Penerapan Algoritma K-Means untuk *Clustering* Dokumen E-Jurnal STMIK GI MDP

Ernie Kurniawan*[1], Maria Fransiska[2], Tinaliah[3], Rachmansyah[4]
[1-4]STMIK Global Informatika MDP Jl. Rajawali No.14 Palembang
[1,2]PS Teknik Informatika STMIK Global Informatika MDP,
[3]PS Manajemen Informatika, [3]PS Teknik Komputer AMIK MDP
e-mail: *[1]deepblue_nie_k@yahoo.com, [2]mariafransiska09@yahoo.com, [3]tinaliah@mdp.ac.id, [4]rachmansyah@gmail.com

Abstrak

Banyaknya dokumen jurnal yang terus bertambah membuat pengelompokkan dokumen jurnal semakin sulit karena memperlambat pencarian dokumen. Oleh karena itu, pengelompokkan dokumen jurnal diperlukan untuk mempercepat pencarian yang diperoleh dari query yang diinput pengguna dan menghasilkan hasil yang relevan dengan query tersebut. Penelitian ini bertujuan untuk menerapkan algoritma K-Means dalam mengelompokkan dokumen jurnal yang sesuai dengan query yang diinput sehingga menghasilkan kelompok-kelompok yang sesuai dengan query. Dalam prosesnya dilakukan tahap preprocessing yaitu tokenization, penghilangan stopwords dan stemming. Selanjutnya, pengelompokkan dokumen dilakukan dengan algoritma K-Means menggunakan bahasa pemrograman PHP dengan menggunakan proses stemming dan non-stemming untuk mengetahui kerelevanan hasil yang diperoleh dari masing-masing proses. Hasil dari pengelompokkan dokumen dengan menggunakan proses stemming menghabiskan waktu lebih banyak dibandingkan dengan proses non-stemming karena proses stemming harus menemukan terlebih dahulu kata dasar dari query yang diinput sedangkan dalam proses non-stemming tidak diperlukan proses pencarian kata dasar. Pengujian dengan menggunakan dataset yang sedikit membuat pengembang kesulitan dalam membedakan hasil kelompok antara proses stemming dan proses non-stemming. Penentuan titik pusat awal sangat berpengaruh terhadap jumlah cluster yang terbentuk serta pengelompokan dokumen ini juga dapat membantu pengguna menemukan dokumen yang relevan sesuai dengan query yang diinput.

Kata kunci— Clustering, K-Means, stemming, PHP

Abstract

The amount of journal documents which increase continously make the classified of journal document more difficult is slow down the document research. Therfore the classification of journal document is needed to speed the research which get from query that input by the user and produce a relevant result from the query. The purpose of the research is to imply K-Means algorithm in classify the journal document based on the query which already input so produce the groups as query. In that process, there is preprocessing step which is call tokenization, the omit of stopwords and stemming. Next, the classification of documents are done with K-Means algorithm use PHP programming language with use stemming and non-stemming process to know the relevance result with get from each process. The result from document classification with use stemming process spend more times compare with non-stemming process because the stemming process should be found the basic words of query which already input. While in non-stemming process isn't needed of the basic words research process. The result of this classification with use stemming process is more relevant compare with the classification by non-stemming process. The testing with use a little dataset make the developer find the difficulty in comparing the result of groups between stemming and non-stemming process, the determination of the beginning of center is very influence to the

amount of cluster which is formed beside that the classification of this document can help the user find the relevant document as suit as query which is input.

Keywords— Clustering, K-Means, stemming, PHP

1. PENDAHULUAN

Clustering dokumen merupakan metode yang digunakan untuk melakukan pengelompokan dokumen, dimana dokumen yang dikelompokan sesuai dengan informasi yang akan dicari [1]. Misalnya pada mesin pencari sering memberikan ribuan halaman dalam menanggapi permintaan pengguna, sehingga sulit bagi pengguna untuk mencari atau mengidentifikasi informasi yang relevan.

