PERBANDINGAN RENTANG DATA METRIC PADA ALGORITMA *WARD'S* TERHADAP DATA CLUSTERING

Diajukan Sebagai Syarat Untuk Menyelesaikan
Pendidikan Program Strata-1 Pada
Jurusan Teknik Informatika

Oleh :

AKHMAD SENA
NIM : 09021381520091

Jurusan Teknik Informatika
FAKULTAS ILMU KOMPUTER UNIVERSITAS SRIWIJAYA
2021

LEMBAR PENGESAHAN TUGAS AKHIR

PERBANDINGAN RENTANG DATA METRIC PADA ALGORITMA *WARD'S* TERHADAP DATA CLUSTERING

Oleh :

AKHMAD SENA
NIM : 09021381520091

Palembang, Maret 2021

Pembimbing I

Yunita, M.Cs.
NIP. 198306062015042002

Pembimbing II

Kanda Januar Miraswan, M.T.
NIP.1671080901900006

Mengetahui,
Ketua Jurusan Teknik Informatika

Alvi Syahrini Utami, M.Kom
NIP. 197812222006042003

TANDA LULUS UJIAN SIDANG TUGAS AKHIR

Pada hari tanggal 08 Januari 2021 telah dilaksanakan ujian sidang tugas akhir oleh Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.

Nama : Akhmad Sena

NIM : 09021381520091

Judul : Perbandingan Rentang Data Metric Pada Algoritma Ward's Terhadap Data Clustering

a.n Pembimbing I,
Yunita, M.Cs.
NIP. 198306062015042002 .....

Pembimbing II
Kanda Januar Miraswan, M.T.
NIP.199001092019031012 .....

Penguji I
Alvi Syahrini Utami, M.Kom.
NIP. 197812222006042003 .....

Penguji II
Desty Rodiah, M.T.
NIP. 1671016112890005 .....

Mengetahui,
Ketua Jurusan Teknik Informatika

Alvi Syahrini Utami, M.Kom.
NIP. 197812222006042003

HALAMAN PERNYATAAN

Yang Bertanda Tangan di bawah ini :

Nama : Akhmad Sena

NIM : 09021381520091

Program Studi : Teknik Informatika

Judul Skripsi : Perbandingan Rentang Data Metric Pada Algoritma Ward's Terhadap Data Clustering

Hasil Pengecekan software *iTheticate* / *Turnitin* : 20%

Menyatakan bahwa laporan saya merupakan hasil karya sendiri dan bukan hasil penjiplakan / plagiat. Apabila ditemukan unsur penjiplakan / plagiat dalam laporan proyek ini, maka saya bersedia menerima sanksi akademik dari Universittas Sriwijaya sesuai dengan ketentuan yang berlaku.

Demikian pernyataan ini saya buat dengan sebenarnya dan tidak ada paksaan dari pihak manapun.

Palembang, Juni 2021

Akhmad Sena

09021381520091

MOTTO DAN PERSEMBAHAN

MOTTO:

- **No matter what happened, what is important is the result**
- **Don’t Rush, you’ll bound to make mistakes.**

Kupersembahkan karya tulis ini kepada :

- ***Kedua orang tuaku***
- ***Keluarga besarku***
- ***Teman - teman Informatika 2015***
- ***Fakultas Ilmu Komputer***
- ***Universitas Sriwijaya***

PERBANDINGAN RENTANG DATA METRIC PADA ALGORITMA *WARD'S* TERHADAP DATA CLUSTERING

Oleh:
Akhmad Sena
0902138152009

ABSTRAK

Clustering merupakan suatu cabang metodologi peneltian data mining untuk mengelompokan objek-objek kedalam sebuah regu dalam *cluster* yang berbeda, beberapa algoritma yang telah diterapkan pada metodologi ini adalah *Ward's*. Pada penerapan metode *clustering* banyak factor yang dapat mempengaruhi hasil perolehan cluster salah satu nya adalah penggunaan metode jarak vektor yang berbeda dapat mempengaruhi hasil *cluster yang didapat* secara drastis. Pada algoritma *Ward's* penggunaan metode jarak vektor *Euclidean distance* seringkali memberikan hasil cluster yang lebih baik dengan nilai Indeks DBI yang relatif lebih rendah dibandingkan saat menggunakan metode jarak vektor lainnya.

