PENERAPAN ALGORITMA K-MEANS UNTUK CLUSTERING PENENTUAN JURUSAN BAHASA MANDARIN GERMAN DAN PRANCIS

Ardi Mardiana
Fakultas Teknik, Universitas Majalengka
Email: aim@ft.unma.ac.id

Abstract

Grouping language department based on academic data using clustering techniques and create applications then analyze the results that are expected to provide the information concerned. K-Means algorithm is a clustering algorithm technique that starts with a random selection of K, which is the number of clusters to be formed from the data to be in clusters, namely student test scores when entering the majors language. System created to show the results of student academic data clustering, ie the pattern of student achievement lusternya remain, down and up, and can be seen from the data value of the test results. From the results of the case study can be obtained information of students who remain in clusters such as early admission, students who ride down clusters and clusters of students.

***Keywords** : clustering, Algoritma K-Means.*

I. PENDAHULUAN

Seleksi masuknya siswa dalam sebuah sekolah menengan atas umumnya dengan memberikan soal-soal test yang harus mereka kerjakan, untuk mengetahui kemampuan dan pengetahuan mereka. Setelah siswa mengalami proses tes untuk masuk jurusan bahasa, maka akan dapat diketahui termasuk jurusan bahasa yang terbagi menjadi 3 yaitu bahasa mandarin, german dan prancis. Jumlah data yang banyak ini membuka peluang untuk dihasilkan informasi yang berguna bagi pihak sekolah.

II.

III.

3.1.

3.2.

Penggalian informasi pada sebuah data yang berukuran besar (mempunyai jumlah record dan jumlah field yang cukup banyak) tidak dapat dilakukan dengan mudah. Teknologi data mining merupakan salah satu alat bantu untuk penggalian data pada basisdata berukuran besar dan dengan spesifikasi tingkat kerumitan yang telah banyak digunakan pada banyak domain aplikasi seperti perbankan maupun bidang telakomunikasi. Algoritma K-Means merupakan algoritma teknik klustering yang berulang ulang . algoritma ini dimulasi dengan pemilihan secara acak K, yang merupakan banyaknya kluster yang ingin dibentuk. Kemudian tetapkan nilai nilai K secara random, untuk sementara nilai tersebut pusat dari kluster atau biasa disebut dengan centroid / mean.

3.3.

Hitung jarak setiap data yang ada terhadap masing-masing centroid menggunakan rumus yang sudah disediakan hingga diketemukan jarak yang paling dekat dari setiap data dengan centroid. Klasifikasi setiap data berdasarkan kedekatannya dengan centroid. Lakukan langkah tersebut sampai nilai centroid tidak berubah (stabil). Data akademik tersebut adalah hasil evaluasi tes masuk penerimaan siswa baru (PSB) berupa nilai tes bahasa mandarin, german dan prancis. Dengan menggunakan data hasil tes masuk, maka dapat diketahui minat belajar dari siswa apakah tetap pada nilai test awal masuk atau ada perubahan yang signifikan.

II. LANDASAN TEORI

Clustering

Clustering adalah suatu metode untuk pengelompokan dokumen dimana dokumen dikelompokan dengan konten untuk mengurangi ruang pencarian yang diperlukan dalam merespon suatu query. Misalnya koleksi dokumen yang berisi dokumen-dokumen medis danhukum dapat dikelompokan sodem ikian rupa sehingga semua dokumen medis ditempatkan dalam satu cluster dan semua dokumen hokum ditempatkan dalam satu cluster hokum (Grossman, David A. dan OphirFrieder, 2004, h,105).

Algoritma K-Mens

Algoritma K-Mens merupakan algoritma yang membutuhkan parameter input sebanyak K dan membagi sekumpulan objek ke dalam k cluster

sehingga tingka kemiripan antar anggota dalam satu cluster tinggi sedangkan kemiripan dengan anggota. Pada cluster lain sangat rendah. Kemiripan anggota terhadap cluster diukur dengan kedekatan objek terhadap nilai mean pada cluster atau dapat disebut sebagai centroid cluster atau pusat massa. (nango, Dwi N viati, 2012). Berikut ini adalah rumus untuk menentukan jumlah cluster :

$$K = \sqrt{n}/2$$

Berikut rumus pengukuran jarak :

$D(x,y) = |x-y||^2 = \sum$

Pembaharuan suatu titik centroid dapat dilakukan dengan rumus berikut :

$$\mu_k = \frac{1}{Nk} \sum_{q=1}^{Nk} x_q$$

Dimana :

μ_k = titik centroid dari cluster ke-K

N_k = banyaknya data pada cluster ke-K

X_q = data ke-q pada cluster ke-K

Algoritma K-Means

Menurut hasn & kember algoritma K-Means bekerja dengan membagi data ke dalam K buah cluster yang telah ditentukan.

