

ABSTRAK

Email sudah menjadi alat penting untuk berkomunikasi dalam kehidupan sehari-hari. Tingginya jumlah *email* yang masuk dapat mempersulit pengguna dalam mengelompokkan *email* secara manual. Melalui proses *clustering* dengan metode K-Means, X-Means Heuristic, dan X-Means Dynamic dibantu dengan fitur pembobotan *Ranking Term Frequency* dan TF-IDF, data *email* dapat diolah agar dapat dikelompokkan secara otomatis. *Cluster* yang terbentuk juga akan diberi label secara otomatis, agar setiap *cluster* memiliki label berdasarkan isi dari data *email* yang sudah di *clustering*. Sehingga perlu dilakukan analisis, perancangan, desain aplikasi, dan pencarian teori-teori yang mendukung untuk membuat aplikasi email clustering dengan menggunakan algoritma k-means, x-means heuristic, dan x-means dynamic. Pengujian yang dilakukan pada 20 akun email yang berbeda, ditarik 100 data *email* lalu diuji dengan 3 tahapan berbeda dan 3 metode *cluster*. Hal ini dilakukan untuk mengimplementasikan aplikasi *email clustering* agar dapat diuji tingkat keakuratannya dalam melakukan proses *email clustering*. Aplikasi *email clustering* ini dapat mengelompokkan *email* dan memberikan label secara otomatis dengan rata-rata akurasi sebesar 90,08%, menggunakan penarikan data *subject* dan *body email*.

Kata Kunci: Algoritma K-Means, Algoritma X-Means Heuristic, Algoritma X-Means Dynamic, *Ranking Term Frequency*, TF-IDF

ABSTRACT

Email has become an important tool for communicating in everyday life. The high number of incoming email may complicate the user to classify the email manually. Through the process of clustering with K-Means method, X-Means Heuristics, and X-Means Dynamic, augmented with Ranking Term Frequency weighting features and TF-IDF, emails data could be grouped automatically. Clusters that formed also be labeled automatically, so that each cluster has a label based on the contents of the email data that is already in cluster. So that needs to be done the analysis, design, application design, and search the theories that support to make the email application clustering using k-means algorithm, x-means heuristic, and x-means dynamic. Tests were performed on 20 different email accounts, then downloaded 100 emails data, and last tested with 3 different stages and three methods of cluster. This was done to implement email clustering application so that it can be tested its accuracy in the process of email clustering. This application can classify emails and give labels automatically with an average accuracy of 90.08%, using downloaded data subject and body emails.

Keywords: K-Means Algorithm, X-Means Heuristic Algorithm, X-Means Dynamic Algorithm, Ranking Term Frequency, TF-IDF

DAFTAR ISI

LEMBAR PENGESAHAN	i
PERNYATAAN ORISINALITAS LAPORAN PENELITIAN	ii
PERNYATAAN PUBLIKASI LAPORAN PENELITIAN	iii
PRAKATA	iv
ABSTRAK	v
ABSTRACT	vi
DAFTAR ISI	vii
DAFTAR GAMBAR	xi
DAFTAR TABEL	xiv
DAFTAR NOTASI/ LAMBANG	xv
BAB 1 PENDAHULUAN	1
1.1 Latar Belakang Masalah	1
1.2 Rumusan Masalah	1
1.3 Tujuan Pembahasan	2
1.4 Batasan Masalah	2
1.5 Ruang Lingkup	2
1.6 Sumber data	3
BAB 2 KAJIAN TEORI	4
2.1 Data <i>Clustering</i>	4
2.2 Algoritma <i>K-Means</i>	4
2.3 Algoritma <i>X-Means</i>	11
2.4 Pembobotan TF-IDF	12
2.5 Email	12
2.6 <i>SMTP</i>	12
2.7 <i>POP3</i>	13
2.8 <i>IMAP</i>	13
2.9 Aplikasi <i>Email Client</i> dan Dokumen <i>Clustering</i>	13
2.9.1 Microsoft Outlook	14
2.9.2 Opera Mail	14
BAB 3 ANALISIS DAN RANCANGAN SISTEM	15

