IMPLEMENTASI METODE HIERARCHICAL CLUSTERING PADA OZONE LEVEL DETECTION DATA SET

Muhammad Yamin[1], Indriati, ST., M.Kom.[2], Candra Dewi, S.Kom., MSc.[3]
Program Studi Ilmu Komputer, Fakultas Program Teknologi Informatika dan Ilmu Komputer
Universitas Brawijaya Malang
Jalan Veteran No.8 Malang 65145, Indonesia
Email:yaammiinn@gmail.com[1], indriati.tif@ub.ac.id[2], d3w1_c4ndr4@yahoo.com[3].

Abstrak

Clustering merupakan 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. Salah satu algoritma clustering adalah Qualitative Hierarchical Clustering. Algoritma tersebut adalah metode analisis cluster yang membangun hirarki dari cluster. Jarak antara tiap obyek merupakan input dari algoritma ini. Iterasi terus berlanjut sampai semua obyek telah dicluster menjadi sebuah cluster saja. Dengan menggunakan sebuah metode Hierarchical Clustering untuk mengukur jarak antara data time series pada ozone level detection dataset yang telah ditentukan. Algoritma Hierarchical dimulai dengan menjadikan tiap objek menjadi sebuah cluster dan secara iterasi menggabungkan tiap cluster yang mirip. Jarak antar objek merupakan input dari algoritma ini. Iterasi terus berlanjut sampai semua objek telah di cluster hingga menjadi cluster saja. Sistem perangkat lunak yang dibangun dengan algoritma pada Data time series, Hierarchical Clustering, dan metode Qualitative dapat melakukan proses clustering pada data time series. Tingkat akurasi dari metode hierarchical clustering dapat dilihat dari nilai rata-rata akurasi f-measure yang di dapat. Akurasi dari masing-masing percobaan yang memiliki nilai f-measure paling tinggi terdapat pada scenario ke dua dimana nilai f-measure yang di dapat adalah 0.9792.

Kata Kunci: *Data Mining, Clustering, Hierarchical Clustering, Time Series, Short Time Series.*

1. Pendahuluan

Clustering merupakan 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*. Salah satu algoritma *clustering* adalah *Qualitative Hierarchical Clustering*. Algoritma tersebut adalah metode analisis *cluster* yang membangun hirarki dari kluster. Jarak antara tiap obyek merupakan input dari algoritma ini. Iterasi terus berlanjut sampai semua obyek telah di*cluster* menjadi sebuah *cluster* saja. Output dari algoritma ini adalah sebuah *dendrogram* (*Hierarchical tree*). *Dendrogram* adalah sebuah *binary tree* di mana setiap *cluster* awalnya berisi satu elemen saja yaitu *leave* dari *tree* tersebut.

Salah satu penggunaan algoritma ini adalah *clustering dataset time series*. Dataset *time series* adalah sebuah kumpulan data dari pengamatan yang dibuat secara berurutan waktunya, misalnya total penjualan per bulan, panggilan telpon per hari, perubahan inventori per minggu dan lain sebagainya.

Pada penelitian ini menggunakan sebuah pendekatan *clustering* untuk menganalisa data *time series*. Dengan menggunakan sebuah metode *Hierarchical Clustering Average Linkage* untuk mengukur jarak antar data time series yang telah ditentukan. Algoritma *Hierarchical* dimulai dengan menjadikan tiap objek dengan menjadikan sebuah cluster dan secara iterasi menggabungkan tiap cluster yang mirip. Jarak antar objek merupakan input dari algoritma ini. Iterasi terus berlanjut sampai semua objek telah dicluster hingga menjadi cluster saja.

2. Metode Penelitian

2.1 Data Mining

Data mining adalah suatu proses menggali (mining) pengetahuan dari sejumlah data yang besar, data mining juga disebut dengan *knowledge mining from databases*, *knowledge extraction*, data/*pattren analysis*, data *archeology*, dan data *design*. Data mining juga merupakan suatu proses *ekstraksi* informasi yang potensial, implisit, dan tidak diketahui sebelumnya (misalnya aturan-aturan, batasan-batasan, dan regularitas) dari sekumpulan data dalam *database* yang besar [4].

Data mining dapat dikatakan sebagai basil dari evolusi teknologi informasi. Yaitu mulai dari sistem data *collection*, *database creation*, data *management* (termasuk *storage* dan *retrieval* dan *databasetransaction processing*), dan data *analysis* dan *understanding*. Semakin berkembangnya kecepatan dalam mengumpulkan dan penyimpanan data dalam jumlah yang besar makin mengakibatkan semakin sulitnya data yang banyak tersebut untuk dianalisa oleh manusia tanpa bantuan alat yang dapat menggali informasi yang ada dalam data yang banyak tersebut [5].

