INTISARI

ALGORITME *CLUSTERING K-MEANS* DENGAN *SEMANTIC SIMILARITY* UNTUK MEMPERKIRAKAN BIAYA RAWAT INAP

Oleh
IDA BAGUS GEDE SARASVANANDA
16/403687/PPA/05204

Besar biaya rawat inap dari seorang pasien dapat diperkirakan dengan melakukan *cluster* pasien. Salah satu algoritme yang banyak digunakan untuk *clustering* adalah *K-means*. Algoritme *K-means* berbasiskan *distance* masih memiliki kelemahan dalam hal mengukur kedekatan makna atau semantik antar data. Padahal, dalam memperkirakan biaya rawat inap pasien, perlu diperhatikan juga kemiripan penyakit pasien, yang dapat dilihat dari kode ICD penyakit pasien. Untuk mengatasi permasalahan tersebut dapat digunakan *semantic similarity* untuk mengukur similaritas antar objek pada *clustering* sehingga kedekatan secara semantik dapat diperhitungkan.

Penelitian ini bertujuan untuk melakukan *clustering* terhadap data pasien dengan memperhatikan kemiripan penyakit pasien. Kode ICD digunakan sebagai pedoman dalam menentukan penyakit pasien. Metode *K-means* digabungkan dengan *semantic similarity* untuk mengukur kedekatan kode ICD pasien. Metode yang digunakan untuk pengukuran kemiripan semantik antar data dalam penelitian ini yaitu *semantic similarity* Girardi, Leacock & Chodorow, Rada, dan *Jaccard Similarity*. Pengukuran kualitas *cluster* menggunakan metode *silhouette coefficient*.

Berdasarkan hasil eksperimen, metode pengukuran data *semantic similarity* mampu manghasilkan kualitas hasil *clustering* yang lebih baik dibandingkan dengan tanpa *semantic similarity*. Akurasi terbaik adalah 91,78% untuk ketiga metode *semantic similarity*. Dari ketiga metode *semantic similarity* yang digunakan, metode *semantic similarity* Rada memiliki ukuran kualitas *cluster* yang lebih baik dibandingkan dengan *semantic similarity* Leacock & Chodorow, dan Girardi.

Kata Kunci: *Clustering* pasien, *K-means*, *Semantic Similarity*, *Girardi, Leacock & Chodorow, Rada, Jaccard Similarity*

ABSTRACT

THE K-MEANS CLUSTERING ALGORITHM WITH SEMANTIC SIMILARITY TO ESTIMATE THE COST OF HOSPITALIZATION

By
IDA BAGUS GEDE SARASVANANDA
16/403687/PPA/05204

The cost of hospitalization from a patient can be estimated by performing a cluster of patient. One of the algorithms that is widely used for clustering is K-means. However, K-means algorithm, based on distance still has weaknesses in terms of measuring the proximity of meaning or semantics between data. In fact, in estimating the cost of hospitalization for patient, it should be noted also the similarity of patient's disease, which can be seen from the ICD code of the patient's disease. To overcome this problem, semantic similarity can be used to measure the similarity between objects in clustering, so that, semantic proximity can be calculated.

This study aims to conduct clustering of patient data by paying attention to the similarity of the patient's disease. ICD code is used as a guide in determining a patient's disease. The K-means method is combined with semantic similarity to measure the proximity of the patient's ICD code. The method used to measure the semantic similarity between data, in this study, is the semantic similarity of Girardi, Leacock & Chodorow, Rada, and Jaccard Similarity. Cluster quality measurement uses the silhouette coefficient method.

Based on the experimental results, the method of measuring semantic similarity data is capable to produce better quality clustering results than without semantic similarity. The best accuracy is 91,78% for the three semantic similarity methods. From the three semantic similarity methods used, the method of semantic similarity Rada has a quality measure that is better than the semantic similarity of Leacock & Chodorow, and Girardi.

Keywords: Patient clustering, K-means, Semantic Similarity, Girardi, Leacock & Chodorow, Rada, Jaccard Similarity