersebut dapat diambil contoh hasil pemrosesan *case folding* seperti pada tabel 1.

Proses selanjutnya setelah melalui tahap *case folding* adalah proses Pemisahan rangkaian term (*tokenization*). Term dapat berupa kata atau frasa di dalam dokumen. Namun, kata-kata yang tidak memberikan perbedaan seperti ini, itu, saya, kamu, serta tanda-tanda baca dihilangkan atau dianggap bukan *term.* Hal ini bertujuan untuk mendapatkan hanya kata – kata tertentu saja yang nantinya didapat dan berkontribusi sebagai ciri – 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 mensortir atau menghilangkan kata salah pengetikan karena terlalu pendek atau terlalu panjang dalam tiap kata.

Stopwords dilakukan dalam penelitian ini untuk mendapatkan kata dasar, dikarenakan ditiap dokumen penelitianbanyak 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. Proses 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.

Dalam proses pencarian kata dasar ini, mengingat kata yang terdapat dalam dokumen penelitiantidak 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.

Tahap pembobotan ini menggunakan Metode TF-IDF dimana terdapat integrasi antar term frequency (tf), dan inverse document frequency (idf) Dengan rumus :

$$w(t,d) = tf(t,d) * log2(N/nt) \text{ ‘}$$

Simbol w(t,d) merupakan bobot dari term *t* dalam sebuah dokumen penelitian*d* sedangkan *tf(t,d)* adalah frekuensi term dalam dokumen penelitian*(tf)* dan *N* merupakan ukuran data training yang digunakan untuk penghitungan IDF. Adapun nt adalah jumlah dari dokumen yang ditraining yang mengandung nilai *t*.

Tabel 1 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

Dari tahapan-tahapan yang telah dilalui dengan proses ekstraksi dokumen, maka 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, D 3- Teknik Komputer, D 3- 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 sebagai berikut:

Tabel 2. Dokumen yang paling cocok berdasarkan pusat klaster

word	in documents	total	in class (komputer)	in class (akutansi)	in class (kebidanan)	in class (farmasi)
aaffecting	1,0	1,0	,0	1,0	,0	,0
abad	1,0	2,0	2,0	,0	,0	,0
abortus	1,0	1,0	,0	,0	1,0	,0
abrasiver	1,0	1,0	,0	,0	,0	1,0

4. Kesimpulan

Dari hasil penelitian yang telah dilakukan pengklasteran dokumen abstrak penelitian berbahasa Indonesia dengan menerapkan Algoritma K Means, klaster yang dihasilkan cukup baik, sehingga dapat dijadikan rekomendasi bahwa metode K-Means klastering baik digunakan dalam penerapan sistem temu kembali. *Pre processing* pada tiap data penelitianharus tetap dilakukan karena proses, *case folding*, *stemming*, *stopwords*, dan term mempengaruhi hasil klaster yang dibentuk. Dalam penelitian ini hasil klaster dari k-means clustering mendapatkan nilai DBI sebesar -6.186.

5. Daftar Pustaka

[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 Clustering 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.