IMPLEMENTASI ALGORITMA IMPROVED K-MEANS PADA PORTAL JURNAL INTERNASIONAL

Amelia Handana Putri[1] , Randy Cahya W, S.ST.,M.Kom[2], M.Ali Fauzi, S.Kom, M.Kom[3]
Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Brawijaya Malang
Jl. Veteran, Malang, Jawa Timur, Indonesia 65145
Telp 0341-551611, Fax 0341-565420
Email:[1]ameliahandana@gmail.com, [2]rendicahya@ub.ac.id, [3]moch.ali.fauzi@ub.ac.id

Abstrak

Penelitian adalah hal utama yang membawa pengaruh cukup besar dalam perkembangan teknologi saat ini. Hasil riset yang telah dilakukan oleh peneliti mampu menjadi bahan referensi oleh peneliti lainnya. Pada umumnya, referensi tersebut berasal dari publikasi jurnal yang telah diterbitkan secara internasional. Saat ini belum banyak dilakukan pengelompokan jurnal berdasarkan kualitasnya. Adanya sistem pengelompokan portal jurnal diharapkan dapat meminimalisasi tingkat kesalahan dalam publikasi jurnal sehingga tidak terjebak dalam jurnal predator termasuk dalam beberapa tindakan yang kurang elegan dalam publikasi jurnal ilmiah, misalnya *citation cartel* dan *junk science*. Sehingga diperlukan suatu sistem yang mampu bekerja secara otomatis untuk mengelompokkan portal jurnal internasional yang dapat membantu peneliti tersebut. Teknik yang digunakan adalah *clustering* menggunakan metode Improved K-Means. Metode Improved K-Means melakukan pengembangan pada inisialisasi *centroid* awal dengan perhitungan jarak antar data. Pengujian dilakukan dengan *silhouette coefficient* dengan range nilai antara -1 sampai 1 dimana semakin tinggi nilai dari *silhouette coefficient* maka semakin bagus *cluster* bentukan dari metode ini. Pada implementasi ini dibandingkan pula hasil pengujian pada Improved K-Means terhadap hasil dari pengujian pada K-Means Standar sehingga dapat diketahui keefektifan dari kedua metode. Pada pengujian menggunakan portal jurnal internasional, penggunaan 2 *cluster* dan 198 data menghasilkan nilai *silhouette coefficient* paling tinggi yaitu 0,869. Penggunaan metode Improved K-Means pada beberapa kasus memiliki kestabilan pada nilai *silhouette* dibandingkan dengan penggunaan metode K-Means Standar dan nilai *silhouette coefficient* yang lebih tinggi.

Kata kunci: Jurnal, *Clustering, K-Means, Improved K-Means, Silhouette Coefficient*

Abstract

Research is the main thing that brings considerable influence in the development of today's technology. Results of research conducted by the researchers were able to be a reference by other researchers. In general, the reference is derived from the publication of journals have been published internationally. Little has been done grouping journals based on its quality. A system of grouping journal portal is expected to minimize the error rate in the publication of journals that do not get stuck in predatory journals included in some of the actions that are less elegant in the publication of scientific journals, such as citation cartel and junk science. So, we need a system that can work automatically to classify international journal portal that can assist researchers. The technique used is clustering using K-Means Improved methods. Improved K-Means method to develop the initialization of initial centroid by calculating the distance between the data. Testing is done with a silhouette coefficient with a value range between -1 to 1 where the higher the value of the coefficient, the better silhouette cluster formation of these methods. In this implementation also compared results of tests on Improved K-Means of the results of testing on K-Means standards so it can know the effectiveness of both methods. In testing using the portal international journals, the use of two cluster and the data yielded values 198 silhouette highest coefficient is 0.869. Use of Improved K-Means method in some cases have stability in silhouette value compared with the use of K-Means method Standards and silhouette coefficient value is higher.

Keywords: *Journal, Clustering, K-Means, Improved K-Means, Silhouette Coefficient*

1. PENDAHULUAN

Dewasa ini, perkembangan teknologi semakin luar biasa seperti halnya perkembangan ilmu pengetahuan dan internet. Dengan adanya perkembangan tersebut, membawa dampak yang cukup besar dalam berbagai bidang salah satunya bidang penelitian. Berbagai macam hasil penelitian yang dilakukan oleh peneliti banyak dipublikasikan dalam bertuk artikel ilmiah. Publikasi hasil penelitian secara digital ini mulai berkembang pesat seiring dengan meningkatnya penggunaan teknologi komputer, tak terkecuali peneliti yang

mempublikasikan hasil penelitiannya. Hal ini menyebabkan banyak dokumen secara digital tersebar luas di internet terutama portal jurnal internasional.

Berdasarkan data dari sebuah lembaga riset internasional (SCimago), publikasi ilmiah para peneliti (khususnya dosen) di Indonesia masih sedikit dibanding dengan negara lain. Hingga tahun 2014, Indonesia mencatat sebanyak 32.355 judul penelitian yang dipublikasikan di jurnal internasional dan terakreditasi dan diindeks di Scopus. Sedangkan untuk negara Malaysia 153.378 judul paper telah dipublikasikan di jurnal internasional dan terakreditasi dan diindeks di Scopus. Sedangkan pada data lainnya, negara indonesia jika dibandingkan dengan negara tetangga malaysia, jumlah dosen di Malaysia yang bergelar doktor 8.000 orang, ternyata bisa menghasilkan 14.103 publikasi setiap tahunnya. Sedangkan di Indonesia, jumlah total dosen di seluruh perguruan tinggi sebanyak 270.000 orang (23.000 diantaranya telah bergelar doktor) hanya menghasilkan publikasi sebanyak 1.975 buah, setiap tahunnya (M. Lutfi, 2012).