Dengan adanya pengelompokan dokumen ini, mahasiswa hanya perlu mengetikkan judul tugas akhirnya, lalu secara otomatis *web* akan memberikan dokumen-dokumen yang mirip dengan judul yang dimasukkan agar tidak terjadi penelitian dengan judul yang sama. Hal ini dikarenakan di dalam *website* terdapat sebuah algoritma yang dapat mengelompokan dokumen berdasarkan kesamaan antar dokumen tersebut atau berdasarkan kelompoknya.

Dengan adanya sebuah sistem yang berfungsi untuk mengelompokan dokumen jurnal ini, dapat mempermudah mahasiswa dalam pencarian jurnal dengan tingkat kemiripan yang paling sesuai dengan *query* yang diberikan oleh pengguna.

2. METODE PENELITIAN

2.1 Clustering

Clustering adalah suatu metode untuk penggelompokan dokumen dimana dokumen dikelompokan dengan konten untuk mengurangi ruang pencarian yang diperlukan dalam merespon suatu *query*. Misalnya koleksi dokumen yang berisi dokumen-dokumen medis dan hukum dapat dikelompokkan sedemikian rupa sehingga semua dokumen medis ditempatkan dalam satu *cluster* dan semua dokumen hukum ditempatkan dalam satu *cluster* hukum [1].

2.2 Algoritma K-Means

Algoritma K-Means merupakan algoritma yang membutuhkan parameter input sebanyak *k* dan membagi sekumpulan *n* objek ke dalam *k cluster* sehingga tingkat kemiripan antar anggota dalam satu *cluster* tinggi sedangkan tingkat kemiripan dengan anggota pada *cluster* lain sangat rendah [2]. Kemiripan anggota terhadap *cluster* diukur dengan kedekatan objek terhadap nilai *mean* pada *cluster* atau dapat disebut sebagai *centroid cluster* atau pusat massa. [2].

Rumus untuk menentukan jumlah *cluster* yang digunakan pada algoritma k-means adalah sebagai berikut [2]:

$$k = \approx \sqrt{\frac{n}{2}} \quad (1)$$

Rumus untuk mengukur jarak antar objek adalah sebagai berikut [2]:

$$d_{(x,y)} = \sqrt{(xi - yi)^2 + (xi - yi)^2} \quad (2)$$

Keterangan :
d = titik dokumen
x = data *record*
y = data *centroid*

Jarak yang terpendek antara *centroid* dengan dokumen menentukan posisi *cluster* suatu dokumen. Misalnya dokumen *A* mempunyai jarak yang paling pendek ke *centroid 1* dibanding ke yang lain, maka dokumen *A* masuk ke *group 1*. Hitung kembali posisi *centroid* baru untuk tiap-tiap *centroid* ($C_{i..j}$) dengan mengambil rata- rata dokumen yang masuk pada *cluster* awal ($G_{i..j}$). Iterasi dilakukan terus hingga posisi *group* tidak berubah.

Rumus dari penentuan *centroid* adalah sebagai berikut :

$$C(i) = \frac{1}{|Gi|} \sum xec \;\; d\bar{x} \quad (3)$$

Adapun rumus iterasi lainnya didefinisikan sebagai berikut :

$$C(i) = \frac{x1+x2+x..+x...}{\sum x} \quad (4)$$

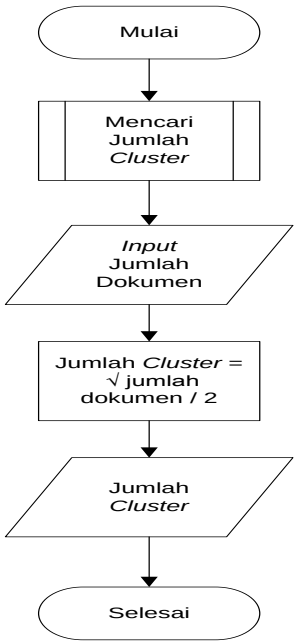
Keterangan :
x_1 = nilai data *record* ke-1
x_2 = nilai data *record* ke-2
Σx = jumlah data *record*

