SKRIPSI

PENGELOMPOKAN DOKUMEN BERBASIS PSO

Siti Khalishah Ulfah

NPM: 2013730016

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS
UNIVERSITAS KATOLIK PARAHYANGAN
2019

UNDERGRADUATE THESIS

PSO-BASED DOCUMENT CLUSTERING

Siti Khalishah Ulfah

NPM: 2013730016

DEPARTMENT OF INFORMATICS
FACULTY OF INFORMATION TECHNOLOGY AND SCIENCES
PARAHYANGAN CATHOLIC UNIVERSITY
2019

LEMBAR PENGESAHAN

PENGELOMPOKAN DOKUMEN BERBASIS PSO

Siti Khalishah Ulfah

NPM: 2013730016

Bandung, 15 Mei 2019

Menyetujui,

Pembimbing

Kristopher David Harjono, M.T.

Ketua Tim Penguji	Anggota Tim Penguji
Natalia, M.Si.	**Dr.rer.nat. Cecilia Esti Nugraheni**

Mengetahui,

Ketua Program Studi

Mariskha Tri Adithia, P.D.Eng

PERNYATAAN

Dengan ini saya yang bertandatangan di bawah ini menyatakan bahwa skripsi dengan judul:

PENGELOMPOKAN DOKUMEN BERBASIS PSO

adalah benar-benar karya saya sendiri, dan saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan.

Atas pernyataan ini, saya siap menanggung segala risiko dan sanksi yang dijatuhkan kepada saya, apabila di kemudian hari ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya, atau jika ada tuntutan formal atau non-formal dari pihak lain berkaitan dengan keaslian karya saya ini.

Dinyatakan di Bandung,
Tanggal 15 Mei 2019

Meterai
Rp. 6000

Siti Khalishah Ulfah
NPM: 2013730016

ABSTRAK

Analisis *cluster* berasal dari antropologi oleh Driver dan Kroeber pada tahun 1932 dan pertama kali diperkenalkan ke bidang psikologi oleh Joseph Zubin pada tahun 1938 dan Robert Tryon pada tahun 1939. Analisis *cluster* sendiri sampai masa kini telah dimanfaatkan dalam banyak bidang seperti pada bidang *machine learning*, *face recognition*, grafik komputer, kompresi data, analisis gambar, bioinformatika, *marketing*, dsb.

Analisis *cluster* memiliki pengertian sebagai proses untuk mengelompokkan kumpulan objek, di mana kumpulan objek pada *cluster* yang sama memiliki kemiripan satu sama lain yang lebih tinggi dibanding objek pada *cluster* lainnya. Salah satu penerapan dari analisis *cluster* adalah dalam memproses kumpulan dokumen berbasis teks.

Sebelum dapat melakukan analisis *cluster* pada skripsi ini, dilakukan terlebih dahulu *text preprocessing*. *Text preprocessing* merupakan proses mengubah bentuk data yang belum terstruktur menjadi data terstruktur. Kumpulan dokumen yang digunakan pada skripsi ini memiliki banyak variasi kata pada tiap dokumennya. Untuk dapat memproses kumpulan dokumen yang memiliki banyak variasi kata pada tiap dokumennya, tentu diperlukan suatu struktur agar dapat dilakukan proses perhitungan dengan tepat dan efisien.

Setelah dilakukkannya *text preprocessing* pada skripsi ini dibangun perangkat lunak yang dapat mengimplementasikan analisis *cluster* pada dokumen berbasis teks menggunakan metode *Particle Swarm Optimization*. PSO adalah teknik optimisasi stokastik berbasis populasi yang dikembangkan oleh Dr. Eberhart dan Dr. Kennedy pada tahun 1995, terinspirasi oleh perilaku sosial burung, dan atau ikan. Perangkat lunak yang dibangun memiliki tujuan untuk menganalisis tingkat efektivitas yang diciptakan pada proses analisis *cluster* menggunakan algoritma *Particle Swarm Optimization*. Kemudian untuk membandingkan tingkat efektifitas dari algoritma PSO dirancang juga perangkat lunak lainnya. Pada skripsi ini digunakan algioritma *K-means* sebagai pembanding.