Dari data tersebut tidak banyak hasil riset akademis Indonesia yang dipublikasikan dan masuk jurnal internasional yang terpercaya. Para peneliti tersebut memiliki peran besar dalam publikasi informasi ilmiah dalam bentuk jurnal di portal jurnal internasional yang diharapkan dapat meningkatkan kuantitas hasil riset akademis di Indonesia. Sehingga hasil riset tersebut mampu menjadi bahan referensi atau rujukan oleh peneliti lainnya. Selain memilih portal jurnal internasional yang terpercaya sebagai tempat publikasi, peneliti juga harus selektid dalam memilih bahan referensi jurnal yang berkualitas sehingga penelitian yang dihasilkan juga optimal dan terpercaya.

Pada beberapa kasus, pengelompokan jurnal internasional sering kali mengelompokkan konten jurnal untuk memperoleh informasi dari jurnal tersebut. Pada saat ini belum banyak ditemukan penglompokkan portal jurnal internasional berdasarkan jumlah kualitasnya seperti jumlah sitasi, jumlah dokumen dan sebagainya yang dapat memudahkan peneliti untuk mempublikasikan jurnal dan memilih referensi jurnal. Maka dari itu, diperlukan suatu sistem yang mampu bekerja secara otomatis untuk mengelompokkan portal jurnal yang dapat membantu para peneliti dalam mengembangkan riset baru, memilih referensi atau rujukan hasil riset dan mempublikasikan hasil riset mereka. Dengan adanya sistem portal jurnal ini diharapkan dapat meminimalisir tingkat kesalahan dalam publikasi jurnal sehingga tidak terjebak dalam jurnal predator termasuk dalam beberapa tindakan yang kurang elegan dalam publikasi jurnal ilmiah, misalnya *Citation Cartel* dan *Junk Science*. Salah satu teknik yang dapat diimplementasikan ke dalam sistem tersebut adalah teknik pengelompokan (*clustering*) yang menggunakan metode *Improved K-Means.*

K-Means Clustering merupakan metode untuk mengklasifikasikan atau mengelompokkan objek-objek data ke dalam K-group (*cluster*) berdasarkan atribut tertentu. Namun, *K-Means* mempunyai kelemahan yang diakibatkan oleh penentuan pusat awal *cluster*. Hasil cluster yang terbentuk dari metode *K-Means* ini sangatlah tergantung pada inisialisasi nilai pusat awal cluster yang diberikan. Hal ini menyebabkan hasil clusternya berupa solusi yang sifatnya *local optimal.* Beberapa upaya dilakukan oleh peneliti untuk meningkatkan akurasi dan efisiensi dari algoritma *K-Means. Improved K-Means* adalah salah satu metode yang dapat diterapkan untuk permasalahan *clustering* (Carlisle, 2001). Dibandingkan dengan *K-Means, Improved K-means* memiliki kelebihan yaitu lebih stabil dan kuat dibandingkan dengan metode *K-means* standar (Frans, 2009).

Berdasarkan latar belakang diatas, maka penulis menggunakan judul **"Implementasi Algoritma Improved K-Means pada Portal Jurnal Internasional "** yang diharapkan dapat membantu peneliti untuk mengembangkan riset baru, memilih referensi atau rujukan hasil riset dan mempublikasikan hasil riset mereka.

2. TINJAUAN PUSTAKA

2.1 *SCImagojr Journal & Country Rank*

SCImagojr Journal & Country Rank merupakan sebuah portal yang mencakup indikator jurnal ilmiah dan negara maju dari informasi yang terdapat dalam basis data *Scopus*® (http://www.scopus.com/) yaitu Elsevier BV (http:www.elsevier.com/). Indikator ini dapat digunakan untuk menilai dan menganalisis *domain* ilmiah.

Nama untuk platform ini diambil dari Scimago Journal Rank (SJR) *indicator*, dikembangkan oleh Scimago algoritma yang dikenal secara luas yaitu *Google Pageran*™. Indikator tersebut menunjukkan visibilitas jurnal yang terdapat dalam database *Scopus*® mulai tahun 1996.

2.2 Jurnal Ilmiah

Jurnal ilmiah dianggap sebagai sumber informasi yang paling penting di dunia ilmu pengetahuan dan teknologi. Jurnal ilmiah berisi kumpulan artikel yang dipublikasikan secara periodik ditulis oleh ilmuwan peneliti untuk melaporkan hasil-hasil penelitian terbarunya. Maka dari itu, keberadaan jurnal ilmiah merupakan hal yang penting untuk memajukan ilmu pengetahuan dan teknologi. Tulisan yang dimuat dalam jurnal ilmiah, sudah mengalami proses *peer-review* dan seleksi ketat dari para pakar di bidangnya masing-masing. Proses *peer-review* dijalankan untuk

menjamin kualitas dan validitas ilmiah artikel yang dimuat (Pustaka, 2014).

