**MAKALAH**

**KLUSTERING PADA TWITTER MENGGUNAKAN BAHASA PEMROGRAMAN R**



**Disusun Oleh :**

**Muhamad Ali Basri (14.01.53.0044)**

**UNIVERSITAS STIKUBANK SEMARANG (UNISBANK)**

**FALKUTAS TEKNOLOGI INFORMASI**

**TEKNIK INFORMATIKA**

**2017**

**BAB I**

**PENDAHULUAN**

1. **Latar belakang**

Clustering adalah metode penganalisaan data, yang sering dimasukkan sebagai salah satu metode *Data Mining*, yang tujuannya adalah untuk mengelompokkan data dengan karakteristik yang sama ke suatu ‘wilayah’ yang sama dan data dengan karakteristik yang berbeda ke ‘wilayah’ yang lain. Ada beberapa pendekatan yang digunakan dalam mengembangkan metode clustering. Dua pendekatan utama adalah clustering dengan pendekatan partisi dan clustering dengan pendekatan hirarki. Clustering dengan pendekatan partisi atau sering disebut dengan *partition-based clustering* mengelompokkan data dengan memilah-milah data yang dianalisa ke dalam cluster-cluster yang ada. Clustering dengan pendekatan hirarki atau sering disebut dengan *hierarchical clustering*mengelompokkan data dengan membuat suatu hirarki berupa dendogram dimana data yang mirip akan ditempatkan pada hirarki yang berdekatan dan yang tidak pada hirarki yang berjauhan. Di samping kedua pendekatan tersebut, ada juga clustering dengan pendekatan *automatic mapping*(Self-Organising Map/SOM).

Twitter merupakan situs jejaring sosial yang keberadaannya masih diminiati oleh masyarkat sampai saat ini. Twitter adalah jejaring sosial berupa blog ukuran kecil yang didirikan oleh Jack Dorsey pada bulan Maret 2006. Melalui Twitter pengguna dapat mengirim dan membaca pesan, berbagi informasi, menjalin relasi bisnis, menuangkan isi hati dan pikiran dalam bentuk tulisan (sering disebut tweet), dengan kapasitas kata yang bisa diunggah dan ditampilkan pada timeline penggunatwitter mencapai 140 karakter. Sama halnya dengan situs jejaring sosial lain dalam Twitter disediakan suatu mesin pencarian (search engine) yang berguna untuk mempermudah pengguna dalam menemukan informasi menggunakan kata kunci. Melalui search engine pengguna dapat menemukan lebih banyak informasi yang dibutuhkan terkait topik yang ingin dicari, yaitu lebih dari satu akun yang ada di twitter.Twitter sebagai hasil dari perkembangan teknologi informasi memungkinkan setiap waktu untuk menghasilkan kumpulan data yang banyak, dimana setiap detiknya pada saat kehidupan normal rata-rata jumlah tweet yang ada dalam twitter adalah 600 tweet. Dengan adanya kumpulan data yang terus meningkat setiap waktunya yaitu berupa data tweet perlu dilakukan suatu penangan menggunakan metode khusus untuk mengalisis data pada twitter sehingga menghasilkan suatu informasi yang bermanfaat dan mengurangi kondisi yang biasa disebut “rich of data but poor of information”.

1. **Rumusan Masalah**

Berdasarkan latar belakang di atas maka rumusan masalahnya adalah bagaimana penjelasan mengenai Klustering K-means,PAM(Partitioning AroundMedoids),Hirarki Agglomerative dan implementasinya menggunakan bahasa R.

1. **Tujuan Penulisan**

Adapun tujuan dari penulisan ini adalah untuk mengklasifikasikan Klustering K-means,PAM(Partitioning AroundMedoids),Hirarki Agglomerative pada sebuah twitter menggunakan bahasa R.

