**MAKALAH KLASTERING TEKS**

**DENGAN MENGGUNAKAN BAHASA R**

**Disusun Oleh :**

**Andre Achmad (15.01.55.0055)**

**UNIVERSITAS STIKUBANK SEMARANG (UNISBANK)**

**FAKULTAS TEKNOLOGI INFORMASI**

**SISTEM INFORMASI**

**2017**

1. **PENDAHULUAN**

Clustering merupakan suatu metode analisa data untuk memecahkan masalah pengelompokan data. Salah satu metode clustering adalah K-Means. K-Means mempunyai kemampuan mengelompokkan data dalam jumlah yang cukup besar dengan komputasi yang relative cepat dan efisien. Akan tetapi, hasil clustering dengan K-Means sangat bergantung pada pusat awal cluster. Hasil clustering dengan metode K-Means baik jika penentuan pusat cluster tepat. Metode Hirerachical Agglomerative dapat digunakan untuk mengatasi masalah penentuan pusat cluster pada K-Means. Penelitian ini mengkombinasikan K-Means dengan Hirerachical Agglomerative. Hasil dari Hirerachical Agglomerative akan digunakan dalam penentuan pusat awal cluster K-Means clustering. Kombinasi antara metode Hirerachical Agglomerative dan K-Means.

1. **LANDASAN TEORI**
   1. **Text mining**

Text mining adalah proses menemukan hal baru, yang sebelumnya tidak diketahui, mengenai informasi yang berpotensi untuk diambil manfaatnya dari sumber data yang tidak terstruktur mencakup dokumen bisnis, komentar customer, halaman web dan file XML.

Text mining hampir sama dengan data mining dalam hal tujuan dan proses, tapi pada text mining inputnya adalah file data tidak terstruktur seperti dokumen dalam bentuk word, PDF, text, XML dan sebagainya . Text mining dapat digunakan dalam beberapa hal yaitu ekstraksi informasi, topic tracking, summarization, kategorisasi dan clustering.

* 1. **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.

Algoritma Partitioning Around Medoids (PAM) dikembangkan oleh Leonard Kaufman dan Peter J. Rousseeuw. 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 kesalahan. Tetapi algoritma Partitioning Around Medoids (PAM) bekerja dengan menggunakan Medoids, yang merupakan entitas dari dataset yang mewakili kelompok dimana ia dimasukkan.

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

* + 1. **Algoritma Partitioning Around Medoids (PAM)**

Algoritma dari Partitioning Around Medoids (PAM) atau K-Medoids adalah sebagai berikut :

1. Secara acak pilih k obyek pada sekumpulan n obyek sebagai medoid.

2. Ulangi langkah 3 hingga langkah 6

3. Tempatkan obyek non – medoid ke dalam cluster yang paling dekat dengan medoid.

4. Secara acak pilih Orandom. Sebuah obyek non-medoid.

5. Hitung total biaya, S, dari pertukaran medoid Oj dengan Orandom.

6. Jika S < 0 maka tukar oj dengan orandom untuk membentuk sekumpulan k obyek baru sebagai medoid.

7. Hingga tidak ada perubahan

Nilai total biaya / cost dinyatakan dengan persamaan



Nilai S dinyatakan dengan persamaan :



Dimana :

Total cost baru = jumlah biaya/cost non-medoids

Total cost baru = jumlah biaya/cost medoids.

K-medoids sangat mirip dengan k-means, perbedaan utama diantara dua algoritma tersebut adalah jika pada k-means cluster diwakili oleh pusat dari cluster. Sedangkan k –medoids cluster diwakili oleh obyek terdekat dari pusat cluster.

* 1. **Clustering**

Clustering adalah proses mengelompokkan atau penggolongan objek berdasarkan informasi yang diperoleh dari data yang menjelaskan hubungan antar objek dengan prinsip untuk memaksimalkan kesamaan antar anggota satu kelas dan meminimumkan kesamaan antar kelas/cluster. Clustering membagi data ke dalam grup-grup yang mempunyai obyek yang karakteristiknya sama.

* + 1. **Hierarchical clustering**

Hierarchical clustering adalah metode analisis kelompok yang berusaha untuk membangun sebuah hierarki kelompok. Hierarchical clustering dibagi menjadi dua yaitu Agglomeratif Clustering dan Difisive Clustering. Agglomeratif Clustering mengelompokkan data dengan pendekatan bawah atas (bottom up), sedangkan Difisif Clustering menggunakan pendekatan atas bawah (top-bottom).