2.3 *Impact Factor*

Impact Factor (IF) oleh sebagian besar peneliti di dunia, dijadikan ukuran kualitas suatu jurnal. Semakin tinggi IF-nya maka jurnal tersebut akan semakin bergengsi. IF diperkenalkan oleh ISI *Web of Science*, sebuah perusahaan di bidang pendidikan.

Penghitungan IF untuk sebuah jurnal (misalnya untuk tahun 2010) dilakukan dengan cara menjumlahkan rata-rata *citation*/rujukan setiap karya yang diterbitkan pada 2 tahun sebelumnya (yaitu tahun 2009 dan 2008) dibandingkan dengan jumlah seluruh karya (yang bisa dirujuk) yang terbit pada tahun 2006 dan 2007. Misal suatu jurnal yang mempunyai nilai IF = 3 berarti setiap paper yang terbit pada tahun 2007 dan 2008 dirujuk oleh rata-rata 3 buah karya ilmiah lainnya. IF tahun 2010 akan keluar pada tahun 2011, dan seterusnya (Rossner, 2007).

2.4 *Data Mining*

*Data minin*g adalah suatu proses menemukan hubungan yang berarti, pola, dan kecenderungan dengan memeriksa dalam sekumpulan besar data yang tersimpan dalam penyimpanan dengan menggunakan teknik pengenalan pola seperti teknik statistik dan matematika (Kusrini, 2009). Definisi lain data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar (Turban, 2005).

2.4 *Clustering*

Clustering adalah proses pengelompokan benda serupa ke dalam kelompok yang berbeda, atau lebih tepatnya partisi dari sebuah *dataset* ke dalam *subset*, sehingga data dalam setiap *subset* memiliki arti yang bermanfaat. Dimana sebuah *cluster* terdiri dari kumpulan benda-benda yang mirip antara satu dengan yang lainnya dan berbeda dengan benda yang terdapat pada *cluster* lainnya. Algoritma *clustering* terdiri dari dua bagian yaitu secara *hirarkis* dan secara *partitional.* Algoritma *hirarkis* menemukan *cluster* secara berurutan dimana *cluster* ditetapkan sebelumnya, sedangkan algoritma *partitional* menentukan semua kelompok pada waktu tertentu (Tahta, 2012).

2.5 *K-Means Clustering*

K-Means merupakan salah satu metode data klustering non hirarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih *cluster/*kelompok. Metode ini mempartisi ke dalam cluster/ kelompok sehingga data yang memiliki karakteristik yang sama (*High intra class similarity*) dikelompokkan ke dalam satu cluster yang sama dan yang memiliki karakteristik yang berbeda (*Low intra class similarity*) dikelompokkan pada kelompok yang lain (Giyanto, 2008). Proses *clustering* dimulai dengan mengidentifikasi data yang akan dikluster, X_{ij} (i=1,...,n; j=1,...,m) dengan n adalah jumlah data yang akan diklaster dan m adalah jumlah variabel. Pada awal iterasi, pusat setiap *cluster* ditetapkan secara bebas (sembarang), C_{kj} (k=1,...,k; j=1,...,m). Kemudian dihitung jarak antara setiap data dengan setiap pusat klaster. Untuk melakukan penghitungan jarak data ke-i (x_i) pada pusat klaster ke-k (c_k), diberi nama (di_k), dapat digunakan formula Euclidean seperti pada persamaan (2.1), yaitu:

$$d_{ik} = \sqrt{\sum_{j=1}^{m}(x_{ij} - c_{kj}{}^2)} \quad (2.1)$$

Suatu data akan menjadi anggota dari klaster ke-k apabila jarak data tersebut ke pusat klaster ke-k bernilai paling kecil jika dibandingkan dengan jarak ke pusat klaster lainnya. Hal ini dapat dihitung dengan menggunakan persamaan (2.2). Selanjutnya, kelompokkan data-data yang menjadi anggota pada setiap klaster.

$$\text{Min } \sum_{k=1}^{k} d_{ik} = \sqrt{\sum_{j=1}^{m}(x_{ij} - c_{kj}{}^2)} \quad (2.2)$$

Nilai pusat klaster yang baru dapat dihitung dengan cara mencari niilai rata-rata dari data-data yang menjadi anggota pada klaster tersebut, dengan menggunakan rumus pada persamaan (2.3):

$$c_{kj} = \frac{\sum_{i=1}^{p} x_{ij}}{p}; \quad (2.3)$$

Dimana x_{ij} ∈ kluster ke-k

P = banyaknya anggota klaster ke-k

2.6 K-Means *Clustering*

K-Means merupakan salah satu metode data klustering non hirarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih *cluster*/ kelompok. Metode ini mempartisi ke dalam *cluster*/ kelompok sehingga data yang memiliki karakteristik yang sama (*High intra class similarity*) dikelompokkan ke dalam satu *cluster* yang sama dan yang memiliki karakteristik yang berbeda (*Low inter class similarity*) dikelompokkan pada kelompok yang lain (Giyanto, 2008). Proses *clustering* dimulai dengan mengidentifikasi data yang akan dikluster, X_{ij} (i=1,...,n; j=1,...,m) dengan n adalah jumlah data yang akan di *cluster* dan m adalah jumlah variabel. Pada awal iterasi, pusat setiap *cluster* ditetapkan secara bebas (sembarang), C_{kj} (k=1,...,k+1; j=1,...,m). Kemudian dihitung jarak antara setiap data dengan setiap pusat *cluster*. Suatu data akan menjadi anggota dari *cluster* ke-*k* apabila jarak data tersebut ke pusat *cluster* ke-*k* bernilai paling kecil

jika dibandingkan dengan jarak ke pusat *cluster* lainnya.