**BAB II**

**PEMBAHASAN**

1. **Algoritma K-means**

K-means merupakan salah satu algoritma clustering .Tujuan algoritma ini yaitu untuk membagi data menjadi beberapa kelompok. Algoritma ini menerima masukan berupa data tanpa label kelas. Hal ini berbeda dengan supervised learning yang menerima masukan berupa vektor (­x­1 , y1) , (­x­2 , y2) , …, (­x­i , yi), di mana xi merupakan data dari suatu data pelatihan dan yi merupakan label kelas untuk xi.

Pada algoritma pembelajaran ini, komputer mengelompokkan sendiri data-data yang menjadi masukannya tanpa mengetahui terlebih dulu target kelasnya. Pembelajaran ini termasuk dalam unsupervised learning. Masukan yang diterima adalah data atau objek dan k buah kelompok (cluster) yang diinginkan. Algoritma ini akan mengelompokkan data atau objek ke dalam k buah kelompok tersebut. Pada setiap cluster terdapat titik pusat (centroid) yang merepresentasikan cluster tersebut.

Algoritma untuk melakukan K-Means clustering adalah sebagai berikut:

* Pilih K buah titik centroid secara acak
* Kelompokkan data sehingga terbentuk K buah cluster dengan titik centroid dari
* setiap cluster merupakan titik centroid yang telah dipilih sebelumnya

Perbaharui nilai titik centroid

* Ulangi langkah 2 dan 3 sampai nilai dari titik centroid tidak lagi berubah

Proses pengelompokkan data ke dalam suatu cluster dapat dilakukan dengan cara menghitung jarak terdekat dari suatu data ke sebuah titik centroid. Perhitungan jarak Minkowski dapat digunakan untuk menghitung jarak antar 2 buah data. Rumus untuk menghitung jarak tersebut adalah:



Di mana:

g = 1, untuk menghitung jarak Manhattan

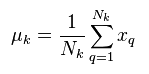
g = 2, untuk menghitung jarak Euclidean

g = ∞, untuk menghitung jarak Chebychev

xi , xj adalah dua buah data yang akan dihitung jaraknya

p = dimensi dari sebuah data

Pembaharuan suatu titik centroid dapat dilakukan dengan rumus berikut:



Di mana:

µk = titik centroid dari cluster ke-K

Nk = banyaknya data pada cluster ke-K

xq = data ke-q pada cluster ke-K

* **Kekurangan dan Kelebihan K-Means**

Ada beberapa kelebihan pada algoritma k-means, yaitu:

1. Mudah untuk diimplementasikan dan dijalankan.
2. Waktu yang dibutuhkan untuk menjalankan pembelajaran ini relatif cepat.
3. Mudah untuk diadaptasi.
4. Umum digunakan.

Algoritma k-means memiliki beberapa kelebihan, namun ada kekurangannya juga. Kekurangan dari algoritma tersebut yaitu :

1. Sebelum algoritma dijalankan, k buah titik diinisialisasi secara random sehingga pengelompokkan data yang dihasilkan dapat berbeda-beda. Jika nilai random untuk inisialisasi kurang baik, maka pengelompokkan yang dihasilkan pun menjadi kurang optimal.
2. Dapat terjebak dalam masalah yang disebut curse of dimensionality. Hal ini dapat terjadi jika data pelatihan memiliki dimensi yang sangat tinggi (Contoh jika data pelatihan terdiri dari 2 atribut maka dimensinya adalah 2 dimensi. Namun jika ada 20 atribut, maka akan ada 20 dimensi). Salah satu cara kerja algoritma ini adalah mencari jarak terdekat antara k buah titik dengan titik lainnya. Jika mencari jarak antar titik pada 2 dimensi, masih mudah dilakukan. Namun bagaimana mencari jarak antar titik jika terdapat 20 dimensi. Hal ini akan menjadi sulit.
3. Jika hanya terdapat beberapa titik sampel data, maka cukup mudah untuk menghitung dan mencari titik terdekat dengan k titik yang diinisialisasi secara random. Namun jika terdapat banyak sekali titik data (misalnya satu milyar buah data), maka perhitungan dan pencarian titik terdekat akan membutuhkan waktu yang lama. Proses tersebut dapat dipercepat, namun dibutuhkan struktur data yang lebih rumit seperti kD-Tree atau hashing.
4. **Algoritma K-medoids/ PAM(Partitioning Around Medoids)**

Algoritma Partitioning Around Medoids (PAM) atau dikenal juga dengan K-Medoids adalah algoritma pengelompokan yang berkaitan dengan algoritma K-means dan algoritma medoidshift. Algoritma K-Medoids ini diusulkan pada tahun 1987.