2.2 *Clustering*

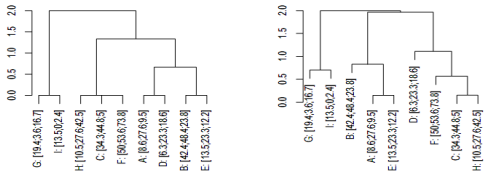
Clustering adalah proses mengelompokkan atau penggolongan objek berdasarkan informasi yang diperoleh dari data yang menjelaskan hubungan antar objek dengan prinsip untuk memaksimalkan kesamaan antar anggota satu kelas dan meminimumkan kesamaan antar kelas/cluster. *Clustering* dalam data mining berguna untuk menemukan pola distribusi di dalam sebuah dataset yang berguna untuk proses analisa data. Kesamaan objek biasanya diperoleh dari kedekatan nilai-nilai atribut yang menjelaskan objek-objek data, sedangkan objek-objek data biasanya direpresentasikan sebagai sebuah titik dalam ruang multi dimensi [1].

Clustering digunakan untuk mendapatkan *high availability* dan scalability. Pada *high available cluster*, dapat digunakan *fail over database cluster*, dimana hanya ada satu *node* yang aktif melayani *user*, sedangkan *node* lainnya *standby*. *Storage* yang digunakan mempunyai koneksi ke setiap *node* pada *cluster*, sehingga jika *primary node* mati, *database* engine, *listener process*, dan *logical host ip address* akan dijalankan pada *secondary node* tanpa perlu menunggu *operating system boot*, sehingga *downtime* dapat diminimalisasi. *Highavailability* mempunyai standar *duptime* 99.999 persen, atau hanya boleh mati selama 5 menit dalam setahun. Beberapa contoh *software* yang

dapat digunakan untuk membuat *HA cluster* adalah *Sun Cluster* dan *Veritas Cluster*. Pada *scalable cluster*, digunakan produk *Oracle RAC*, dimana setiap node aktif melayani user, sehingga diperoleh performa yang semakin baik dengan menggunakan lebih banyak node. *Sun cluster* dapat digunakan sampai 16 node, sedangkan *Veritas Storage Foundation for Oracle RA*C bias sampai 32 node. Jika ada *node* yang mati, tentu akan menurunkan performa, namun tidak terjadi *downtime*. Pada *scalable cluster*, seluruh *node* dapat terhubung secara langsung ke share di *storage*, namun dapat juga tidak mempunyai koneksi fisik ke *storage*, melainkan melalui *private cluster transport* [1].

2.3 Hierarchical Clustering

Algoritma *Hierarchical* dimulai dengan menjadikan tiap obyek menjadi sebuah *cluster* dan secara iterasi menggabungkan tiap *cluster* yang mirip. Jarak antara tiap obyek merupakan inputdari algoritma ini. Iterasi terus berlanjut sampai semua obyek telah di*cluster* menjadi sebuah *cluster* saja. Output dari algoritma ini adalah sebuah *dendrogram* (*Hierarchical tree*) [3]. *Dendrogram* adalah sebuah *binary tree* di mana setiap *cluster* awalnya berisi satu elemen saja yaitu *leave* dari *tree* tersebut. Setiap internal *node* mewakili *cluster* yang merupakan gabungan dari obyek dari dua *cluster* pada *node* anak. Tingkat dari *node* tersebut sesuai dengan jarak antara *cluster* yang digabung. *Dendrogram* memiliki dua keterbatasan, karena masing-masing pengamatan harus ditampilkan sebagai daun yang mereka hanya dapat digunakan untuk sejumlah kecil pengamatan. Stata 7 memungkinkan hingga 100 pengamatan, bahkan dengan 75 pengamatan sulit untuk membedakan individu daun. Sumbu vertikal mewakili kriteria tingkat di mana setiap dua *cluster* dapat bergabung. Berturut-turut bergabung dengan *cluster* menyiratkan struktur hirarkis, yang berarti bahwa *dendrogram* hanya cocok untuk analisis *cluster hirarki. Dendrogram* seperti ponsel dan dapat secara bebas diputar di setiap *node*. Dalam contoh di atas, *dendrogram* bisa berputar sedemikian rupa sehingga sampel muncul dalam urutan yang berbeda, ditunjukkan di bawah ini, dengan kendala yang *dendrogram* tidak salib itu sendiri. Seperti pemintalan *dendrogram* adalah cara yang berguna untuk menonjolkan pola *chaining* atau kekhasan *cluster* (meskipun tidak membantu dalam kasus ini). Untuk program analisis *cluster* banyak, berputar harus dilakukan (*tediously*) dalam sebuah program grafis yang terpisah, seperti *Illustrator*, tetapi dapat berputar dilakukan jauh lebih mudah langsung [2].