2.7 Improved K-Means

Improved K-Means merupakan modifikasi algoritma K-Means Standar untuk meningkatkan akurasi dan efisiensi. Metode modifikasi ini diuraikan sebagai berikut (Nazeer, 2009):

Masukan:
D = {d1, d2,...,dn} //kumpulan dari *n* data
k //banyaknya *cluster*
Keluaran:
Satu set *k clusters*
Langkah-langkah:
Fase 1: Menentukan pusat awal *centroid* dari *cluster*.
Fase 2: Menempatkan tiap *record* pada *cluster*.

Tahap pertama Improved K-Means, *centroid* awal ditentukan sistematis sehingga menghasilkan *cluster* dengan akurasi yang lebih baik. Tahap kedua memanfaatkan variasi dari metode pengelompokan dimulai dengan membentuk *cluster* awal berdasarkan jarak relatif setiap *record* dari *centroid* awal. *Cluster* ini kemudikan diperbaiki dengan menggunakan pendekatan heuristik, dengan demikian dapat meningkatkan efisiensi.

Dua fase algoritma Improved K-Means diuraikan sebagai berikut (Nazeer, 2009):

1. Fase menentukan pusat awal *centroid.*
   Inputkan *dataset* dan k *cluster* yang diinginkan.

   Set m=1.

   Hitung jarak tiap data dengan seluruh himpunan di D.

   Temukan pasangan data terdekat dan bentuk himpunan data-point Am ($1 \leq m \leq k$) yang berisi dua data-point tersebut. Hapus baris data dari D.

   Temukan data di D yang dekat dengan data-point Am. Tambahkan pada Am dan hapus dari D. Ulangi langkah ini hingga jumlah data-point mencapai *threshold**(n/k).

   Jika m<k, maka m=m+1, tentukan pasangan data-point yang lain, kembali ke langkah 4.

   Untuk setiap data-point Am, temukan rata-rata dari tiap vektor data-point pada Am, rata-rata ini akan menjadi inisial *centroid* (*centroid* awal).

2. Fase menempatkan tiap *record* pada *cluster*
   1. Input himpunan *dataset* D, himpunan *cluster* C.
   2. Hitung jarak tiap data-point di($1 \leq i \leq n$) dengan semua *centroid* cj($1 \leq j \leq k$) sebagai d(di, cj).
   3. Untuk setiap data-point di, temukan *centroid* yang terdekat cj dan masukan di ke *cluster* j.
   4. Set *Cluster*ID[i]=j;
   5. Set NearestDist[i]=d(di, cj);
   6. Untuk tiap *cluster* j($1 \leq j \leq k$), hitung ulang *centroid* dengan data-data yang ada pada tiap *cluster*.
   7. ULANG.
   8. Untuk tiap data-point di:

      8.1Hitung jarak data dengan *cluster* akhir yang terdekat
      8.2 IF jarak sama atau kurang dari *cluster* akhir yang terdekat, data tetap berada dalam *cluster*
      ELSE
      Data berpindah *cluster*, lalu ulangi langkah 2 hingga 5.
   9. UNTIL kriteria konvergen ditemui, yaitu ketika tidak ada data yang berpindah *cluster*.

2.8 Analisis *Cluster*

Analisis *cluster* merupakan teknik statistik multivariat yang pada awalnya dikembangkan klasifikasi biologis. Dalam penelitian untuk *clustering* ini, digunakan ukuran kualitas atau *goodness*. Jenis ukuran ini memungkinkan untuk membandingkan suatu kelompok yang berbeda dari kelompok lain berdasarkan kesamaan berpasangan dokumen dalam *cluster*. Jenis lain dari ukuran kualitas ini memungkinkan mengevaluasi seberapa baik pengelompokan tersebut dengan membandingkan kelompok yang dihasilkan oleh teknik *clustering* untuk kelas yang sudah ada (Saracli, 2013).

2.9 *Silhouette Coefficient*

Silhouette Coefficient digunakan untuk melihat kualitas dan kekuatan *cluster*, seberapa baik suatu objek ditempatkan dalam suatu *cluster*. Metode ini merupakan gabungan dari metode *cohesion* dan *separation*. Tahapan perhitungan *Silhouette Coefficient* adalah sebagai berikut (Al Zhoubi, 2008):

Untuk setiap obyek *i*, hitung rata-rata jarak antar obyek pada *cluster* tersebut. Beri nama α(i).

Untuk setiap obyek *i*, hitung rata-rata jarak ke semua obyek pada *cluster* lain. Kemudian diambil nilai mininummya. Beri nama *b*(i).

Nilai *Silhouette* Coefficient nya adalah :

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \qquad (2.4)$$

Nilai *silhouette coefficient* dapat bervariasi antara -1 dan 1. Nilai -1 tidak diinginkan dimana α(i) lebih besar dari *b*(i). Hal yang diinginkan adalah nilai *silhouette* positif (α(i)< *b*(i)), dan untuk α(i). Ukuran keseluruhan dari pengelompokan dapat diperoleh dengan menghitung *silhouette coefficient*

rata-rata semua titik. Semakin tinggi nilai *silhouette* semakin bagus juga kekuatan dan kualitas *cluster*.