Kata-kata kunci: Particle Swarm Optimization, Clustering, K-means

ABSTRACT

Cluster analysis is a process for grouping a collection of objects, where a collection of objects on the same cluster has similarities to each other which are higher than other objects on cluster. One of the application using cluster analysis is in documents text-based.

Before the data can be clustered, the data need to be process by text preprocessing first. Text preprocessing is the process of changing the form of unstructured data into structured data. The collection of document used in this thesis has a variety of words in each document. To be able to process a document that has a lot of word variations on each document, of course a structure is needed so that the process of calculating properly and efficiently can be done.

In this thesis, designed a software that implement analysis of cluster on text-based documents using the Particle Swarm Optimization method. PSO is a population-based stochastic optimization technique developed by Dr. Eberhart and Dr. Kennedy in 1995, inspired by the social behavior of birds, and or fish. The software that was built has the aim to analyze the level of effectiveness created in the process of analysis cluster using the algorithm Particle Swarm Optimization. Then to compare the effectiveness of the PSO algorithm, the writer of this thesis also build a software to implement K-means algorithm.

Keywords: Particle Swarm Optimization, Clustering, K-means

Kepada kedua orang tua tersayang.

KATA PENGANTAR

Pertama tentu saja saya ingin mengucapkan terima kasih kepada Mama dan Papa saya yang telah membesarkan saya. Terima kasih juga kepada Kakak dan Abang saya, yang keduanya telah menjadi panutan saya sejak kecil.

Terima kasih kepada teman saya sejak SMP hingga kini, Nisa, Shinta, Nindia, Aini, Dea, Fitri, Gaby (terutama atas obrolan obrolan *update* yang menarik), dan lainnya yang mungkin tidak dapat saya sebutkan satu persatu.

Terima kasih kepada Rachael, Vica, Glorya, Gavrila, Jessica, Ilham, Fadel, Mesa, Deta, David, Renal, banyak bgt gasi? Terima kasih kepada IT UNPAR 13.

Terima kasih sangat teramat kepada Media Parahyangan, terutama Axel, Katya, Devina, Oni, Naning, Vincent, Zico, Tanya, Fiqih, Nisa, dan semua anggota lainnya yang telah memberi saya sangat teramat banyak pelajaran yang sangat bermakna (melebihi pelajaran dikelas).

Terima kasih kepada Distra, Nancy, Ka Meke, Ninet, yang sempat menjadi teman sekaligus rekan kerja yang tidak terlupakan pengalamannya (pahit tapi luar biasa mempengaruhi karir ku hingga kini) sekali lagi terima kasih banyak atas segala pelajarannya. Terima kasih juga kepada Ijal yang telah banyak membagi ilmunya.

Terima kasih kepada Dedi yang telah menemani 80% pengerjaan skripsi ini, baik secara langsung, telfon, maupun lewat doa (geer bgt gasi gw padahal ga didoain).

Terima kasih kepada semua orang yang pernah maupun masih men-*support* saya.

Terima kasih juga kepada diri saya sendiri, HEHE.

Sekian dan terima kasih telah membaca. Mohon maaf jika ada kesalahan kata pada skripsi ini.

Bandung, Mei 2019

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Berkembangnya teknologi pada masa kini mempengaruhi tingkat keberagaman dari informasi yang beredar. Meningkatnya keberagaman informasi ini dapat dimanfaatkan untuk mengoptimalkan sistem *marketing*, proses bisnis, dan sebagainya. Salah satu pengembangan dalam memanfaatkan keberagaman informasi ini dapat dilakukan dengan analisis *cluster*.