3.1	Analisis.....	15
3.1.1	Preprocessing	15
3.1.2	Proses perhitungan kata.....	17
3.1.3	<i>Ranking</i>	18
3.1.4	<i>Clustering</i>	18
3.2	Gambaran Keseluruhan.....	21
3.2.1	Persyaratan Antarmuka Eksternal	21
3.2.2	Antarmuka dengan Pengguna	21
3.2.3	Antarmuka Perangkat Keras	22
3.2.4	Antarmuka Perangkat Lunak.....	22
3.2.5	Fitur-fitur Produk Perangkat Lunak	23
3.3	Disain Perangkat Lunak	31
3.3.1	Pemodelan Perangkat Lunak.....	31
3.4	Rancangan Antarmuka	57
3.4.1	<i>Form Login</i>	57
3.4.2	Form Email Client Cluster Mode	57
3.4.3	Form Send New Email	58
3.5	Rencana Pengujian	59
3.5.1	Pengambilan Data	59
3.5.2	Pembersihan Data <i>Email</i>	59
3.5.3	Proses <i>Clustering</i>	60
3.5.4	Proses <i>Auto Labeling Cluster</i>	60
3.5.5	Pembandingan <i>Directory</i> yang Terbentuk	60
BAB 4	PENGEMBANGAN PERANGKAT LUNAK.....	61
4.1	Implementasi Algoritma <i>Clustering</i>	61
4.1.1	Tokenisasi dan <i>Stopping</i>	61
4.1.2	Pembobotan.....	62
4.1.3	Pembobotan TF-IDF	63
4.1.4	Membuat Titik <i>Centroid</i>	64
4.1.5	Proses <i>Clustering</i>	64
4.1.6	<i>Clustering</i> K-Means	66
4.1.7	<i>Clustering</i> X-Means <i>Heuristic</i>	66

4.1.8	<i>Clustering X-Means Dynamic</i>	67
4.1.9	<i>Algoritma Euclidean Distance</i>	67
4.1.10	<i>Auto Labeling Cluster</i>	68
4.2	Implementasi <i>Class</i>	69
4.2.1	Class AppHelper	69
4.2.2	Class EmailAddressRegister	70
4.2.3	Class EmailVector.....	70
4.2.4	Login	70
4.2.5	Class EmailMessage.....	71
4.2.6	Class Cluster	71
4.2.7	Class EmailConnection	72
4.2.8	Class ClusteringProcess	72
4.2.9	Class SaveData.....	73
4.2.10	Class VectorSpaceModel	73
4.3	Implementasi Antar Muka.....	74
4.3.1	Implementasi Form Login.....	74
4.3.2	Implementasi Form Email Client Cluster Mode	75
4.3.3	Form Send New Email	76
BAB 5	Testing dan Evaluasi Sistem	77
5.1	Pengujian Aplikasi	77
5.1.1	Pengambilan Data <i>Email</i>	77
5.1.2	Pembuatan <i>Directory Email</i> Pada Akun Pengguna	77
5.1.3	Pembuatan Model <i>Cluster</i>	79
5.1.4	Pengujian Tahap Satu Menggunakan Data <i>Subject</i> dan <i>Body Email</i> ...	79
5.1.5	Pengujian Tahap Dua Menggunakan Data <i>Subject Email</i>	85
5.1.6	Pengujian Tahap Tiga Menggunakan Data <i>Body Email</i>	90
5.1.7	Hasil Seluruh Pengujian Tahap 1	95
5.1.8	Hasil Seluruh Pengujian Tahap 2.....	96
5.1.9	Hasil Seluruh Pengujian Tahap 3.....	98
5.1.10	Akurasi Pengujian Secara Keseluruhan	100
5.1.11	Pengujian dengan <i>Unit Testing</i>	101
BAB 6	Simpulan dan Saran.....	104