Beberapa cara penghitungan jarak yang biasa digunakan yaitu :

✓ *Euclidean distance*

Formula jarak antara dua titik dalam satu, dua dan tiga dimensi secara berurutan ditunjukan pada formula 1,2,3 berikut ini :

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2}$$

Atau

$$d(x,y) = \left[\sum_{i=1}^{p} |x_i - y_i|^2\right]^{1/2}$$

Pengelompokan Siswa Menggunakan K-means

Klasifikasi K-means

Beberapa teknik klastering yang paling sederhana dan umum adalah klastering K-means. Secara detail teknik ini menggunakan ukuran ketidak miripan untuk mengelompokan obyek. Ketidak miripan dapat diterjemahkan dalam konsep jarak. Dua obyek di katakn miripjika jarak dua objek tersebut dekat. Semakin tinggi nilai jarak, seakin tinggi nilai ketidak miripannya.

Algoritma Klastering K-means dapat di ringkas sebagai berikut: (santosa, 2007)

1. Pilih jumlah klaster.
2. Inisialisasi k pusat klaster (di beri nilai-nilai random).
3. Tempatkan setiap data/objek ke klaster terdekat. Kedekatan dua obyek di tentukan berdasarkan kedua obyek tersebut. Jarak jarak paling dejatantara satu data dengan satu klaster tertentu akan menentukan suatu data masuk dalam klaster mana.
4. Hitung kembali pusat klaster dengan anggota klaster yang sekarang. Pusat klaster adalah rata-rata semua data/obyek dalam klaster.
5. Tugaskan lagi setiap obyek memakai pusat klaster yang baru. Jika pusat klaster sudah tidak berubah lagi, maka proses prengklasteran selesai.
6. Kembali ke langkah 3 pusat klaster tidak berubah lagi.

3.4.

3.5. Visual Basic

Menurut andi (2002) Microsoft visual basic 6.0 adalah bahasa pemrograman yang digunakan untuk membuat aplikasi windows yang berbasis grafis (GUI-Grapical User Interface). Microsoft visual basic merupakan event-driven programming (pemrograman terkendali kejadian) artinya program nunggu sampai adanya respon dari pemakai berupa event atau kejadian tertentu (tombol diklik, menu dipilih, dan lain-lain)

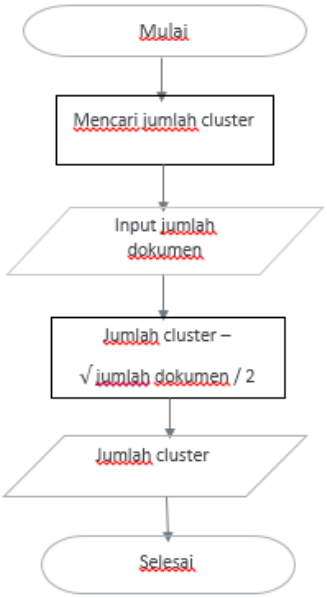
IV. PERANCANGAN ALGORITMA DAN PROGRAM

3.1 Flowchart

Flowchart adalah sebuah diagram dengan symbol-simbol grafis yang menyatakan aliran algoritma atau proses yang menampilkan langkah-langkah di simbolkan dalam bentuk kotak, beserta urutannya dengan menghubungkan masing-masing langkah tersebut mengguanakan tanda panah.

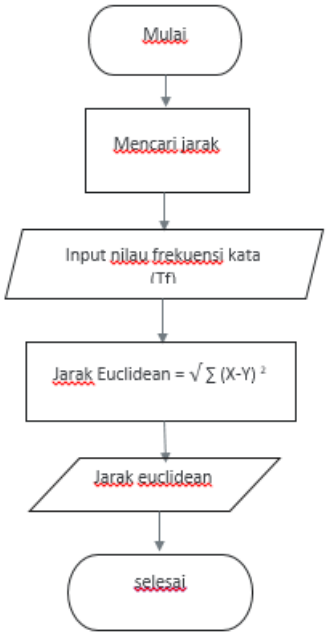
3.1.1 Flowchart mencari jumlah cluster

Flowchart mencari jumlah cluster merupakan flowchart yang berisi proses pencarian jumlah cluster dengan cara membagi dua jumlah dari seluruh document kemudian diakarkan.



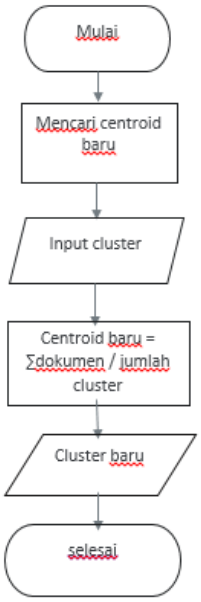
3.1.2 Flowchart mencari jarak

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 tiap dokumen, kemudian dilakukan perhitungan jarak dengan Euclidean.