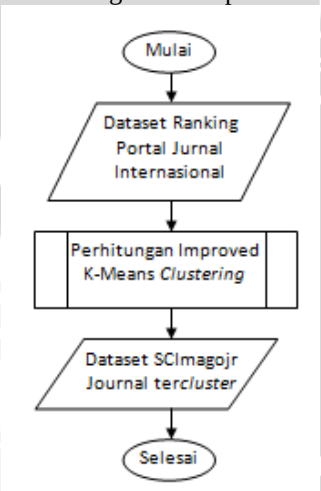
3. METODE PENELITIAN

Aliran proses dari sebuah sistem secara terstruktur sering disebut dengan Diagram Blok. Diagram Blok menjelaskan cara kerja dari sistem secara umum, mulai dari tahap inputan, proses, hingga menghasilkan keluaran (output). Secara garis besar perancangan Diagram Blok terlihat seperti pada Gambar 3.1

Gambar 3.1 Diagram Blok Sistem

Input
Masukan dalam sistem ini yaitu Dataset SCImagojr Journal yang diambil dari *The SCIMago Journal & Country Rank* dalam format *.xls.*
Process
Pada tahap pemrosesan, digunakan metode Improved K-Means dan menghitung nilai Silhouette Coefficient untuk pengujiannya.
Output
Keluaran yang dihasilkan dari sistem ini adalah Dataset SCImagojr Journal yang telah terklaster.

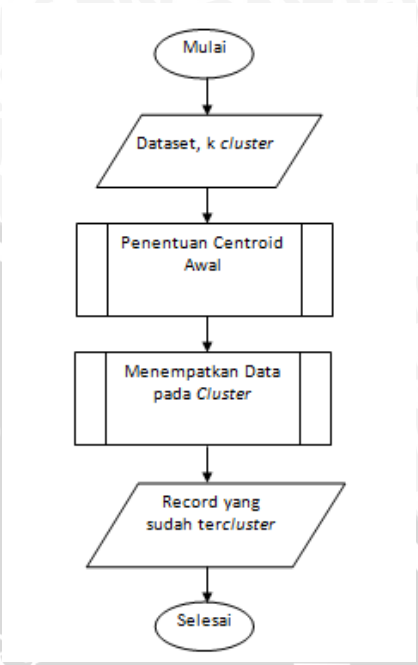
Secara umum, langkah –langkah pada algoritma *Improved K-Means* digambakan pada Gambar 3.2.



Gambar 3.2 Diagram Alir *Improved K-Means*

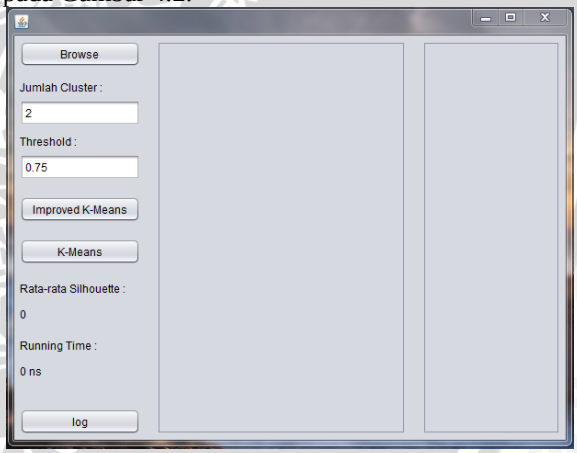
4. PERANCANGAN

Algoritma *clustering* pada penelitian ini adalah algoritma Improved K-Means. Langkah-langkah pada Improved K-Means ada dua, yaitu menemukan *centroid* awal dan menaruh data pada *cluster*. Gambar 4.1 adalah *flowchart clustering* Improved K-Means.



Gambar 4.1 *Flowchart* Improved K-Means

Adapun desain antarmuka sistem digambarkan pada Gambar 4.2.



Gambar 4.2 Desain Antarmuka Sistem

Terdapat tujuh bagian yang berfungsi dalam jalannya aplikasi, antara lain *Button Browse*, *Combo Box K*, *Combo Box threshold*, *Datagridview Dataset*, *Button Execute by* K-Means, *Button Execute by* Improved K-Means, dan *RichTextBox Result.*

5. IMPLEMENTASI

Implementasi perangkat keras yang dipakai dalam proses pembuatan sistem dijelaskan pada Tabel 5.1.

Tabel 5.1 Implementasi Perangkat Keras Komputer

Ultrabook Dell Inspiron 14R – 5420	
Nama *Hardware*	**Spesifikasi**
Processor	Intel® Core™ i3-3317U (2M Cache, up to 2.4 GHz)
Memory (RAM)	4GB DDR3
Harddisk	500GB SATA (5400RPM)
Motherboard	Dell Inc.
Graphic Card	NVIDIA® GeForce® GT 630M with 1GB GDDR5 VRAM
Monitor	14.0″ HD+ (900p) Truelife Infinity Display
Webcam	Skype-Certified Hi-Def Webcam

Implementasi perangkat lunak yang dipakai dalam proses pembuatan sistem dijelaskan pada Tabel 5.2.