Kata kunci: *Data mining,* Jarak Vector , *Ward's* , *Davies Bouldin Index*

Palembang, Januari 2021

Pembimbing I	Pembimbing II
Yunita, M.Cs.	Kanda Januar Miraswan, M.T.
NIP. 198306062015042002	NIP.1671080901900006

Mengetahui,
Ketua Jurusan Teknik Informatika

Alvi Syahrini Utami, M.Kom
NIP. 197812222006042003

COMPARISON OF METRIC DATA RANGES IN WARD'S ALGORITHM TO DATA CLUSTERING

By:
Akhmad Sena
09021381520091

ABSTRACT

Clustering is a branch of data mining methodology which objective is to cluster multiple objects to form a group of different clusters, some of the algorithm that have been applied in this methodology is *Ward's*. In the application of clustering method, a lot of factor can affect the results of cluster, one of them is distance vector method which might affect the cluster result by a huge margin. In *Ward's* algorithms, the use of *Euclidean* distance often gives better cluster results with relatively lower DBI index score compared to when using another distance vector methods.

Key Word: *Data mining, Distance* Vector , *Ward's* , *Davies Bouldin Index*

Palembang, Januari 2021

Pembimbing I

Yunita, M.Cs.
NIP. 198306062015042002

Pembimbing II

Kanda Januar Miraswan, M.T.
NIP.1671080901900006

Mengetahui,
Ketua Jurusan Teknik Informatika

Alvi Syahrini Utami, M.Kom
NIP. 197812222006042003

KATA PENGANTAR

Puji syukur kepada Allah atas berkat dan rahmat-Nya yang telah diberikan kepada Penulis sehingga dapat menyelesaikan Tugas Akhir ini dengan baik. Tugas akhir ini disusun untuk memenuhi salah satu syarat guna menyelesaikan pendidikan program Strata-1 pada Fakultas Ilmu Komputer Program Studi Teknik Informatika di Universitas Sriwijaya.

Dalam menyelesaikan Tugas Akhir ini banyak pihak yang telah memberikan bantuan dan dukungan baik secara langsung maupun secara tidak langsung. Penulis ingin menyampaikan rasa terima kasih kepada:

1. Orang tuaku, Abdullah dan Masnun serta seluruh keluarga besarku yang selalu mendokan serta memberikan dukungan baik moral maupun materil.
2. Bapak Jaidan Jauhari selaku Dekan Fakultas Ilmu Komputer Universitas Sriwijaya, Ibu Alvi Syahrini Utami selaku Ketua Jurusan Teknik Informatika, dan Ibu Anggina Primanita selaku Sekretaris Jurusan Teknik Informatika.
3. Ibu Yunita selaku dosen pembimbing I dan Bapak Kanda Januar Miraswan selaku pembimbing II yang telah membimbing, mengarahkan, dan memberikan motivasi penulis dalam proses perkuliahan dan pengerjaan Tugas Akhir.
4. Bapak Kanda Januar Miraswan selaku dosen pembimbing akademik, yang telah membimbing, mengarahkan dan memberikan motivasi penulis dalam proses perkuliahan dan pengerjaan Tugas Akhir.
5. Bapak Ali Ibrahim selaku ketua dosen penguji , Ibu Alvi Syahrini Utami selaku dosen penguji I dan Ibu Desty Rodiah selaku dosen penguji II yang telah memberikan masukan dan dorongan dalam proses pengerjaan Tugas Akhir.
6. Seluruh dosen Program Studi Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.
7. Mbak Winda, Mbak Wiwin, dan seluruh staf tata usaha yang telah membantu dalam kelancaran proses administrasi dan akademik selama masa perkuliahan.

8. Teman-teman jurusan Teknik Informatika yang telah berbagi keluh kesah, motivasi, semangat, dan canda tawa selama masa perkuliahan.

Penulis menyadari dalam penyusunan Tugas Akhir ini masih terdapat banyak kekurangan disebabkan keterbatasan pengetahuan dan pengalaman, oleh karena itu kritik dan saran yang membangun sangat diharapkan untuk kemajuan penelitian selanjutnya. Akhir kata semoga Tugas Akhir ini dapat berguna dan bermanfaat bagi kita semua.

Palembang, Januari 2021

Akhmad Sena

DAFTAR ISI

Halaman

TANDA LULUS UJIAN SIDANG TUGAS AKHIR ..... **iii**
MOTTO DAN PERSEMBAHAN..... **v**
ABSTRACT..... **vi**
ABSTRAK ..... **vii**
KATA PENGANTAR..... viii
DAFTAR ISI..... x
DAFTAR TABEL xiv
DAFTAR GAMBAR..... xvi
DAFTAR ALGORITMA..... xvi
DAFTAR LAMPIRAN xix
DAFTAR ISTILAH, SINGKATAN, DAN LAMBANG..... xx
BAB I I-1
1.1. Pendahuluan I-1
1.2. Latar Belakang I-1
1.3. Perumusan Masalah..... I-6
1.4. Tujuan Penelitian..... I-6
1.5. Manfaat Penelitian..... I-7
1.6. Batasan Masalah..... I-7
1.7. Sistematika Penulisan..... I-8
1.8. Kesimpulan..... I-9
BAB II II-1
2.1. Pendahuluan II-1

