

## INTISARI

### **PENGELOMPOKAN DOKUMEN BERDASARKAN METADATA MENGGUNAKAN ALGORITME *K-MEANS* DENGAN METODE DENSITAS DAN PROBABILITAS**

Oleh

Arya Kresna Wijaya  
14/364160/PA/15921

Banyaknya dokumen digital dengan berbagai jenis tipe menimbulkan permasalahan pencarian informasi pada dokumen jenis tertentu. Proses pencarian informasi tersebut dapat dipermudah dengan mengelompokkan dokumen. *Clustering* merupakan salah satu metode yang dapat digunakan untuk mengelompokkan dokumen. Akan tetapi, dokumen dengan jenis berbeda memiliki jenis fitur internal yang berbeda sehingga fitur internal tersebut tidak dapat digunakan sebagai fitur untuk mengelompokkan dokumen. Metadata sebagai fitur eksternal yang dimiliki oleh setiap jenis dokumen dapat digunakan sebagai fitur alternatif untuk mengelompokkan dokumen dengan jenis yang berbeda.

Pada penelitian ini proses *clustering* dilakukan menggunakan algoritme *K-means* dengan metode densitas dan probabilitas pada proses pemilihan pusat kluster awal. Dokumen yang digunakan pada penelitian ini adalah dokumen pdf, docx, jpg, dan png. Hasil penelitian menunjukkan bahwa proses *clustering* dengan menggunakan metode probabilitas memiliki performa yang lebih baik dibandingkan dengan metode densitas. Hal tersebut ditunjukkan dari rata-rata nilai SSE metode probabilitas yang lebih rendah dibandingkan metode densitas yaitu sebanyak 2,9951, sedangkan nilai rata-rata SSE dengan metode densitas sebanyak 3,7509. Selain itu, dokumen docx dan jpg cenderung mengelompok ke dalam kluster yang sama dikarenakan memiliki kesamaan atribut metadata yang banyak, sedangkan dokumen pdf dan png mengelompok membentuk kluster yang berbeda.

**Kata kunci:** Dokumen, *Clustering*, Metadata, Algoritme *K-means*

## **ABSTRACT**

### **DOCUMENT CLUSTERING BASED ON METADATA USING K-MEANS ALGORITHM WITH DENSITY AND PROBABILITY METHOD**

by

Arya Kresna Wijaya  
14/364160/PA/15921

The number of digital documents with various types causes problems of information retrieval in a particular type of document. The information retrieval process can be made easier by grouping documents. Clustering is one method that can be used to group documents. However, documents with different types have different types of internal features so that they can not be used as a feature for grouping documents. Metadata as an external feature owned by each document type can be used as an alternative feature to group documents with different types.

In this research, the process of clustering is done using K-means algorithm with density and probability methods in the process of selecting the initial cluster centers. The documents that used in this research are pdf, docx, jpg, and png documents. The results showed that the clustering process using the probability method has better performance than the density method. This is shown by the average SSE value of the lower probability method than the density method of 2.9951, whereas the average value of SSE with the density method is 3.7509. In addition, docx and jpg documents tend to group into the same cluster because they have many common metadata attributes, whereas pdf and png documents tend to form different clusters.

**Keywords:** Document, Clustering, Metadata, K-means Algorithm