Algorima Partitioning Around Medoids (PAM) dikembangkan oleh Leonard Kaufman dan Peter J. Reusseeuw. Algoritma ini sangat mirip dengan algoritma K-Means, terutama karena kedua algoritma ini partitional. Dengan kata lain, kedua algoritma ini memecah dataset menjadi kelompok-kelompok dan kedua algoritma ini berusaha untuk meminimalkan keslahan. Tetapi algoritma Algorima Partitioning Around Medoids (PAM) bekerja dengan menggunakan Medoids, yang merupakan entitas dari dataset yang mewakili kelompok dimana ia dimasukkan.

Algorima Partitioning Around Medoids (PAM) menggunakan metode partisi clustering untuk mengelompokkan sekumpulan n objek menjadi sejumlah cluster. Algoritma ini menggunakan objek pada kumpulan objek untuk mewakili sebuah cluster. Objek yang terpilih untuk mewakili sebuah cluster disebut dengan medoid. Cluster dibangun dengan menghitung kedekatan yang dimiliki antara medoids dengan objek non-medoids.

Algoritma dari Algorima Partitioning Around Medoids (PAM) atau K-Medoids adalah sebagai berikut (Han & Kamber, 2006) :

1. Secara acak pilih k objek pada sekumpulan n objek sebagi medoid.
2. Ulangi langkah 3 hingga langkah 6.
3. Tempatkan objek non-medoid ke dalam cluster yang paling dekat dengan medoid.
4. Secara acak pilih Orandom sebuah objek non-medoid.
5. Hitung total biaya, S, dari pertukaran medoid o, dengan Orandom.
6. Jika S<0 maka tukar o, dengan Orandom untuk membentuk sekumpulan k objek baru sebagai medoid.
7. Hingga tidak ada perubahan.

Nilai total biaya/cost dinyatakan dengan persamaan :

Total cost =

Dimana :

Dist = merujuk pada rumus -

Nilai S dinyatakan dengan persamaan :

S = Total cost baru – total cost lama

Dimana :