2.2. Data Mining II-1

2.3. Clustering II-3

2.4. Hierarchial Clustering II-4

2.5. *Agglomerative Hierarchial Clusteringz* II-6

2.6. Metode *Linkage* II-7

2.7. Jarak Vektor II-8

2.7.1. Euclidean Distance II-9

2.7.2. *Manhattan Distance* 10

2.7.3. *Minkowski Distance* 11

2.7.3. *Mahalanobis Distance* II-12

2.9. Data *Pre-processing* II-13

2.10. Validasi Cluster II-14

2.10.1. *Davies-Bouldin Index* II-16

2.11. Penelitian Lain yang Relevan II-17

2.12. Kesimpulan II-22

BAB III III-1

3.1. Pendahuluan III-1

3.2. Pengumpulan Data III-1

3.3. Struktur Penelitian III-1

3.4. Tahapan Penelitian III-3

3.4.1. Menetapkan Kerangka Kerja III-3

3.4.2. Menetapkan Kriteria Pengujian III-6

3.4.3. Menetapkan Format Data Pengujian III-7

3.4.4. Menentukan Alat yang Digunakan dalam Pelaksanaan Penelitian . III-8

3.4.5. Melakukan Pengujian Penelitian III-8

3.4.6. Melakukan Analisa Hasil Pengujian dan Membuat Kesimpulan Penelitian III-8

3.5. Metode Pengembangan Perangkat Lunak III-9

3.5.1. *SCRUM* III-10

3.5.2. Manajemen Proyek Penelitian III-12

BAB IV IV-1

4.1. Pendahuluan IV-1

4.2. *SCRUM* IV-1

4.2.1. *Backlog* IV-1

4.2.2. Perancangan *Sprint* IV-3

BAB V IV-1

5.1. Pendahuluan V-1

5.2. Data Hasil Percobaan/Penelitian V-1

5.2.1. Konfigurasi Pengujian V-1

5.2.2. Data Hasil Perhitungan I V-2

5.2.3. Data Hasil Perhitungan II V-3

5.2.4. Data Hasil Perhitungan III V-4

5.2.5. Data Hasil Pengujian IV V-5

5.3. Analisis Hasil Penelitian V-6

5.4. Kesimpulan V-11

BAB VI VI-1

6.1. Pendahuluan VI-1

6.2. Kesimpulan VI-1

DAFTAR TABEL

Halaman

Tabel II-1. Hasil Perbandingan Validasi *Cluster* menggunakan *Euclidean Distance* dan *Manhattan Distance* ……... II-17

Tabel III-1. Contoh Data sebelum sebelum normalisasi menggunakan metode *Min-Max* …… III-4

Tabel III.2. Data set setelah nilai attribut dinormalisasi menggunakan metode *Min-Max* 5…….. III-5

Tabel III-2. Rancangan Tabel Hasil Pengujian Perubahan pada *Plaintext*…….III-7

Tabel III- 3. Rancangan Tabel Hasil Perbandingan……III-9

Tabel III-4. *Work Breakdown Structure* (WBS) Penelitian…….. III-16

Tabel IV-1. Kebutuhan Pengguna…….. IV-2

Tabel IV-2. Kebutuhan Sistem……... IV-2

Tabel IV-3. *Sprint Backlog* Pertama……...IV-3

Tabel IV-4. *Sprint Backlog* Kedua……. IV-3

Tabel IV-5. *Sprint Backlog* Ketiga……IV-4

Tabel IV-6. Definisi Aktor……. IV-6

Tabel IV-7. Definisi *Use Case Sprint Backlog* Kedua……... IV-7

Tabel IV-8. *Use Case* Memilih *File*……... IV-7

Tabel IV-9. *Use Case Euclidean Distance*……IV-9

Tabel IV-10. *Use Case Manhattan Distance*……..IV-9

Tabel IV-11. *Use Case Minkowski Distance*……IV-10

Tabel IV-12. *Use Case Mahalanobis* Distance…… IV-11

Tabel IV-13. Deskripsi Kelas *Sprint Backlog* Kedua……...IV-15

Tabel IV-14. Rencana Pengujian *Sprint Backlog* Kedua……. IV-19

Tabel IV-15. Pengujian *Sprint Backlog* Kedua…… IV-20

Tabel IV-16. Definisi *Use Case Sprint Backlog* Ketiga……...IV-23

Tabel IV-17. *Use Case* Melakukan Perhitungan cluster terhadap data…… IV-23

Tabel IV-18. Deskripsi Kelas *Sprint Backlog* Ketiga……...IV-25

Tabel IV-19. Rencana Pengujian *Sprint Backlog* Ketiga……. IV-27

Tabel IV-20. Pengujian *Sprint Backlog* Ketiga…… IV-28

Tabel V-1. Perhitungan Algoritma *Ward's* terhadap metode *Euclidean Distance*……... V-2