Tabel 5.2 Implementasi Perangkat Lunak Komputer

Ultrabook Dell Inspiron 14R – 5420	
Nama *Software*	**Spesifikasi**
Sistem operasi	Microsoft Windows 7 Home Basic 64-bit
Bahasa pemrograman	JAVA
Tool Pemrograman	Netbeans

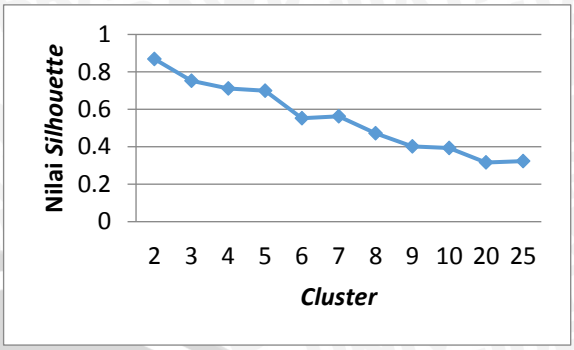
Beberapa batasan dalam mengimplementasikan perangkat lunak *clustering* jurnal internasional adalah sebagai berikut:

1. Perangkat Lunak sistem *clustering* jurnal internasional dirancang dan dijalankan dengan menggunakan *Java Apllication*.
2. File yang digunakan memiliki format *.java* dan *.xls*.

6. PENGUJIAN

Proses pengujian dilakukan dengan 3 skenario, yaitu pengujian 198 data dengan menggunakan percobaan 2 sampai 25 *cluster*, pengujian menggunakan 10, 30, 50, 80, 110, 150, 170 dan 198 data dan pengujian ketiga yaitu percobaan 2 sampai 25 *cluster* pada metode Improved K-Means kemudian dibandingkan dengan hasil *silhouette coefficient* pada pengujian K-Means Standar.

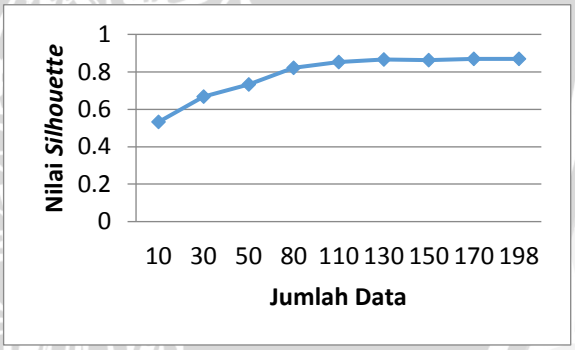
Adapun hasil penelitian dari pengujian *cluster* dengan Improved K-Means ditunjukkan pada Gambar 6.1.



Gambar 6.1 Statistik Nilai *Silhouette Cluster* 2 sampai 25

Berdasarkan grafik hasil pengujian yang ditunjukkan Gambar 6.1 dapat dilihat bahwa pengujian dengan menggunakan jumlah *cluster* sebanyak 2 sampai dengan 25 *cluster* menghasilkan nilai *silhouette* paling tinggi pada jumlah *cluster* 2, hal tersebut dipengaruhi oleh jarak kerapatan data pada satu *cluster* memiliki jarak yang cukup baik dan jarak kerapatan data *cluster* satu dengan *cluster* lainnya cukup besar.

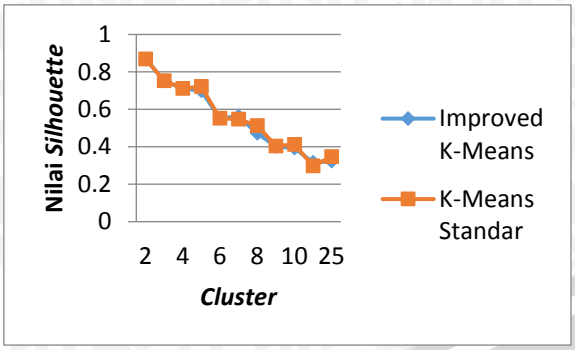
Adapun hasil penelitian dari pengujian jumlah data Improved K-Means ditunjukkan pada Gambar 6.2.



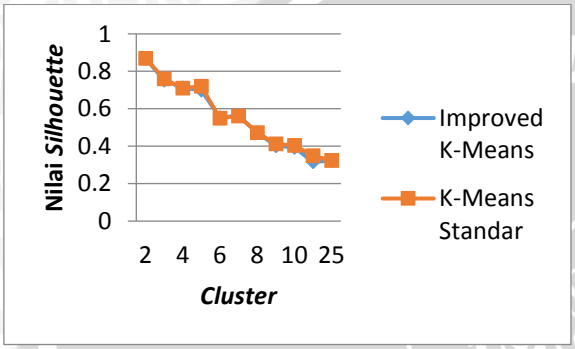
Gambar 6.2 Statistik Nilai *Silhouette* Data 10 sampai 198

Berdasarkan grafik hasil pengujian yang ditunjukkan Gambar 6.2 dapat dilihat bahwa pengujian dengan menggunakan jumlah data sebanyak 10 sampai dengan 198 data menghasilkan nilai *silhouette* paling tinggi pada jumlah data 198, hal tersebut dipengaruhi oleh jarak kerapatan data pada satu *cluster* memiliki jarak yang cukup baik dan jarak kerapatan data *cluster* satu dengan *cluster* lainnya cukup besar.