Total cost baru = jumlah biaya / cost non medoid

Total cost lama = jumlah biaya / cost medoid

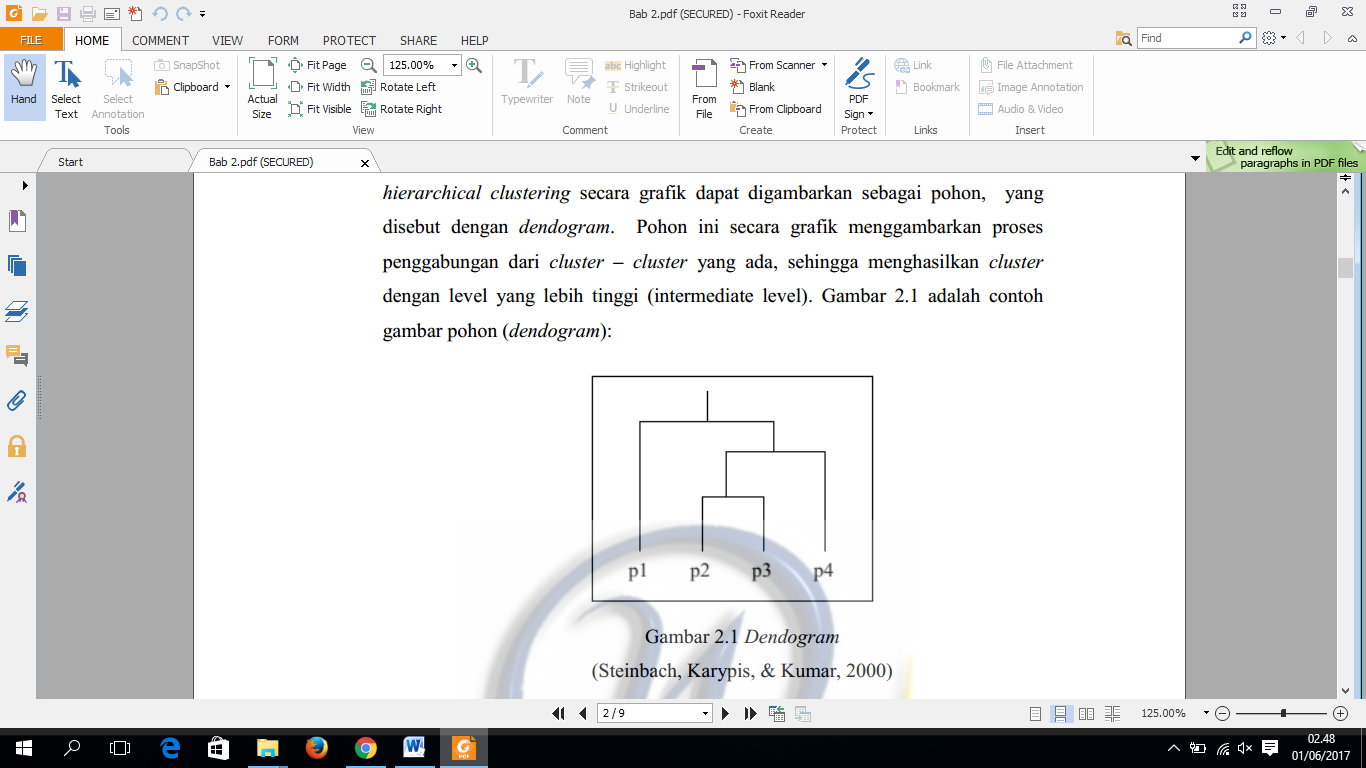
K-Medoid sangat mirip dengan K-Means, perbedaan utama diantara dua algoritma tersebut adalah jika pada K-Means cluster diwakili dengan pusat dari cluster, sedangkan pada K-Medoid cluster diwakili objek terdekat dari pusat cluster.

1. **Hirarki Agglomerative**

Hierarchical Agglomerative Clustering (HAC) adalah suatu metode clustering yang bersifat bottom-up yaitu menggabungkan sebuah cluster menjadi satu cluster tunggal. Metode ini dimulai dengan meletakkan setiap obyek data sebagai sebuah cluster tersendiri dan selanjutnya menggabungkan cluster-cluster tersebut menjadi cluster yang lebih besar dan lebih besar lagi sampai akhirnya semua objek data menyatu dalam sebuah cluster tunggal. Secara logika semua obyek pada akhirnya hanya akan membentuk sebuah cluster . Metode ini dimulai membentuk cluster pada setiap obyek. Kemudian dua obyek dengan memiliki jarak terdekat digabungkan. Selanjutnya obyek ketiga digabung dengan obyek lain memiliki jarak terdekat dan membentuk cluster baru. Proses akan berlanjut hingga akhirnya terbentuk satu cluster yang terdiri dari keseluruhan obyek.

Hierarchical clustering merupakan salah satu algoritma clustering yang fungsinya dapat digunakan untuk mengcluster dokumen (dokumen clustering). Dari teknik Hierarchical clustering, dapat dihasilkan suatu kumpulan partisi yang berurutan, dimana dalam kumpulan tersebut terdapat cluster-cluster yang mempunyai poin-poin individu, cluster-cluster ini berada di level yang paling bawah. Selain itu ada juga cluster yang didalamnya terdapat poin-poin yang dipunyai semua cluster didalamnya, cluster ini disebut single cluster, terletak di level yang paling atas.