Analisis *cluster* berasal dari antropologi oleh Driver dan Kroeber pada tahun 1932 dan pertama kali diperkenalkan ke bidang psikologi oleh Joseph Zubin pada tahun 1938 dan Robert Tryon pada tahun 1939. Analisis *cluster* sendiri sampai masa kini telah dimanfaatkan dalam banyak bidang seperti pada bidang *machine learning*, *face recognition*, grafik komputer, kompresi data, analisis gambar, bioinformatika, *marketing*, dsb. Dengan banyaknya data yang beredar, tentunya meningkat juga informasi yang dapat diproses. Dilakukannya proses analisis *cluster* diharapkan dapat meningkatkan efesiensi dan efektivitas pemrosesan data dibandingkan melakukan pengelompokan atau analisis secara manual.

Analisis *cluster* memiliki pengertian sebagai proses untuk mengelompokkan kumpulan objek, di mana kumpulan objek pada *cluster* yang sama memiliki kemiripan satu sama lain yang lebih tinggi dibanding objek pada *cluster* lainnya. Pada skripsi ini, dibangun program untuk melakukan analisis *cluster* pada kumpulan dokumen berbasis teks yang selanjutnya akan disebut korpus. Korpus yang digunakan merupakan kumpulan artikel bahasa Indonesia yang memiliki beragam topik. Pada skripsi ini dibangun perangkat lunak yang dapat melakukan analisis *cluster* untuk mengelompokkan kumpulan dokumen berbasis teks berdasarkan kemiripan topik yang dimiliki pada setiap dokumen.

Sebelum dapat melakukan analisis *cluster* pada skripsi ini, dilakukan terlebih dahulu *text preprocessing*. *Text preprocessing* merupakan proses mengubah bentuk data yang belum terstruktur menjadi data terstruktur. Kumpulan artikel yang digunakan pada skripsi ini memiliki banyak variasi kata pada tiap dokumennya. Untuk dapat memproses kumpulan dokumen yang memiliki banyak variasi kata pada tiap dokumennya, tentu diperlukan suatu struktur agar dapat dilakukan proses perhitungan dengan tepat dan efisien.

Maka dari itu, sebelum dapat melakukan analisis *cluster* kumpulan dokumen berbasis teks yang memiliki banyak variasi kata ini perlu diproses terlebih dahulu menggunakan proses yang disebut *text preprocessing*. Pada *text preprocessing*, data yang diproses akan tersimpan dalam bentuk vektor. Pada skripsi ini, dimensi dari vektor yang dihasilkan akan sangat banyak, yaitu sejumlah variasi kata yang ada pada keseluruhan korpus. Untuk dapat melakukan analisis *cluster* pada data yang didapat melalui *text preprocessing*, pada skripsi ini digunakan algoritma *Particle Swarm Optimization*.

PSO merupakan teknik optimisasi stokastik berbasis populasi yang dikembangkan oleh Dr. Eberhart dan Dr. Kennedy pada tahun 1995, terinspirasi oleh perilaku sosial burung, dan atau ikan. Selain untuk melakukan analisis *cluster* menggunakan data berupa kumpulan dokumen berbasis teks, dibangunnya perangkat lunak pada skripsi ini memiliki tujuan untuk menganalisis tingkat efektivitas pada proses analisis *cluster* menggunakan algoritma *Particle Swarm Optimization*. Kemudian untuk membandingkan tingkat efektivitas dari algoritma PSO dirancang juga perangkat lunak lainnya. Pada skripsi ini, digunakan algoritma *K-means* sebagai pembanding.

K-means merupakan metode yang dikenal luas dan mudah untuk diimplementasikan, namun memiliki masalah pada saat menginisialisasi titik pusat *cluster*.

1.2 Rumusan Masalah

Berdasarkan latar belakang, rumusan masalah pada skripsi ini adalah sebagai berikut:

1. Bagaimana cara merepresentasikan dokumen sehingga dapat diproses menggunakan *Particle Swarm Optimization*?
2. Bagaimana perbandingan hasil kinerja dari *Particle Swarm Optimization* dibandingkan dengan *K-means*?
3. Apa fungsi *fitness* yang akan digunakan dalam implementasi *Particle Swarm Optimization*?