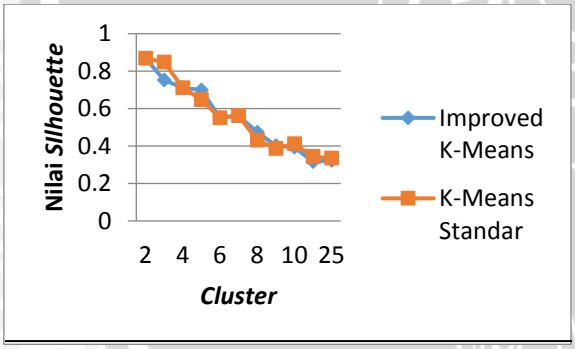
Adapun hasil penelitian dari perbandingan Improved K-Means dan K-Means Standar ditunjukkan pada Gambar 6.3 sampai 6.7.



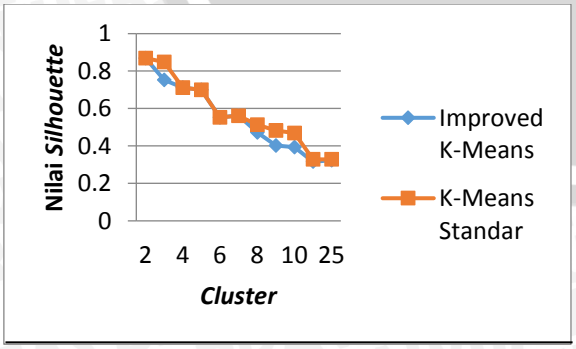
Gambar 6.3 Statistik Perbandingan Nilai *Silhouette* Pengujian *Cluster* Ke-1



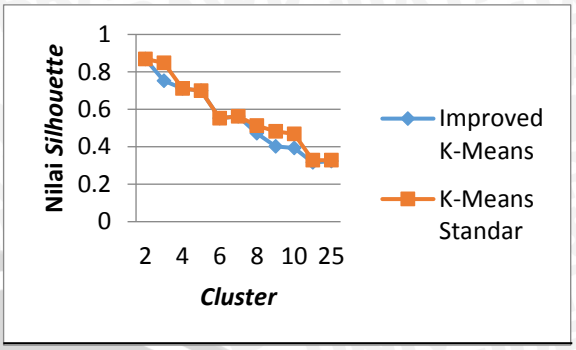
Gambar 6.4 Statistik Perbandingan Nilai *Silhouette* Pengujian *Cluster* Ke-2



Gambar 6.5 Statistik Perbandingan Nilai *Silhouette* Pengujian *Cluster* Ke-3



Gambar 6.6 Statistik Perbandingan Nilai *Silhouette* Pengujian *Cluster* Ke-4



Gambar 6.7 Statistik Perbandingan Nilai *Silhouette* Pengujian *Cluster* Ke-5

Berdasarkan grafik hasil perbandingan nilai *silhouette* pengujian *cluster* yang ditunjukkan pada Gambar 6.3 sampai Gambar 6.7 dapat dilihat bahwa pengujian jumlah *cluster* metode Improved K-Means memiliki kecenderungan nilai *silhouette* yang hampir sama terhadap nilai *silhouette* yang dihasilkan pada metode K-Means Standar pada penggunaan 2 sampai 25 *cluster*. Akan tetapi, nilai *silhouette* pada K-Means Standar tidaklah mutlak karena pada K-Means Standar inisialisasi *centroid* awal ditentukan secara acak sehingga memungkinkan memiliki nilai *silhouette* yang lebih baik ataupun lebih buruk.

7. PENUTUP

7.1 KESIMPULAN

Berdasarkan hasil perancangan, implementasi dan pengujian yang telah dilakukan, maka didapatkan kesimpulan antara lain sebagai berikut:

1. Metode Improved K-Means dapat diterapkan pada *clustering* portal jurnal internasional dimana memiliki hasil yang lebih stabil dibandingkan dengan metode K-Means Standar.
2. Metode Improved K-Means dapat di implementasikan pada *clustering* portal jurnal internasional dengan nilai kualitas *cluster* terbaik yang didapatkan adalah 0,869 untuk 198 data dengan penggunaan 2 *cluster*.
3. Penambahan dan pengurangan jumlah *cluster* pada metode Improved K-Means pada *clustering* portal jurnal internasional memiliki pengaruh terhadap nilai dari kualitas *cluster*. Pada penggunaan 2 *cluster* memiliki nilai kualitas yang paling baik daripada penggunaan jumlah *cluster* yang lain.

4. Penambahan dan pengurangan data pada Improved K-Means pada *clustering* portal jurnal internasional memiliki pengaruh terhadap meningkat atau menurunnya kualitas *cluster*. Pada penggunaan 198 data dengan 2 *cluster* memiliki nilai kualitas yang lebih baik daripada penggunaan data yang lebih sedikit.
5. Pada pengujian jumlah *cluster* metode Improved K-Means memiliki kecenderungan nilai *silhouette* yang hampir sama terhadap nilai *silhouette* yang dihasilkan pada metode K-Means Standar pada penggunaan 2 sampai 25 *cluster*. Akan tetapi, hasil pada K-Means Standar tersebut tidaklah mutlak karena pada metode K-Means Standar inisialisasi *centroid* awal ditentukan secara acak.
6. Dari hasil pengujian yang telah dilakukan sebelumnya, penggunaan metode Improved K-Means memiliki kestabilan pada nilai *silhouette* dibandingkan dengan penggunaan metode K-Means Standar. Penggunaan metode K-Means Standar memiliki kecenderungan nilai *silhouette* yang berubah-ubah karena inisialisasi *centroid* dilakukan secara acak.