Tahap penyelesaian algoritma K-Means adalah sebagai berikut [2]:

a. Menentukan *K* buah titik yang merepresentasikan obyek pada setiap *cluster* (*centroid* awal).
b. Menetapkan setiap objek pada *cluster* dengan posisi *centroid* terdekat.
c. Jika semua objek sudah dikelompokkan maka dilakukan perhitungan ulang dalam menentukan *centroid* yang baru.
d. Ulangi langkah ke-2 dan ke-3 sampai *centroid* tidak berubah.

2.2.1 Flowchart Mencari Jumlah *Cluster*

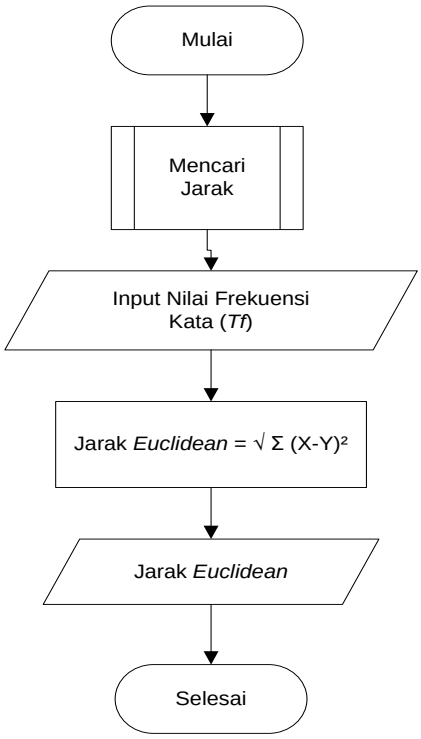
Flowchart mencari jumlah *cluster* merupakan *flowchart* yang berisi proses pencarian jumlah *cluster* dengan cara membagi dua jumlah dari seluruh dokumen kemudian diakarkan. *Flowchart* mencari jumlah *cluster* dapat dilihat pada Gambar 1.



Gambar 1 *Flowchart* Mencari Jumlah *Cluster*

2.2.2 Flowchart Mencari Jarak

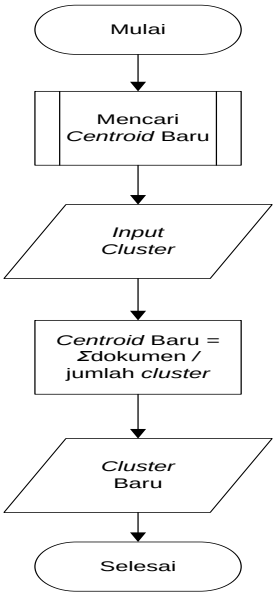
Flowchart mencari jarak merupakan *flowchart* yang berisi proses pencarian jarak antara dokumen dengan titik *centroid* dimana proses dilakukan dengan menghitung nilai frekuensi kata yang ada pada tiap dokumen, kemudian dilakukan perhitungan jarak dengan *Euclidean*. *Flowchart* mencari jarak dapat dilihat pada Gambar 2.