Dalam algoritma Hierarchical clustering, cluster yang berada dilevel lebih atas (intermediate level) dari cluster yang lain, dapat, diperoleh dengan cara mengkombinasikan dua buah cluster yang berada pada level dibawahnya (Steinbach, Karypis, & Kumar, 2000). Hasil keseluruhan dari algoritma Hierarchical clustering secara grafik dapat digambarkan sebagai pohon, yang disebut dengan dendogram. Pohon ini secara grafik menggambarkan proses penggabungan dari cluster-cluster yang ada, sehingga menghasilkan cluster dengan level yang lebih tinggi (intermodiate level).



Ada dua metode yang sering diterapkan yaitu agglomerative hierarchical clustering dan divisive hierarchical clustering. Pada agglomerative hierarchical, proses hierarchical clustering dimulai dari cluster-cluster yang mempunyai poin-poin individu yang berada di level paling bawah. Pada setiap langkahnya dilakukan penggabungan sebuah cluster dengan cluster lainnya, dimana cluster-cluster yang digabungkan berada saling berdekatan atau mempunyai tingkat / sifat kesamaan yang paling tinggi (Steinbach, Karypis, & Kumar, 2000).

Agglomerative hierarchical clustering bekerja dengan sederetan dari penggabungan yang berurutan atau sederetan dari pembagian yang berurutan dan berawal dari objek-objek yang individual. Jadi, pada awalnya banyaknya cluster sama dengan banyaknya objek. Objek-objek yang paling mirip dikelompokkan, dan kelompok-kelompok awal ini digabungkan sesuai dengan kemiripannya. Sewaktu kemiripan berkurang, semua sub kelompok digabungkan menjadi satu cluster tunggal. Hasil-hasil dari clustering dapat disajikan secara grafik dalam bentuk dendogram. Cabang-cabang dalam pohon menyajikan cluster dan bergabung pada node yang posisinya sepanjang sumbu jarak (similaritas) menyatakan tingkat dimana penggabungan terjadi.

Langkah-langkah dalam algoritma clustering hirarki agglomerative untuk mengelompokkan N objek (item/variabel) :

1. Mulai dengan N cluster, setiap cluster mengandung entiti tunggal dan sebuah matriks simetrik dari jarak (similarities) D = {dix} dengan tipe NxN.
2. Cari matrix jarak untuk pasangan cluster yang terdekat (paling mirip). Misalkan jarak antara cluster U dan V yang paling mirip adalah duv.
3. Gabungkan cluster U dan V. Label cluster yang baru dibentuk dengan (UV). Update entries pada matrik jarak dengan cara :
4. Hapus baris dan kolom yang memberikna yang memberikan jarak-jarak antara cluster U dan V
5. Tambahkan baris dan kolom yang meberikan jarak-jarak antara cluster (UV) dan cluster-cluster yang tersisa.
6. Ulangi langkah 2 dan 3 sebanyak (N-1) kali. (semua objek akan berada dalam cluster tunggal setelah algoritma berakhir). Catat identitas dari cluster yang digabungkan dan tingkat-tingkat (jarak atau similaritas) dimana penggabungan terjadi.

Beberapa metode hierarchical clustering yang sering digunakan dibedakan menurut cara mereka untuk menghitung tingkat kemiripan atau jarak antar kelompok. Ada yang menggunakan ward’s linnkage, centroid linkage, single linkage, complete linkage, average linkage, median linkage, dan lain-lainnya.

**BAB III**

**PENUTUP**

* **Kesimpulan**

Klustering adalah mengelompokkan variabel ke dalam kelompok berdasarkan kesamaan tertentu.Sedangkan R adalah suatu perangkat lunak yang digunakan untuk manipulasi data, perhitungan, simulasi, penayangan grafik, dan sekaligus sebagai bahasa pemograman yang bersifat interpreter.R adalah bahasa pemograman yang bersifat object oriented, sehingga seluruh variable, data, fungsi, dan lain sebagainya disimpan dalam memori komputer sebagai objek.Program merupakan serangkaian instruksi yang memberitahu komputer sesuatu yang harus dilakukan (Zelle, 2002). Pemrograman dengan bahasa R dapat mengklasifikasikan Klustering K-means,PAM(Partitioning AroundMedoids),Hirarki Agglomerative pada sebuah twitter menggunakan bahasa R.