Tabel V-2. Perhitungan Algoritma *Ward's* terhadap metode Manhattan……....V-3

Tabel V-3. Perhitungan Algoritma *Ward's* terhadap metode *Minkowski* Distance……....V-4

Tabel V-4. Perhitungan Algoritma *Ward's* terhadap metode *Mahalanobis* Distance……V-5

Tabel V-5. Hasil Perbandingan…… V-6

DAFTAR GAMBAR

Halaman

Gambar II-1. Proses dalam Knowledge Discovery Data …… II-3

Gambar II-2. Visualisasi Plot Hierarchial Clustering …….... II-5

Gambar II-3. Pengaplikasian Metode Agglomerative dan Divisive Clustering.. II-7

Gambar II-4. Visualisasi Perhitungan Euclidean Distance Pada Diagram

Matriks Dua Dimensi ……... II-10

Gambar II-5. Visualisasi Perhitugan *Manhattan Distance* Pada Diagram Matriks Dua Dimensi …….....II-11

Gambar II-6. Dendogram Algoritma *Ward's* menggunakan *Euclidean Distance*

……II-18

Gambar II.7. Dendogram Algoritma *Ward's* menggunakan ……II-19

Gambar II-8. Visualisasi diagram pada data set Dars Sill menggunakan metode Generalized Ward pada 4 grup cluster …….. II-21

Gambar II-9. Visualisasi diagram pada data set Dars Sill menggunakan metode *Generalized Ward* pada 8 grup *cluster* ……..... II-21

Gambar III-1. Struktur Penelitian …….. III-2

Gambar III-2. Struktur Tahapan Penelitian ……... III-3

Gambar III-2. Struktur Tahapan Penelitian ……... III-6

Gambar III-4. Arsitektur *SCRUM* …….. III-10

Gambar III-5. WBS dan Gantt Chart …….. III-13

Gambar III-6. Lanjutan Gambar III-5……...III-14

Gambar III-7. Lanjutan Gambar III-6……...III-15

Gambar IV-1. Rancangan Antarmuka Perangkat Lunak……IV-4

Gambar IV-2. Implementasi Antarmuka…… IV-5

Gambar IV-3. Diagram *Use Case Sprint Backlog* Kedua…….. IV-6

Gambar IV-4. Diagram Aktivitas Memilih *File*……...IV-12

Gambar IV-5. Diagram Aktivitas *Euclidean distance*……. IV-12

Gambar IV-6. Diagram Aktivitas *Manhattan distance*……... IV-13

Gambar IV-7. Diagram Aktivitas *Minkiowski distance*…….IV-13

Gambar IV-8. Diagram Aktivitas *Mahalanobis distance*……IV-14

Gambar IV-9. Kelas Diagram *Sprint Backlog* Kedua…….. IV-15

Gambar IV-10. Diagram *Sequence* Memilih *File*…….IV-16

Gambar IV-11. Diagram *Sequence* Memilih Metode pengukuran jarak

Euclidean Distance……... IV-17

Gambar IV-12. Diagram *Sequence* Memilih Metode pengukuran jarak

Manhattan Distance……..IV-17

Gambar IV-13. Diagram *Sequence* Memilih Metode pengukuran jarak

Minkowski Distance……IV-18

Gambar IV-14. Diagram *Sequence* Memilih Metode pengukuran jarak
Mahalanobis Distance…….. IV-18

Gambar IV-15. *Use Case Sprint Backlog* Ketiga…….IV-22

Gambar IV-16. Diagram Aktivitas Melakukan perhitungan cluster *Ward's*... IV-24

Gambar IV-17. Kelas Diagram *Sprint Backlog* Ketiga…… IV-26

Gambar IV-18. Diagram *Sequence* Melakukan Perhitungan *cluster* algoritma
Ward's…….. IV-27

Gambar V-1. menunjukkan dendrogram perhitungan algortima menggunakan
Euclidean Distance. …… V-9

Gambar V-2. menunjukkan dendrogram perhitungan algortima menggunakan
Manhattan Distance……. V-9

Gambar V-3. menunjukkan dendrogram perhitungan algortima menggunakan
Minkowski Distance……... V-10

Gambar V-4. menunjukkan dendrogram perhitungan algortima menggunakan
Mahalanobis Distance……V-10

BAB I

PENDAHULUAN

1.1. Pendahuluan

Bab ini menjelaskan mengenai latar belakang, perumusan masalah, tujuan dan manfaat penelitian, batasan masalah/ruang lingkup, dan sistematika penulisan dari penelitian dan kesimpulan mengenai BAB I.