Gambar 2 *Flowchart* Mencari Jarak

2.2.3 Flowchart Mencari *Centroid* Baru

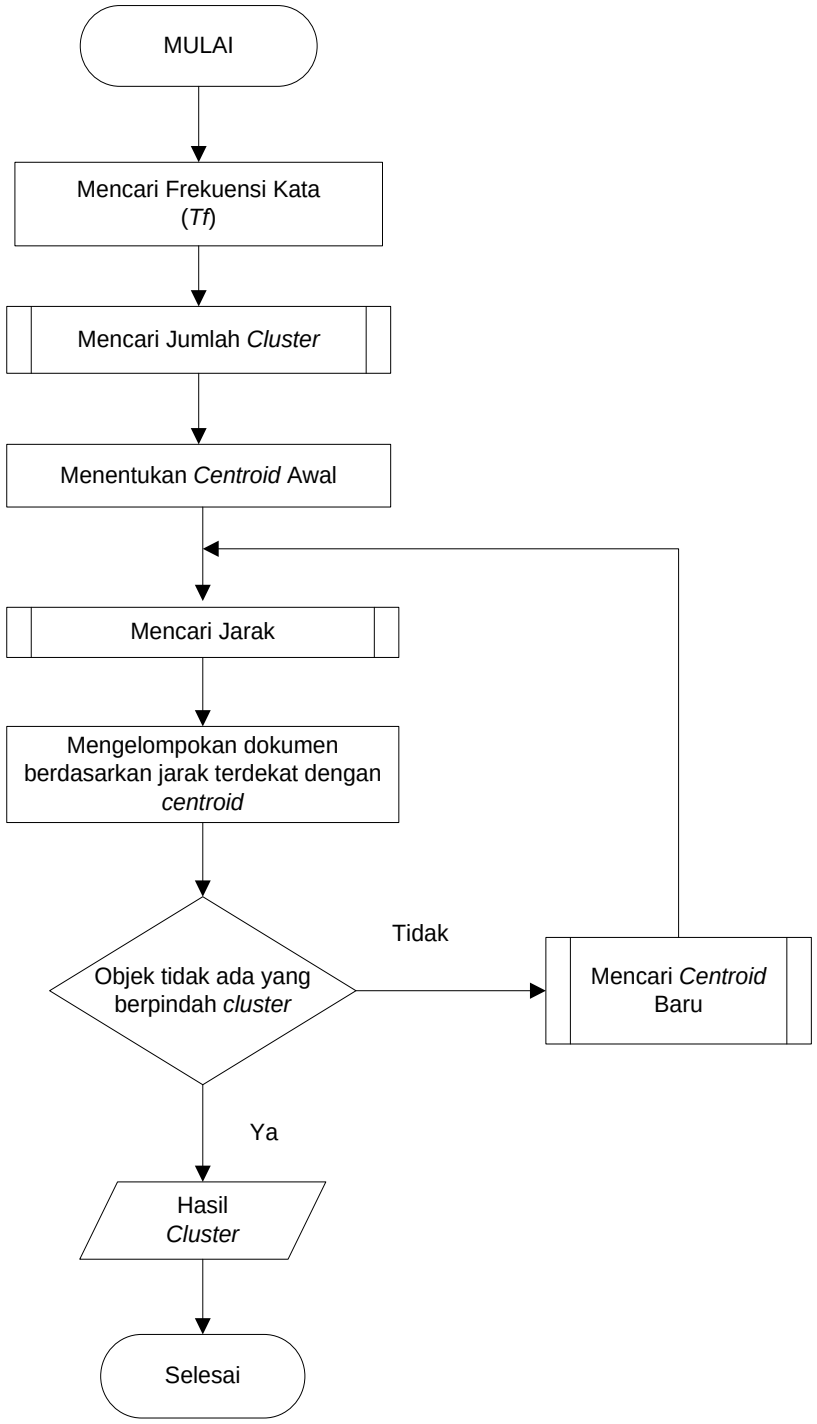
Flowchart mencari *centroid* baru merupakan *flowchart* yang berisi proses pencarian *centroid* (titik pusat) baru dengan cara membagi jumlah seluruh dokumen dengan jumlah *cluster* yang terbentuk. *Flowchart* mencari *centroid* baru dapat dilihat pada Gambar 3.



