ABSTRAK

Nama : Septian Arif Maulana

Nim : 1167020062

Judul : Algoritma *Clustering K-Means++* pada Data Terjemahan Hadis

Perkembangan teknologi mengakibatkan ketersediaan data yang semakin meningkat, peningkatan jumlah data ini juga terjadi pada data hadis. Dibutuhkan solusi matematis untuk memanfaatkan dan menggali informasi yang terdapat dalam data tersebut. Algoritma *clustering K-Means++* merupakan salah satu metode yang dapat digunakan untuk mengatasi permasalahan ini. Pada penelitian ini digunakan algoritma *clustering K-Means++* yang dipadukan dengan dua *proximity measure* berbasis *dissimilarity* yaitu *cosine distance* dan *euclidean distance*. Salah satu masalah utama yang harus diperhatikan dalam proses *clustering* adalah dimensi ruang fitur yang tinggi. Reduksi dimensi dapat dijadikan sebagai salah satu langkah optimasi pada algoritma *clustering* untuk mengurangi jumlah fitur (dimensi). Pada penelitian ini akan digunakan metode *Principal Component Analysis* (PCA) untuk mereduksi fitur-fitur yang kurang berpengaruh dan *redudant* tanpa mengurangi karakteristik data tersebut secara signifikan. Selanjutnya dilakukan perbandingan hasil *clustering* pada data terjemahan hadis yang berjumlah 5344 hadis menggunakan algoritma *K-Means++* dengan masing-masing kombinasi *dissimilarity measure* tanpa reduksi dan menggunakan reduksi untuk mengetahui hasil *clustering* yang terbaik pada algoritma tersebut. Hasil *clustering* yang didapat setelah dilakukan evaluasi menggunakan *Davies-Bouldin Index* (DBI) dan *Silhouette Coefficient* (SC) menunjukkan *clustering* algoritma *K-Means++* dengan kombinasi *cosine distance* menggunakan reduksi mempunyai hasil yang lebih baik jika dibandingkan dengan hasil masing-masing algoritma *clustering K-Means++* dengan kombinasi *dissimilarity measure* tanpa reduksi dan menggunakan reduksi dengan nilai DBI terbaik 1,61217321 dan SC terbaik 0,02996795. Dengan demikian, kombinasi *proximity measure* menggunakan *cosine distance* dan reduksi dimensi menggunakan PCA dalam proses *text clustering* dapat meningkatkan akurasi *clustering*.

Kata Kunci: *K-Means++, Proximity, Principal Component Analysis, Davies-Bouldin Index, Silhouette Coefficient*

ABSTRACT

Name **: Septian Arif Maulana**

Nim : 1167020062

Title **: Algoritma *Clustering K-Means++* pada Data Terjemahan Hadis**

Technological developments have resulted in increased availability of data, this increase in the amount of data also occurs in hadith data. Mathematical solutions are needed to utilize and explore the information contained in the data. Algorithm K-Means++ clustering is one of the methods in text clustering that can be used to overcome these problems. In this study, the K-Means++ clustering which is combined with two proximity measures -based dissimilarity namely cosine distance and euclidean distance. One of the main problems that must be considered in the clustering is the high dimension of the feature space. Dimensional reduction can be used as an optimization step in the clustering to reduce the number of features (dimensions). Method will be used Principal Component Analysis to reduce the features that are less influential and redundant without significantly reducing the characteristics of the data. Furthermore, a comparison of the results of clustering on the hadith translation data which amounted to 5344 hadiths was carried out using the K-Means++ algorithm with each combination of dissimilarity measures without reduction and using reduction to find out the clustering in the algorithm. The clustering obtained after an evaluation using the Davies-Bouldin Index (DBI) and Silhouette Coefficient (SC) show that clustering algorithm the K-Means++ combination cosine distance using reduction has better results when compared to the results of each K-Means++ clustering with a combination of dissimilarity measure without reduction and using reduction with the best DBI value 1.61217321 and the best SC 0.02996795. Thus, the combination of proximity measure using cosine distance and dimension reduction using PCA in the text clustering accuracy clustering.

Keywords*: K-Means++, Proximity, Principal Component Analysis, Davies-Bouldin Index, Silhouette Coefficient*