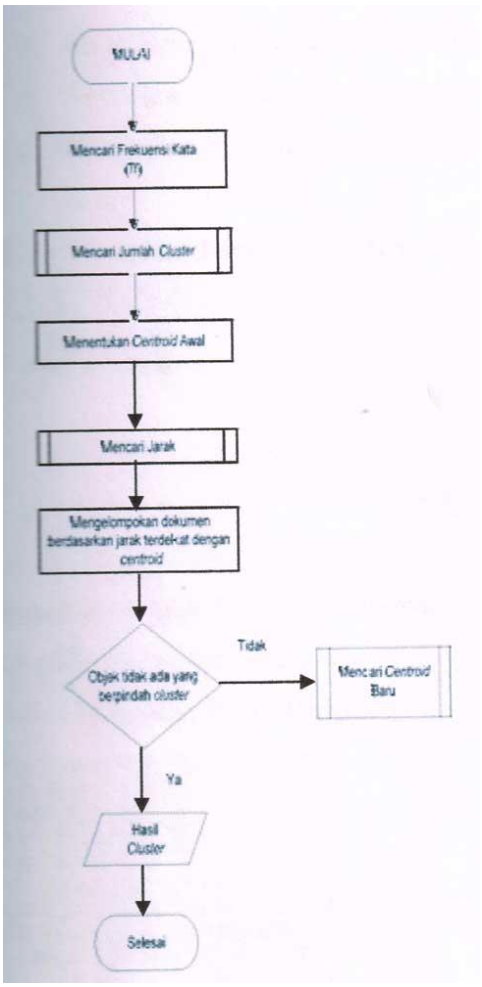
3.1.3 Flowchart mencari centroid baru

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.



3.1.4 Flowchart Algoritma K-Means

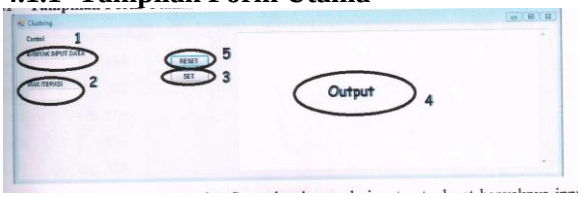
Flowchart Algoritma K-Means merupakan flowchart yang berisi urutan proses darii mencari frekuensi kemunculan kata (Tf), mencari jumlah cluster, menentukan centroid (titik pusat) awal, mencari jarak, mengelompokkan dokumen berdasarkan jarak terdekat dengan centroid serta proses mencari centroid baru.



IMPLEMENTASI DAN ANALISIS PROGRAM

4.1 Prosedur Uji Coba Program

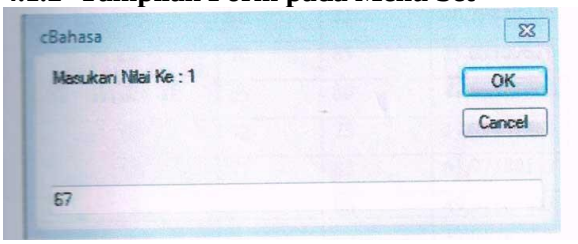
4.1.1 Tampilan Form Utama



Pada form terdapat menu reset dan set sedangkan pada inputan terdapat banyaknya input data daan max iterasi dan text view untuk output. berikut kegunaan dari penjabaran diatas :

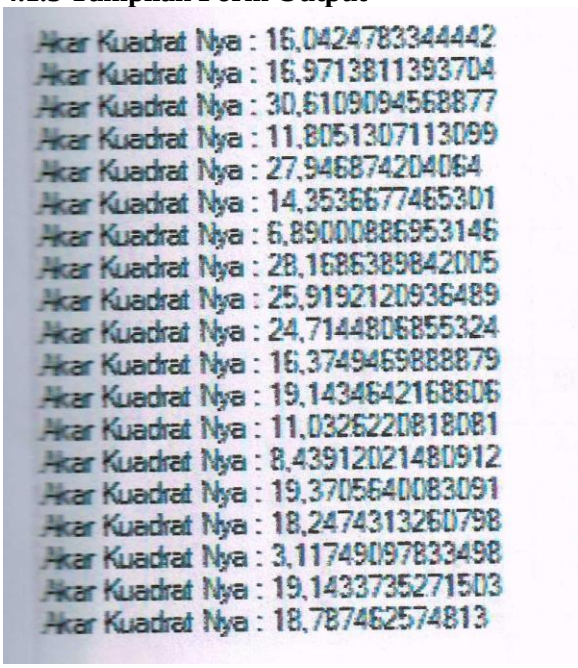
1. TextBox banyaknya input data : banyaknya data nilai siswa yang akan di inputkan
2. Textbox Max iterasi : Maksimal iterasi
3. Menu Set : Untuk memulai perhitungan
4. TextView : untuk menampilkan hasil output perhiyungan
5. Menu Reset : Untuk membersihkan semua input maupun output kembali ke form utama

4.1.2 Tampilan Form pada Menu Set



Memasukan nilai Mandarin, German dan Prancis dari masing-masing siswa sebanyak input data yang telah ditentukan.