6.1	Simpulan	104
6.2	Saran.....	104
DAFTAR PUSTAKA		106
RIWAYAT HIDUP PENULIS		107

DAFTAR GAMBAR

Gambar 2.1 Diagram alir algoritma <i>K-Means</i>	6
Gambar 2.2 Contoh <i>Cluster</i>	7
Gambar 2.3 Rumus <i>Euclidean Distance</i>	8
Gambar 2.4 Jarak data dengan <i>Cluster 1</i>	8
Gambar 2.5 Jarak data dengan <i>Cluster 2</i>	9
Gambar 2.6 Metode Pembobotan TF-IDF	12
Gambar 3.1 Contoh <i>Stopping Word</i>	16
Gambar 3.2 Contoh data <i>text</i> yang belum ditokenisasi.....	17
Gambar 3.3 Rancangan <i>Use Case Diagram</i>	32
Gambar 3.4 Rancangan <i>Class Diagram</i>	38
Gambar 3.5 <i>Activity Diagram Login</i>	39
Gambar 3.6 <i>Activity Diagram Email Client</i>	40
Gambar 3.7 <i>Activity Diagram Download Email</i>	41
Gambar 3.8 <i>Activity Diagram Send Email</i>	42
Gambar 3.9 <i>Activity Diagram K-Means Clustering</i>	44
Gambar 3.10 <i>Activity Diagram X-Means Clustering</i>	44
Gambar 3.11 <i>Activity Diagram X-Means Dynamic</i>	45
Gambar 3.12 <i>Activity Diagram Save Email Cluster</i>	46
Gambar 3.13 <i>Activity Diagram Email Cluster Labeling</i>	47
Gambar 3.14 <i>Sequence Diagram Login</i>	48
Gambar 3.15 <i>Sequence Diagram Email Client</i>	49
Gambar 3.16 <i>Sequence Diagram Download Email</i>	50
Gambar 3.17 <i>Sequence Diagram Send Email</i>	51
Gambar 3.18 <i>Sequence Diagram K-Means Clustering</i>	52
Gambar 3.19 <i>Sequence Diagram X-Means Heuristic</i>	53
Gambar 3.20 <i>Sequence Diagram X-Means Dynamic</i>	55
Gambar 3.22 <i>Sequence Diagram Save Email Cluster</i>	56
Gambar 3.23 <i>Sequence Diagram Email Cluter Labeling</i>	56
Gambar 3.24 Rancangan <i>Form Login</i>	57
Gambar 3.25 Rancangan <i>Form Email Client Cluster Mode</i>	58

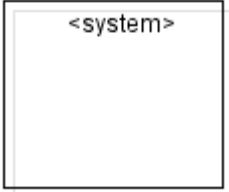
Gambar 3.26 Rancangan <i>Form Send New Email</i>	59
Gambar 4.1 <i>Class AppHelper</i>	69
Gambar 4.2 <i>Class EmailAddressRegister</i>	70
Gambar 4.3 <i>Class EmailVector</i>	70
Gambar 4.4 <i>Class Login</i>	70
Gambar 4.5 <i>Class Email Message</i>	71
Gambar 4.6 <i>Class Cluster</i>	72
Gambar 4.7 <i>Class EmailConnection</i>	72
Gambar 4.8 <i>Class ClusteringProcess</i>	73
Gambar 4.9 <i>Class SaveData</i>	73
Gambar 4.10 <i>Class VectorSpaceModel</i>	74
Gambar 4.11 Implementasi <i>Form Login</i>	75
Gambar 4.12 Implementasi <i>Form Email Client Cluster Mode</i>	76
Gambar 4.13 Implementasi <i>Form Send Email</i>	76
Gambar 5.1 <i>Directory Bimbingan</i>	77
Gambar 5.2 <i>Directory Tugas</i>	78
Gambar 5.3 <i>Directory Hack</i>	78
Gambar 5.4 <i>Directory Kerjaan</i>	78
Gambar 5.5 <i>Directory PubNub</i>	79
Gambar 5.6 Hasil pengujian tahap 1 dengan metode K-Means.....	80
Gambar 5.7 Hasil pengujian tahap 1 dengan metode X-Means <i>Heuristic</i>	82
Gambar 5.8 Hasil pengujian tahap 1 dengan metode X-Means <i>Dynamic</i>	84
Gambar 5.9 Hasil pengujian tahap 2 dengan metode K-Means.....	86
Gambar 5.10 Hasil pengujian tahap 2 dengan metode X-Means <i>Heuristic</i>	87
Gambar 5.11 Hasil pengujian tahap 2 dengan metode X-Means <i>Dynamic</i>	89
Gambar 5.12 Hasil pengujian tahap 3 dengan metode K-Means.....	91
Gambar 5.13 Hasil pengujian tahap 3 dengan metode X-Means <i>Heuristic</i>	92
Gambar 5.14 Hasil pengujian tahap 3 dengan metode X-Means <i>Dynamic</i>	94
Gambar 5.15 <i>Line Chart</i> hasil akurasi tahap 1.....	96
Gambar 5.16 <i>Line Chart</i> hasil akurasi tahap 2.....	98
Gambar 5.17 <i>Line Chart</i> hasil pengujian tahap 3	99
Gambar 5.18 <i>Chart</i> hasil keseluruhan akurasi	101