1.2. Latar Belakang

Clustering adalah proses mengelompokan objek-objek yang mempunyai kemiripan kedalam grup yang berbeda atau lebih tepatnya data dipartisi kedalam beberapa himpunan untuk menciptakan sebuah subset data sesuai dengan ukuran jarak yang ditentukan (Omran, Mahamed & Engelbrecht, Andies & Salman, Ayed 2007). *Clustering* sendiri telah diterapkan di berbagai bidang ilmu dan penelitian dan telah diaplikasikan oleh berbagai instansi dan organisasi bahkan pada aplikasi umum yang sering kita jumpai.

Analisis *clustering* mempunyai lima tipe metode penerapan, dengan metode yang paling sering digunakan adalah *centroid-based clustering* (*partitional clustering*) dan *connectivity-based clustering* (*hierarchial clustering*), dan tiga metode lainnya berupa *fuzzy clustering, density-based clustering* dan *model based clustering. Partitional clustering* merupakan Teknik clustering yang menentukan cluster berdasarkan titik centroid cluster dimana setiap jarak node dihitung berdasarkan kedekatan pada centroid. Proses pembentukan centroid pada

clustering sering kali melibatkan proses yang *arbitrary*, hal ini membuat hasil proses *clustering* dengan data yang sama menghasilkan hasil yang berbeda, sehingga proses akan melibatkan proses iterasi. Berkebalikan dari *hieararchial clustering, hieararchial clustering* merupakan metode yang menentukan *cluster* dengan menghitung jarak antar node yang bertetanggaan, perhitungan ini dilakukan karena setiap node pada proses *cluster* bisa jadi adalah kelompok *cluster* tersendiri, setiap node data akan dihitung dengan setiap node yang bertetangga, dengan kata lain setiap node mempunyai kemungkinan sebagai cluster itu tersendiri.

Proses Perhitungan *hierarchial* dilakukan dengan mengitung semua objek data sebagai objek yang dapat membuat sebuah cluster baru, objek *cluster* tersebut dihiting berulang secara iterasi hingga jumlah cluster yang diinginkan. Hierarchial clustering bedasarkan pola operasi nya dibagi menjadi dua metode yaitu *agglomerative* dan *divisive*. Perbedaan kedua pola ini adalah dimana *agglomerative* menginisiasi perhitungan dengan memperlakukan setiap *data node* sebagai clusternya tersendiri, kemudian pada setiap perhitungan dua data yang paling berdekatan akan membentuk cluster baru. Proses Perhitungan akan dilakukan secara terus-menerus hingga semua objek data membentuk sebuah satu cluster besar, pada beberapa kasus proses perhitungan juga dapat dimatikan jika telah mencapai beberapa batasan kondisi tertentu (Henry, Edi, 2009). Pada *Divisive clustering* proses yang sama juga terjadi namun inisialisasi cluster dan bentuk akhir cluster berkebalikan dengan *Agglomerative clustering*, objek data

diinisiasi sebagai suatu cluster besar dan berakhir dengan setiap *node* merupakan *cluster* mereka tersendiri.

Berdasarkan metode perhitungan yang diterapkan, *hierarchial clustering* pada beberapa kasus menghasilkan waktu komputasi yang relative lebih lama dari *partitional clustering*. Hal ini dikarenakan oleh lama dan jumlah proses perhitungan bergantung dari jumlah objek data yang akan di-*cluster*, dimana jumlah perhitungan yang harus dilakukan adalah *n* sebagai data objek, *n*-1. Hal ini memberikan masalah pada analisis *cluster* dengan jumlah data yang besar memakan biaya dan waktu yang lama dimana pada proses perhitungan *partional clustering* mungkin hanya membutuhkan iterasi yang lebih sedikit.

Clustering melibatkan beberapa tahapan/proses yaitu *data preprocessing*, inisialisasi *cluster*, perhitungan jarak vektor, dan evaluasi. Dalam analisis *cluster*, langkah pertama yang harus diambil oleh peneliti adalah menentukan data poin yang akan dihitung jarak dan kemiripannya (Clatworthy, Buick, Hankins, Weinmann, & Horne, 2005). Perhitungan jarak vector cluster merupakan salah satu tahapan yang paling penting, tahap ini berperan besar dalam menemukan hasil clustering terbaik. Hasil dari pengukuran jarak vektor dapat bergantung dari metode yang digunakan, pengukuran jarak vektor yang berbeda dapat menghasilkan nilai yang berbeda, hal yang sama juga terjadi pada effisiensi dari metode *cluster* analisis yang digunakan.

Metode Perhitungan jarak vektor yang paling sering digunakan adalah *Euclidean distance*. Algoritma ini melakukan perhitungan antar dua data poin dan mengembalikan nilai dalam bentuk satuan jarak, algoritma ini memproses setiap

perhitungan jarak terhadap semua bentuk pasangan data poin dan menampilkan hasilnya pada *proximity matrix*. Metode perhitungan ini tidak melibatkan perhitungan yang kompleks, namun dikarenakan perhitungan harus dilakukan berulang-ulang terhadap setiap node menyebabkan perhitungan pada clustering terutama pada clustering yang bersifat *hierarchial* membutuhkan proses yang lama.