Gambar 3 *Flowchart* Mencari *Centroid* Baru

2.3 Flowchart Algoritma K-Means

Flowchart Algoritma K-Means merupakan *flowchart* yang berisi urutan proses dari mencari frekuensi kemunculan kata *(Tf)*, mencari jumlah *cluster*, menentukan *centroid* (titik pusat) awal, mencari jarak, mengelompokkan dokumen berdasarkan jarak terdekat dengan *centroid*, serta proses mencari *centroid* baru. *Flowchart* algoritma k-means dapat dilihat pada Gambar 4.



Gambar 4 *Flowchart* Algoritma K-Means

3. HASIL DAN PEMBAHASAN

3.1 Prosedur Uji Coba Program

3.1.1 Tampilan Antarmuka Menu Utama

Ketika *user* menjalankan aplikasi maka halaman pertama yang akan tampil adalah halaman menu utama. Tampilan antarmuka menu utama dapat dilihat pada Gambar 5.

Gambar 5 Antarmuka Halaman Menu Utama

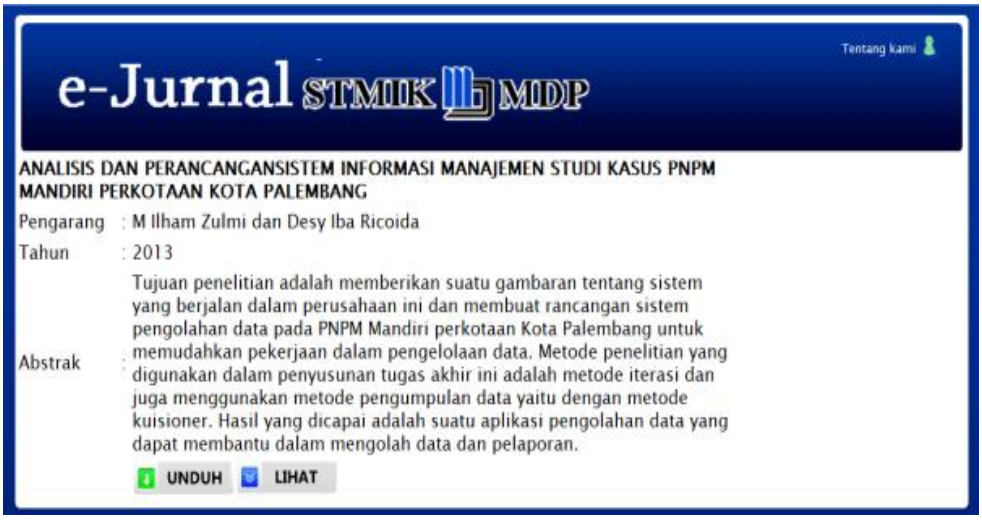
3.1.2 Tampilan Antarmuka Halaman Pencarian

Pada saat *user* memasukkan *query* yang ingin dicari dan menekan tombol cari maka akan tampil halaman pencarian. Tampilan antarmuka halaman pencarian dapat dilihat pada Gambar 6.

Gambar 6 Tampilan Antarmuka Halaman Pencarian

3.1.3 Tampilan Antarmuka Halaman Isi Dokumen

Pada saat *user* memilih salah satu jurnal maka akan menampilkan halaman isi dokumen. Tampilan antarmuka halaman isi dokumen dapat dilihat pada Gambar 7.



Gambar 7 Tampilan Antarmuka Halaman Isi Dokumen

3.2 Analisis Hasil Pengujian Program

3.2.1 Uji Coba Pertama

Uji coba pertama dilakukan untuk menguji keakuratan sistem dengan membandingkan hasil yang diperoleh dari aplikasi menggunakan algoritma K-Means dengan klasifikasi judul yang ada pada *database* aplikasi dimana dilakukan proses pencarian dengan cara memasukkan 5 *query* yang sama ke dalam masing-masing aplikasi.

$$\text{Nilai akurasi} = \frac{\text{jumlah dokumen yang sama antara } \textit{clustering} \text{ dan klasifikasi}}{\text{jumlah dokumen keseluruhan}}$$