1.3 Tujuan

Berdasarkan rumusan masalah, maka tujuan dari skripsi ini adalah sebagai berikut:

1. Mempelajari proses *preprocessing* pada dokumen berbasis teks.
2. Mempelajari algoritma *Particle Swarm Optimization* dan *K-means*.
3. Mempelajari fungsi *fitness* yang tepat untuk diimplementasikan pada algoritma *Particle Swarm Optimization*.
4. Membangun perangkat lunak yang dapat melakukan *preprocessing* pada dokumen berbasis teks.
5. Mencari tahu dara mengembangkan algoritma PSO untuk dapat diimplementasikan pada analisis *cluster*
6. Membangun perangkat lunak yang dapat melakukan *clustering* menggunakan algoritma *Particle Swarm Optimization* dan *K-means*.
7. Membandingkan kinerja *clustering* menggunakan algoritma *Particle Swarm Optimization* dan *K-means*.

1.4 Batasan Masalah

Batasan-batasan masalah dalam skripsi ini adalah sebagai berikut:

1. Dokumen yang diproses merupakan dokumen .txt
2. Isi dari dokumen yang diproses berbahasa Indonesia

1.5 Metodologi

Metodologi yang digunakan dalam penyusunan skripsi ini adalah sebagai berikut:

1. Melakukan studi literatur mengenai *clustering* menggunakan algoritma *K-means*.
2. Melakukan studi literatur mengenai algoritma *Particle Swarm Optimization*.
3. Melakukan studi literatur *preprocessing*.

4. Mengimplementasikan *preprocessing* dokumen.
5. Melakukan perancangan kelas yang akan digunakan untuk mengimplementasikan *preprocessing* dan clustering dengan menggunakan algoritma *K-means* dan algoritma *Particle Swarm Optimization*.
6. Mengimplementasikan hasil perancangan kelas ke dalam bahasa pemograman *Java*.
7. Melakukan pengujian fungsional dan pengujian eksperimental terhadap perangkat lunak yang dibangun.
8. Menarik kesimpulan berdasarkan hasil pengujian.

1.6 Sistematika Pembahasan

Skripsi ini akan tersusun dalam enam bab secara sistematis. Enam bab tersebut terdiri dari pendahuluan, dasar teori, analisis, perancangan, implementasi dan pengujian, dan kesimpulan. Berikut merupakan sistematika pembahasan dalam skripsi ini.

1. Bab 1 Pendahuluan

   Bab ini berisi latar belakang, rumusan masalah, tujuan, batasan masalah, metodologi, dan sistematika pembahasan.

2. Bab 2 Dasar Teori

   Bab 2 berisi dasar teori mengenai *preprocessing* pada dokumen berbasis teks, *Particle Swarm Optimization*, analisis *cluster* menggunakan *Particle Swarm Optimization*.

3. Bab 3 Analisis

   Bab 3 berisi analisis masalah, analisis *preprocessing*, analisis *cluster* menggunakan metode *Particle Swarm Optimzation*.

4. Bab 4 Perancangan

   Bab 4 berisi perancangan perangkat lunak yang dibangun, meliputi kebutuhan masukan dan keluaran, rancangan antarmuka, diagram kelas rinci, dan rincian metode yang digunakan pada perangkat lunak.

5. Bab 5 Implementasi dan Pengujian

   Bab 5 berisi implementasi antarmuka perangkat lunak, pengujian fungsional terhadap perangkat lunak yang mengimplementasikan *clustering* menggunakan algoritma *K-means* dan *Particle Swarm Optimization*, serta pengujian eksperimental terhadap perangkat lunak yang mengimplementasikan *clustering* menggunakan algoritma *K-means* dan *Particle Swarm Optimization*.

6. Bab 6 Kesimpulan dan Saran

   Bab 6 berisi kesimpulan dari awal hingga akhir skripsi beserta saran untuk pengembangan selanjutnya.