4.1.3 Tampilan Form Output



4.2 Analisis Hasil Pengujian Program

4.2.1 Tabel Interasi Ke-1

				cl1			**cl1**		
Siswa	mandarin	german	prancis	45	50	55	40	55	65
a	50	60	70	18,70828693			12,24744871		
b	65	80	73	40,29888336			36,24913792		
c	72	70	65	35,05709629			35,34119409		
…	…	…	…	…			…		
T	77	71	55	38,27531842			41,53311931		

4.2.2 Tabel Clustering Ke-1

cl1	cl1
	OK
	OK
OK	
…	…
OK	

4.2.4 Tabel Clustering Ke-2

siswa	cl1			cl2		
	0	0	0	73,4	71,8	74,2
a	104,8808848			26,54128859		
b	126,3091446			11,8		
c	119,6202324			9,478396489		
d	132,34047			13,11640195		
e	116,8460526			34,94338278		
f	145,9143584			24,1627813		
g	135,2479205			25,68345771		
h	140,406,5526			18,15599075		
i	121,7127767			19,8		
j	126,9094165			23,10064934		
k	140,605832			32,91261156		
l	125,1279345			17,46538436		
m	123,7982229			6,988562084		
n	138,4666025			15,3961034		
o	122,3560379			13,6762568		
p	133,37541			26,26480535		
q	124,9959999			24,14207945		
r	136,1249426			17,15925406		
s	127,0905189			5,407402334		
t	118,3004649			19,55095906		

4.2.5 Tabel Centroid Baru Iteras Ke-3

Dari data table diatas diketahui bahwa pada centroid iterasi ke-2 dan ke-3 tidak ada perubahan.

	CENTROID 1			CENTROID 2		
	0	0	0	73,4	71,8	74,2

V. PENUTUP

5.1 Kesimpulan

Telah dibuat program penentuan nilai ujian dengan klastering menggunakan metode *K-means.* Program dapat dikembangkan menjadi lebih komunikatif tidak hanya nilai jurusan bahasa mandarin, german, prancis saja tetapi dapat ditambahkan komponen nilai yang lain. Dari hasil pembuatan aplikasi dapat disimpulkan hal-hal sebagai berikut ;

1. Program aplikasi menghasilkan pola dari prestasi mahasiswa yang klusternya tetap, turun dan naik. Pola mahasiswa tersebut dapat terlihat dari data nilai tes jurusan bahasa mandarin, german dan prancis
2. *Data cleaning*, proses menghapus data yang tidak konsisten dan kotor
3. *Data integration*, penggabungan beberapa sumber data
4. *Data selection*, pengambilan data yang akan dipakai dari sumber data
5. *Data transformation*, proses dimana data ditransformasikan menjadi bentuk yang sesuai untuk dioroses dalam data mining
6. *Data mining*, suatu proses yang penting dengan melibatkan metode untuk menghasilkan suatu pola data
7. *Pattern evaluation*, proses untuk menguji kebenaran dari ppola data yang mewakili knowledge yang ada didalam data itu sendiri
8. *Knowledge presentation*, proses visualisasi dan teknik menyajikan knowledge digunakan untuk menampilkan knowledge hasil mining kepada user

DAFTAR PUSTAKA

Bertalya. 2009. Konsep Data Mining, Klasifikasi: Pohon Keputusan. Jakarta: Universitas Gunadarma.

Bramer, Max. 2007. Principles of Data Mining. London: Springer.

Cheung, Yiu-Ming. 2003. A New Generalized K-Means Clustering Algorithm. Pattern Recognition Letters 24. Hongkong: Elsevier B. V.

Dwi, A.S, Dimas. 2013. Model Prediksi Tingkat Kelulusan Mahasiswa dengan Teknik Data Mining Menggunakan Metode Decision Tree C4.5. Skripsi. Yogyakarta: Universitas Negeri Yogyakarta.

Dzacko, Haidar. 2007. Basis Data (Database). Indonesia:Mangosoft.

Frank, Eibe, et al. 2004. The WEKA Data Mining Software: An Update. Department of Computer Science. New Zealand: University of WaikatoHamilton.