Gambar 1 Contoh sebuah dendrogram

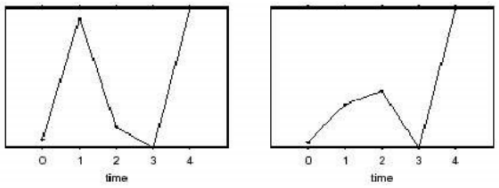
Metode Clustering terbagi menjadi dua yaitu *Partition Clustering* dan *Hierarchical Clustering*.

2.4 Data Time Series

Data *time* series adalah sebuah kumpulan dari pengamatan yang dibuat secara berurutan berdasarkan waktu, misalnya total penjualan per bulan, panggilan telepon per hari, perubahan inventori per minggu dan lain sebagainya. Jadi frekuensi dari *time* series tergantung dari masalah atau problem yang di hadapi. Satuan waktu pada data *time* series dapat berupa harian, mingguan, bulanan, perquater, tahunan dan satuan waktu lainnya. Pemilihan satuan waktu ini sangat penting untuk permodelan dalam pengambilan keputusan nantinya [4]. *Time series* melakukan pencarian terhadap *event-event* pada satuan waktu tertentu. Dengan memakai metode *time* series pada data *time* series maka akan membantu mengurangi jumlah informasi yang harus di analisa. Salah satu metode yang digunakan adalah *trend analysis*, yang merepresentasikan pembahan berdasarkan waktu. Ada empat komponen dalam karakteristik *time* series yaitu: (1) *long term or trend movements*, yang melakukan identifikasi dalam jangka waktu yang lama, di mana ditampilkan dalam bentuk kurva atau garis, (2) *cyclic movements or cyclic variations*, yang menampilkankurva seperti long term namun dapat dilakukan secara unperiodic, (3) *seasonal movements or seasonal variations*, berdasarkan kejadian tertentu, seperti pada saat Natal, (4) *irregular or random movements*, merupakan pengambilan kejadian secara random. Untuk menghasilkan analisis tentang trend yang terjadi dalam sistem lnformasi, digunakan pendekatan *cluster*ing berdasarkan analisis *qualitative* perhitungan jarak [2].

2.4 Perhitungan Jarak Untuk Short *Time* Series

Pada sub bab 2.3 telah dijelaskan tentang algoritma *Hierarchical*, yang menjadi pertanyaan adalah bagaimana menghitung jarak antar obyek dan jarak antar *cluster* objek. Perhitungan jarak dapat dilakukan dengan menggunakan *qualitative analysis* dan membandingkan antar suatu *time* series [2].



Gambar 2 *Time* Series dengan panjang 5

Pada Gambar 2.2, dimisalkan bahwa gambar yang scbclah kiri adalah *time* series X dan gambar sebelah kanan adalah *time* series Y dengan panjang 5. Apabila dipilih suatu titik i dan j, maka pcrubahan yang tcrjadi pada X dan Y yang dilambangkan dengan q (Xi,Xj) (dan q(Yi,Yj)) mempunyai 3 macam kcmungkinanyaitu kenaikan, ketika Xi> Xj; tidak bcrubah, ketika Xi = Xj; pcnurunan, ketika Xi< Xj. Scbagai contoh, antara titik l dan 3 pada X dan Y, keduanya mcngalamikcnaikan; antara titik 2 dan 3 pada X dan Y, X mengalami penurunan dan Y mengalami kenaikan. Untuk menghasilkan jarak tiap *time* series, digunakan persamaan Qualitative [2].

$$Dq(X,Y) = \frac{4}{N.(N-1)} . \sum_{i<j} \text{Diff}\,(q(Xi,Xj), q(Yi,Yj)) \quad (2.2)$$

Dimana persamaan (2.2) digunakan untuk melihat perubahan pada V, dan Diff merupakan fungsi untuk menentukan nilai dari 3 kemungkinan persamaan (2.2). Faktor 4/N (N-1) digunakan untuk melakukan normalisasi terhadap nilai jarak yang ada. Fungsi *Diff* merupakan hasil dari perbedaan titik tersebut [2].