Metode hierarchical agglomeratif clustering, mengasumsikan setiap data yang ada sebagai cluster di awal proses. Jika jumlah data adalah n, dan jumlah cluster adalah k, maka besarnya n = k. Kemudian dihitung jarak antar clusternya dengan menggunakan Euclidean distance berdasarkan jarak rata-rata antar objek. Selanjutnya, dari hasil perhitungan jarak dipilih jarak yang paling minimal dan digabungkan sehingga besarnya n = n -1. Ketika dua cluster digabungkan, jarak antara dua cluster yang digabungkan dengan cluster yang lain di-update. Penggabungan cluster akan terus dilakukan dan akan berhenti jika memenuhi kondisi jumlah k = 1. Pada akhir tahap hierarchical clustering diperoleh dendrogram yang menunjukkan urutan pengelompokan masing-masing anggota dalam cluster. Penelitian ini menggunakan metode ward sebagai metode update jarak. Metode Ward dapat membentuk cluster berdasarkan jumlah total kuadrat deviasi tiap pengamatan dari rata-rata cluster yang menjadi anggotanya. Metode Ward berusaha untuk meminimalkan variasi antar objek dalam satu cluster dan memaksimalkan variasi dengan objek yang ada di cluster lainnya. Jarak antara dua cluster yang terbentuk pada metode Ward adalah sum of squares diantara dua cluster tersebut. Metode Ward didasarkan pada kriteria sum square error (SSE) dengan ukuran kehomogenan antara dua objek berdasarkan jumlah kuadrat kesalahan minimal. Perhitungan pada metode ward menggunakan rumus berikut :



Dengan u dan v cluster yang digabung, w cluster lain yang dicari jaraknya dengan cluster gabungan uv, 𝐼 𝑢𝑣 𝑤 jarak antara cluster uv dan cluster w, 𝐼𝑢𝑤 jarak antara cluster u dan cluster w, 𝐼𝑣𝑤 jarak antara cluster v dan cluster w, 𝐼𝑢𝑣 jarak antara cluster u dan cluster v, 𝑛𝑢 , 𝑛𝑣 , 𝑛𝑤 dan adalah banyaknya objek pada cluster ke-u, ke-v dan ke-w.

* + 1. **K-Means clustering**

K-Means adalah suatu metode penganalisaan data atau metode data mining yang melakukan proses pemodelan tanpa supervisi (*unsupervised*) dan merupakan salah satu metode yang melakukan pengelompokan data dengan sistem partisi. Metode ini berusaha untuk meminimalkan variasi antar data yang ada di dalam suatu *cluster* dan memaksimalkan variasi dengan data yang ada di *cluster* lainnya.

Berikut adalah langkah-langkah dari algoritma K-Means:

1. Menentukan banyak k-cluster yang ingin dibentuk.
2. Membangkitkan nilai random untuk pusat cluster awal (centroid) sebanyak k-cluster.
3. Menghitung jarak setiap data input terhadap masingmasing centroid menggunakan rumus jarak Eucledian (Eucledian Distance) hingga ditemukan jarak yang paling dekat dari setiap data dengan centroid. Berikut adalah persamaan Eucledian Distance:



dengan 𝑑 𝑥𝑖 , 𝜇𝑖 adalah jarak antara cluster 𝑥 dengan pusat cluster 𝜇 pada kata ke i , 𝑥𝑖 adalah bobot kata ke i pada cluster yang ingin dicari jaraknya, 𝜇𝑖 bobot kata ke i pada pusat cluster.

1. Mengklasifikasikan setiap data berdasarkan kedekatannya dengan centroid (jarak terkecil).
2. Mengupdate nilai centroid. Nilai centroid baru diperoleh dari rata-rata cluster yang bersangkutan dengan menggunakan rumus:



Dimana 𝑛𝑘 = jumlah data dalam cluster

𝑑𝑖 = jumlah dari nilai jarak yang masuk dalam masingmasing cluster

1. Melakukan perulangan dari langkah 2 hingga 5 hingga anggota tiap cluster tidak ada yang berubah.
2. Jika langkah 6 telah terpenuhi, maka nilai rata-rata pusat cluster (μj) pada iterasi terakhir akan digunakan sebagai parameter untuk menentukan klasifikasi data.