Penelitian Trudie Strauss dan Michael Johan von Maltiz pada tahun 2017 denga judul *"Generalizing Ward's Method for Use With Manhattan Distance"* menyatakan kepada pengarang lain yaitu Vogt W. Nagel D. dalam penelitiannya yang berjudul *"Cluster Analysis in Diagnosis"* berargumen bahwa penggunaan metode *ward* dalam analisis *cluster* hanya terbatas dalam memnggunakan *Euclidean distance* saja, beliau juga menyatakan bahwa dasar dari *linkage method* pada *ward* berbasis dari *Euclidean distance*. Teori tersebut juga diperkuat dengan manuals yang dicantumkan didalam beberapa software seperti R. Selepas dari penelitian tersebut Strauss dan Johan menyangkal bahwa penggunaan metode *ward* terbatas dalam *Euclidean distance* saja. Dalam penelitiannya beliau melakukan percobaan terhadap analisis *cluster* sensus etnis keluarga eropa menggunakan metode *ward* dengan pengukuran jarak menggunakan *Manhattan distance*. Dalam hasil penelitiannya, beliau menyatakan bahwa penggunaan *Manhattan distance* memberikan hasil yang lebih baik dibandingkan saat menggunakan *Euclidean distance* terutama dalam bagian keeratan stabilitas dan keterhubungan antar *cluster* meskipun hasil perbandingan kedua metode *distance* tidak jauh berbeda.

Penelitian lainnya yang similarjuga pernah dilakukan oleh Srividya Ramaswami dengan judul "*Comparing the Efficieny of Two Clustering Techniques*" pada penelitiannya beliau ingin mengetahui bagaimana pemilihan *distance function* dapat mempengaruhi sifat dari *hierarchy clustering.* Penelitiannya dilakukan dengan mengambil beberapa sampel dari twitter, data sampel yang terkumpul akan di proses menggunakan metode ward. Pada hasil penelitiannya pada percobaan 1-Cluster dan 2-Cluster tidak ditemukan perbedaan yang siginifikan, namun pada percobaan 3-Cluster, 4-Cluster dan 5-Cluster ditemukan sedikit perbedaan yang signifikan terhadap hasil clustering, namun hasil penelitian tersebut memeberikan bayang-bayang dari bagaimana jika penelitian dilakukan dengan data yang lebih besar.

Penelitian serupa lainnya pernah diterapkan oleh Toon Van Craenendonk dan Hendrick Blockeel pada tahun 2015 dengan penelitian yang berjudulkan "*Using Internal Validity Measures to Compare Clustering Algorithms*". Penelitian ini dilakukan dengan membandingkan beberapa metode clustering dengan menggunakan *internal validity measures*. Hasil dari penelitiannya menyatakan bahwa hasil cluster yang tidak seimbang memberikan hasil clustering yang bagus, hasil perhitungan dari *clustering performance* sangat bergantung dari jumlah *noise* dan metode seperti *k-means* dinyatakan sudah cukup untuk menghasilkan cluster yang baik.

Berdasarkan dari penelitian yang telah dilakukan sebelumnya dapat disimpulkan banyak faktor yang dapat mempengaruhi hasil dari clustering seperti perhitungan jarak vektor, kualitas data dan preprosessing data. Dari kesimpulan

tersebut, penulis akan melakukan penelitian analisis dari performa dua metode hierarchial clustering menggunakan beberapa tahapan yang berbeda, umumnya pada penggunaan perhitungan jarak vektor.

1.3. Perumusan Masalah

Berdasarkan latar belakang di atas, maka dapat diidentifikasi permasalahan yang muncul pada performa algoritma *Ward's* pada clustering terhadap *distance method* yang berbeda.

Untuk menyelesaikan masalah di atas maka disusun beberapa pertanyaan penelitian (Research Questions):

1. Bagaimana mengukur jarak vector data pada *hierarchial* Clustering?
2. Bagaimana jarak vektor data dapat dapat mempengaruhi sifat dan ciri-ciri dari *hierachial* clustering?
3. Mengetahui penggunaan metode pengukuran jarak vektor mana yang mempunyai performa terbaik terhadap algoritma *ward's*?

1.4. Tujuan Penelitian

Tujuan yang harus dicapai dalam penelitian ini adalah menganalisa hasil cluster terhadap algoritma clustering *ward's* terhadap penggunaan metode pengukuran jarak vektor yang berbeda.