Query = " perancangan sistem informasi "

Tabel 1 Hasil Kesamaan *Clustering* dan Klasifikasi

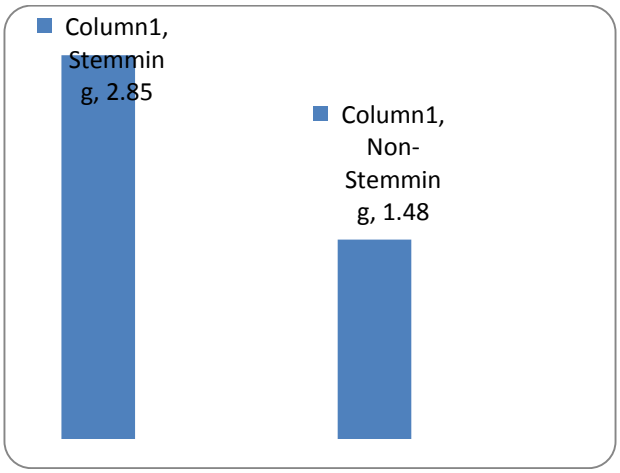
Clustering	Klasifikasi
J0001	J0004
J0007	J0007
J0016	J0008
J0017	J0016
J0028	J0017
J0040	J0018
J0060	J0026
J0074	J0027
J0076	J0028
J0077	J0077

Pada Tabel 1 diperoleh bahwa terdapat 5 buah dokumen yang sama pada aplikasi *clustering* dan klasifikasi sehingga dapat diperoleh nilai akurasinya adalah sebagai berikut :

$$\text{Nilai akurasi} = \frac{5}{10} \text{ x } 100\% = 50\%$$

3.2.2 Uji Coba Kedua

Hasil uji coba waktu antara proses stemming dengan proses non-stemming dengan *query*: "aplikasi pembelian barang" terhadap 300 dokumen jurnal. Grafik hasil ujicoba pengujian waktu clustering dokumen dapat dilihat pada Gambar 8.



Gambar 8 Grafik Hasil Uji Coba Pengujian Waktu

4. KESIMPULAN

1. Algoritma K-Means dapat melakukan pengelompokan dokumen dalam jumlah yang banyak akan tetapi belum efisien dalam mengelompokan dokumen secara tepat.
2. Penentuan *centroid* (titik pusat) pada tahap awal Algoritma K-Means sangat berpengaruh pada hasil *cluster* seperti pada hasil pengujian yang dilakukan dengan menggunakan 300 dataset dengan *centroid* yang berbeda menghasilkan hasil *cluster* yang berbeda juga.
3. Proses *clustering* menggunakan *stemming* akan menghabiskan waktu lebih lama dibandingkan dengan *non-stemming*, hal ini dapat dilihat pada hasil uji coba 2.
4. Semakin sedikit dokumen yang dipakai, maka semakin sulit untuk membedakan *cluster* antara *stemming* dan *non-stemming*.

5. SARAN

1. Untuk meningkatkan hasil pengelompokan dokumen yang lebih relevan sebaiknya algoritma K-Means digabung dengan algoritma lain seperti Algoritma *Hierarchical Clustering*.
2. Aplikasi ini dapat dikembangkan dengan cara menambah fitur *convert file* dan standarisasi sehingga dapat mempermudah kerja admin.
3. Agar aplikasi dapat digunakan untuk umum, sebaiknya aplikasi dibuat secara *online*.

DAFTAR PUSTAKA

[1] Grossman, David A. dan Ophir Frieder 2004. *Information Retrieval Algorithms and Heuristics Second Edition*. Springer, The Netherlands.

[2] Nango, Dwi Noviati 2014. *Penerapan Algoritma K-means untuk Clustering Data Anggaran Pendapatan Belanja Daerah di Kabupaten XYZ*. Universitas Negeri Gorontalo.