7.2 SARAN

Berdasarkan hasil penelitian, penulis menyarankan agar hasil penelitian tentang Implementasi Algoritma Improved K-Means pada inisialisasi *centroid* awal untuk *clustering* portal jurnal internasional digunakan di berbagai macam database untuk membuktikan ketangguhan dan kestabilan nilai *silhouette* algoritmanya. Kemudian pada sistem didesain lebih *user friendly* agar mempermudah *user* untuk menggunakan sistem ini.

8. DAFTAR PUSTAKA

Ahmed, A.H. & Ashour, W., 2011. An Initialization Method for the K-means Algorithm using RNN and Coupling Degree. International Journal of Computer Applications, XXV(1), pp.1-6.

Al-Zoubi, Moh'd Belal, Mohammad al Rawi, 2008. An Efficient Approach for Computing *Silhouette*Coefficients, Journal of Computer Science.

Berkhin, Pavel, 2002. *Survey of Clustering Data Mining Techniques.* Accrue Software, Inc.

Carlisle, A. and Dozier, G., 2001. An Off-The Shelf PSO, Proceedings of the 2001 Worksop on Particle Swarm Optimization.pp. Indianapolis, IN.

Falagas, M. E., Kouranos, V. D., Arencibia-Jorge, R., & Karageorgopoulos, D. E., 2008. *Comparison of SCImago journal rank indicator with journal impact factor. The FASEB Journal*, *22*(8), 2623-2628.

Frand, Jason, 2009. Data Mining: What is Data Mining?. http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologite/palace/datamining.htm. Diakses pada 3 Mei 2016.

Garfield, E, 2006. The History and Meaning of The Journal Impact Factor. Journal of the American Medical Association (JAMA), (293):90-93, January 2006. (Abridged version Published).

Han, Jiawei; & Kamber, Micheline, 2001. Data Mining Concepts and Techniques Second Edition. San Francisco: Morgan Kauffman.

Giyanto, Herribertus, 2008. Penerapan Algoritma *Cluster*ing K-Means, K-Medoid, Gath Geva. Tesis Tidak terpublikasi. Yogyakarta: Universitas Gajah Mada.

Joshi, K.D & Nalwade, P.S., 2013. Modified K-Means for Better Initial *Cluster* Centres. International Journal of Computer Science and Mobile Computing, II(7), pp.219-23.

Kantardzic, Mehmed, 2003. Data Mining Concepts Models, Methods, and Algorithm. New Jersey: IEEE.

Kusrini, 2009. Algoritma Data Mining, Penerbit Andi, Yogyakarta.

Manadokota, 2013. Apakah Diseminasi Informasi Itu. http://www.manadokota.go.id/berita-1194-apakah--diseminasi--informasi--itu.html. Diakses Mei 2016.

Mathworks, 2014. Mathworks statistics toolbox: http://www.mathworks.com/help/stats/cophenet.html .

M. Lutfi Firdaus, D.Sc., 2012. Teknik Publikasi Karya Ilmiah di Jurnal Nasional dan Internasional. FKIP UNB Press.

Nazeer and Sebastian, 2009. Improving the Accuracy and Efficiency of the K-Means *Cluster*ing Algorithm. Proceedings of the

World Congress on Engineering 2009 Vol I WCE 2009, July 1 - 3, 2009, London, U.K.

Oded Maimon and Lior Rokach, 2010. Data Mining and Knowledge Discovery Handbook Second Edition. Springer Science+Business Media, LLC 2005.2010.

Osmar R. Zaiane, 1999. CMPUT690 *Principles of Knowledge Discovery in Database Chapter I: Introduction to Data Mining.*University of Alberta.

Pustakaristek, 2016. http://pustaka.ristek.go.id/main/about. Akses pada 3 Mei 2016.

Rossner M., Van Epps H., Hill E, 2007. Show me the data.J. Cell Biol.179, 1091-1092.doi: 10.1083/jcb.200711140.

Saracli et al.,2013. Comparison of Hierarchical *Cluster* Analysis Methods by Cophenetic Correlation. Journal of Inequalities and Applications 2013:203

Scimagojr, 2014. About Us. http://scimagojr.com/aboutus.php. Diakses pada 3 Mei 2016.

Scimagojr, 2016. SCImago Journal & Country Rank. http://www.scimagojr.com/countryrank.php . Diakses pada 3 Mei 2016.

Steinbach, M., G. Karypis and Vipin Kumar, 2000. A comparison of document *clustering* techniques, Minnesota: University of Minnesota, Department of Computer Science and Engineering. http://glaros.dtc.umn.edu/gkhome/fetch/papers/doc*cluster*.pdf

Tahta Alfina, Budi Santosa, dan Ali Ridho Barakbah, 2012. Analisa Perbandingan Metode Hierarchical *Cluster*ing, K-Means dan Gabungan Keduanya dalam *Cluster* Data (Studi kasus: Problem Kerja praktek Jurusan Teknik Industri ITS). Jurnal Teknik ITS Vol. 1, (Sept,2012) ISSN:

Turban, Efraim, et al., 2005. Decision Support Systems and Intelligent Systems 7th Ed. New Jersey: Pearson Education.