1.5. Manfaat Penelitian

Adapun manfaat yang diharapkan tercapai dari penulisan dan pelaksanaan penelitian ini adalah :

1. Memberikan dokumentasi mengenai kinerja algoritma ward menggunakan metode pengukuran jarak data yang berbeda.
2. Memberikan referensi terhadap penelitian setelahnya yang memiliki kaitan terhadap *clustering* allgoritma *ward's* dan penggunaan metode pengukran jarak data

1.6. Batasan Masalah

Beberapa batasan masalah pada penelitian ini, adalah sebagai berikut.

1. Data yang terlibat dalam proses clustering adalah file berekstensi *.csv* dan memiliki content yang bersifat mixed dari text, nominal dan beberapa tipe data lainnya
2. Data yang digunakan merupakan data sekunder yang diambil dari *UCI Machine Learning Repository* (http:// http://archive.ics.uci.edu/ml/index.php).
3. Metode yang akan digunakan adalah metode *hierarchial clustering, ward's linkage method*.
4. Pengukuran performa clustering akan dilakukan menggunakan metode *Davies Bouldin Index* (DBI).
5. Pengukuran jarak vector data akan menggunakan *Euclidean distance, Manhattan distance, Minkowski distance, dan Mahalanobis distance*.

1.7.Sistematika Penulisan

Sistematika penulisan pada dokumentasi adalah sebagai berikut :

BAB I. PENDAHULUAN

Bab ini memberikan penjelasan mengenai latar belakang, perumusan masalah, tujuan dan manfaat penelitian, batasan masalah/ruang lingkup, dan sistematika penulisan.

BAB II. KAJIAN LITERATUR

Bab ini membahas dasar-dasar teori dari penelitian. Kajian Lietraur meliput hasil penelitian terdahulu yang relevan dengan penelitian ini, penjelasan mengenai clustering, *hierarchial clustering* dan *linkage distance method*, *ward's* clustering, pengukuran jarak vektor dan penelitian lainnya yang berkaitan dengan penelitian ini.

BAB III. METODOLOGI PENELITIAN

Bagian ini membahas tahapan proses pelaksanaan pada penelitian ini. masing-masing rencana tahapan penelitian dideskripsikan dengan rinci dengan mengacu pada suatu kerangka

kerja. Di akhir bab ini berisi perancangan manajemen proyek pada pelaksanaan penelitian.

BAB IV. PENGEMBANGAN PERANGKAT LUNAK

Bab ini membahas proses perancangan perangkat lunak yang akan dibangun pada penelitian ini.

BAB V. HASIL DAN ANALISIS PENELITIAN

Bab ini menampilkan hasil pengujian/penelitian berdasarkan langkah-langkah yang telah direncanakan. Analisis dari penelitian tercapai sebagai basis dari kesimpulan yang diambil dalam penelitian ini.

BAB VI. KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan dari penelitian berdasarkan hasil yang didapat pada bab-bab seebelumnya dan juga berisi saran-saran yang diharapkan berguna dalam pengembangan penelitian yang memperlibatkan *distance metric*.

1.8. Kesimpulan

Hierarchial Clustering merupakan metode clustering dimana performa perhitungan secara besar dipengaruhi dari ukuran data. Pada penelitian ini akan membandingkan dua algoritma *hierarchial cluster* yang berbeda dengan

menggunakan tiga *distance method* yang berbeda sehingga dapat melakukan analisis bagaimana pendekatan yang berbeda pada suatu algoritma *clustering* dapat mempengaruhi performa pada metode tersebut.

DAFTAR PUSTAKA

Ali, M. M. (2017). Cryptography : A Comparative Analysis for Modern Techniques, *8*(6), 442–448.

Archana Singh, Avantika Yadav, and Ajay Rana (April, 2013). K-means with Three different Distance Metrics. *In International Journal of Computer Applications (0975 – 8887) Volume 67– No.10*. 13-17.

Binbin Lu, Martin Charlton, Chris Brunsdon & Paul Harris (September, 2015). The Minkowski approach for choosing the distance metric in geographically weighted regression. *In International Journal of Geographical Information Science*. 1-18.

D. Saraswathi1 and Dr. L.Mary Immaculate Sheela (2016). Lung Image Segmentation Using K-Means Clustering Algorithm With Novel Distance Metric. *In International Journal of Recent Trends in Engineering & Research*. 236-245.

Daniel Mullner (October, 2011). Fastcluster: Fast Hierarchical, Agglomerative Clustering Routines for R and Python. In *Journal of Statistical Software May 2013, Volume 53, Issue 9*. 1-17.

Darshan Sonagara and Soham Badheka (October 2014). Comparison of Basic Clustering Algorithms. In *International Journal of Computer Science and Mobile Computing, Vol.3 Issue.10*. 58-61.

Dengsheng Zhang and Guojun Lu (2003). Evaluation of Similarity Measurement For Image Retrieval

Dharmendra S. Modha and W. Scott Spangle (September, 2003). Feature Weighting in k-Means Clustering.