3. Uji Coba

Untuk mengetahui bagaimana kinerja sistem rekomendasi, dibutuhkan *dataset* yang menentukan level *Ozone*. Dataset yang digunakan dalam penelitian ini diambil dari *UCI* (*www.archive.ics.UCI.edu.com*) yang diterbitkan pada tahun 1998. Pada skripsi ini terdapat statistik dataset yang akan digunakan pada penelitian, dimana jumlah data keseluruhan adalah 2536 dataset, jumlah keseluruhan atribut 73 dimana jumlah atribut akan dijelaskan pada lembar lampiran, dan jumlah keseluruhan *class* ada 2 *class*. Hasil *clustering* dataset oleh sistem dibandingkan dengan hasil dari dataset yang lain, hal ini bertujuan untuk mengetahui hasil akurasi dari system.

4. Analisa Hasil Uji Coba

Proses pengujian dilakukan dengan melakukkan pengujian sebanyak 5 kali pada masing-masing dataset pada Tabel 5.1, setiap pengujian akan menggunakan data random sesuai banyak data yang diujikan. Kemudian dari hasil percobaan didapatkan nilai akurasinya yakni *precision, recall* dan *f-measure*. Dimana nilai precision akan menghasilkan nilai titik estimasi yang mana akan menyeimbangkan fungsi dari *precision* dan *recall*, nilai recall akan menghasilkan nilai perbandingan antara jumlah dataset yang benar/relevan yang ditemukan oleh sistem (A) dengan jumlah dataset pada kategori, dan nilai f-measure akan menghasilkan nilai kemiripan dari hasil kategorisasi suatu dataset. Pada proses perhitungan analisis recall, precision, f-measure akan menggunakan nilai inputan dari *class* asli pada dataset pengujian yang sudah ditentukan dari dataset *ozone level detection dataset*. Dari hasil evaluasi yang akan diketahui kelayakan dari metode *Hierarchical Clustering*. Untuk mengambil kesimpulan, dari hasil analisa terebut dibuat sebuah grafik *precision*, *recall* dan *f-measure* dan *time*.

4.1 Pengujian pertama dengan jumlah dataset 100 akan menghasilkan precision = 0.995, recall =1, F-measure = 0.997 dengan waktu pengujian 00:00:00.52832370. Pengujian pertama ini menghasilkan satu *cluster* dan dua kelas.

4.2 Pengujian kedua dengan jumlah dataset 200 akan menghasilkan precision = 0.995, recall =1, F-measure = 0.98 dengan waktu pengujian 00:00:59.9133420. Pengujian ketiga ini menghasilkan satu *cluster* dan dua kelas.

4.3 Pengujian ktiga dengan jumlah dataset 500 akan menghasilkan precision = 0.995, recall =1, F-measure = 0.997 dengan waktu pengujian 00:01:44.5055167. Pengujian kedua ini menghasilkan satu *cluster* dan dua kelas.

5. Kesimpulan

- Dengan menggunakan sebuah metode Hierarchical Clustering untuk mengukur jarak antara data time series yang telah ditentukan.Algoritma Hierarchical dimulai dengan menjadikan tiap objek menjadi sebuah cluster dan secara iterasimenggabungkan tiap cluster yang mirip. Jarak antar objek merupakan input dari algoritma ini. Iterasi terus berlanjut sampai semua objek telah dicluster hingga menjadi cluster saja.
- Dari penelitian ini dapat di ketahui hasil dari kelayakan suatu metode Hierarchical Clustering ini.
- Algoritma hierarchical clustering mempunyai kelemahan dimana saat metode melakukan salah satu penggabungan/pemecahan dilakukan pada tempat yang salah, tidak dapat didapatkan cluster yang optimal.

6. Daftar Pustaka

[1] Han, Jiawei. 2006. Data Mining Concepts and Techniques Second Edition, Bloomington, USA.

[2] Ljupco Todorovski, 2002, *Qualitative Clustering of Short Time-Series: A Case Study of Firms Reputation Data.*

[3] Edelstein, Herbert A. 1999. Introduction to Data Mining and Knowledge Discovery, Third Edition. Two Crows Corporation. USA.

[4] Larose, Daniel T. 2005. Discovering Knowledge in Data: An Introduction to Data Mining, John Wiley & Sons, Inc. Canada.

[5] Santoso, Budi. 2007. Data Mining: Teknik Pemanfaatan Data Untuk Keperluan Bisnis. Yogyakarta.