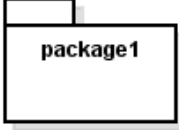
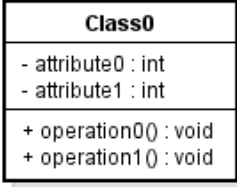
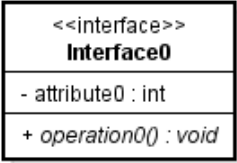
Gambar 5.19 Hasil Unit Testing dari Visual Studio	103
---	-----

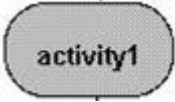
DAFTAR TABEL

Tabel 2.1 Data Sumber.....	7
Tabel 2.2 <i>Centroid</i> pada pengulangan ke-0	8
Tabel 2.3 Hasil Perhitungan Jarak	9
Tabel 2.4 Hasil perhitungan jarak dan pengelompokkan data ke-0	9
Tabel 2.5 <i>Centroid</i> pada pengulangan ke-1	10
Tabel 2.6 Hasil perhitungan jarak dan pengelompokkan data ke-1	10
Tabel 2.7 <i>Centroid</i> pada pengulangan ke-2	10
Tabel 2.8 Hasil perhitungan jarak dan pengelompokkan data ke-2	10
Tabel 2.9 <i>Centroid pada pengulangan ke-3</i>	11
Tabel 3.1 Contoh kata-kata yang sudah ditokenisasi.	17
Tabel 3.2 Contoh kata yang sudah dihitung.....	17
Tabel 3.3 Contoh <i>Top 4 Ranking</i>	18
Tabel 3.4 Deskripsi <i>use case diagram</i> untuk fitur <i>login email</i>	32
Tabel 3.5 Deskripsi <i>use case diagram</i> untuk fitur <i>Open email client</i>	33
Tabel 3.6 Deskripsi <i>use case diagram</i> untuk fitur <i>download email</i>	33
Tabel 3.7 Deskripsi <i>use case diagram</i> untuk fitur <i>send email</i>	34
Tabel 3.8 Deskripsi <i>use case diagram</i> untuk fitur <i>email clustering with K-Means</i> ...	34
Tabel 3.9 Deskripsi <i>use case diagram</i> untuk fitur <i>email clustering with X-Means</i> <i>Heuristic</i>	35
Tabel 3.10 Deskripsi <i>use case diagram</i> untuk fitur <i>email clustering with X-Means</i> <i>Dynamic</i>	36
Tabel 3.11 Deskripsi <i>use case diagram</i> untuk fitur <i>save email cluster</i>	36
Tabel 3.12 Deskripsi <i>use case diagram</i> untuk fitur <i>load email cluster</i>	37
Tabel 5.1 Hasil Pengujian Tahap 1	95
Tabel 5.2 Hasil pengujian tahap 2.....	97
Tabel 5.3 Hasil pengujian tahap 3.....	98
Tabel 5.4 Hasil Akurasi Keseluruhan	100

DAFTAR NOTASI/ LAMBANG

Use Case Diagram (UML 2.0)		
No	Gambar	Keterangan
1.		Menggambarkan aktor atau pengguna aplikasi.
2.		Menggambarkan proses atau aksi yang dapat dilakukan oleh aktor pada aplikasi.
3.		Menggambarkan sistem tempat proses dijalankan

Class Diagram (UML 2.0)		
No	Gambar	Keterangan
1.		Menggambarkan paket tempat menyimpan sekumpulan kelas
2.		Menggambarkan sebuah kelas beserta atribut dan <i>method</i> -nya
3.		Menggambarkan sebuah <i>interface</i> beserta atribut dan <i>method</i> -nya

Activity Diagram (UML 1.3)		
No	Gambar	Keterangan
1.		Menandakan dimulainya aktivitas pada sebuah sistem.
2.		Menandakan aktivitas apa yang akan dilakukan oleh pengguna aplikasi.
3.		Menandakan akhir aliran proses sistem