Dr. Luai Al Shalabi and Dr. Zyad Shaaban (2006). Normalization as a Preprocessing Engine for Data Mining and the Approach of Preference MatrixInternational.

Erhard Rahm and Hong Hai Do (2000). Data Cleaning: Problems and Current Approaches. In *Bulletin of the Technical Committee on Data Engineering1*. 3-13.

Fionn Murtagh, Pierre Legendre (2014). Ward's Hierarchical Agglomerative Clustering Method:Which Algorithms Implement Ward's Criterion?. *In Journal of Classification 31*. 274-295.

Herny Februariyanti and Edi Winarko (2013). Klastering Dokumen Menggunakan Hierarchial Agglomerative Clustering.

J. A. Hartigan (Maret, 2012). Consistency of Single Linkage for High-Density Clusters. In *Journal of the American Statistical Association*. 388-394.

K.Sasirekha and P.Baby (March, 2013). Agglomerative Hierarchical Clustering Algorithm- A Review. In *International Journal of Scientific and Research Publications, Volume 3, Issue 3, March 2013*. 1-2.

Kalaivany Natarajan, Jiuyong Li, Andy Koronios (September 2009). Data Mining Techniques For Data Cleaning. In *4th World Congress on Engineering Asset Management*. 796-804.

Kimberly L. Elmore and Michael B. Richman (September 2000). Euclidean Distance as a Similarity Metric for Principal Component Analysis.

Liwei Wanng, Yan Zhang, and Jufu Feng (August, 2005). On the Euclidean Distance of Images. In *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 27, NO. 8*. 1334-1339.

M. Kuchaki Rafsanjani, Z. Asghari Varzaneh, N. Emami Chukanlo (2012). A survey of hierarchical clustering algorithms. In *The Journal of Mathematics and Computer Science Vol. 5 No. 3*. 229-240.

Odilia Yim and Kylee T. Ack Baraly (February, 2015). Hierarchical Cluster Analysis: Comparison of Three Linkage Measures and Application to Psychological Data

Refat Aljumily (2016). Agglomerative Hierarchical Clustering: An Introduction to Essentials. Proximity Coefficients and Creation of a Vector-Distance Matrix and Construction of the Hierarchical Tree and a Selection of Methods. In *Global Journal of HUMAN-SOCIAL SCIENCE: G Linguistics & Education Volume 16 Issue 3 Version 1.0 Year 2016*.

Renato Cordeiro de Amorim (January, 2011). Learning feature weights for K-Means clustering using the Minkowski metric.

Richa Loohach and Kanwal Garg (July 2012). Effect of Distance Functions on K-Means Clustering Algorithm. In *International Journal of Computer Applications (0975 – 8887) Volume 49– No.6, July 2012*. 7-9.

Sadaaki Miyamoto and Akihisa Terami (July, 2010). Semi-Supervised Agglomerative Hierarchical Clustering Algorithms with Pairwise Constraints.

Sepandar D. Kamvar, Dan Klein and Christopher D. Manning (October, 2002). Interpreting and Extending Classical Agglomerative Clustering Algorithms using a Model-Based Approach.

Shiming Xiang, Feiping Nie and Changshui Zhang (2008). Learning a Mahalanobis distance metric for data clustering and classification. In *Pattern Recognition 41*. 3600-3612.

Shraddha Pandit, and Suchita Gupta (2011). A Comparative Study on Distance Measuring Approaches for Clustering. In *International Journal of Research in Computer Science eISSN 2249-8265 Volume 2 Issue 1*. 29-31.

Shweta Srivastava (2014). Weka: A Tool for Data preprocessing, Classification, Ensemble, Clustering and Association Rule Mining.

Silke Wagner and Dorothea Wagner (January, 2007). Comparing Clusterings - An Overview.

Sung-Hyuk Cha (2007). Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. In *International JOurnal of Mathematical Models and Methods In Applied Sciences*. 300-307.

Timur Sattarov and Mircea-Serban Pavel (August, 2015). On the importance of preprocessing and initialization in k-means.

Toon Van Craenendonck and Hendrik Blockeel (2015). Using Internal Validity Measures to Compare Clustering Algorithms.

Trudie Strauss, and Michael Johan von Maltitz (2017). Generalising Ward's Method for Use with Manhattan Distances.

Vaishali R. Patel and Rupa G. Mehta (2011). Impact of Outlier Removal and Normalization Approach in Modified k-Means Clustering Algorithm. In *International Journal of Computer Science Issues Volume 8, Issue 5, No 2, September 2011*. 331-343.

Yogita Rani and Dr. Harish Rohil (2013). A Study of Hierarchical Clustering Algorithm. In *International Journal of Information and Computation Technology.ISSN 0974-2239 Volume 3, Number 11 (2013).*