IJCCS (Indonesian Journal of Computing and Cybernetics Systems)
Vol.13, No.4, October 2019, pp. 313~322
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: https://doi.org/10.22146/ijccs.45093

The K-Means Clustering Algorithm With Semantic Similarity To Estimate The Cost of Hospitalization

Ida Bagus Gede Sarasvananda*[1], Retantyo Wardoyo[2], Anny Kartika Sari[3]
[1]Master Program of Computer Science, FMIPA UGM, Yogyakarta, Indonesia
[2,3]Department of Computer Science and Electronics, FMIPA UGM, Yogyakarta, Indonesia
e-mail: ***[1]sarasvananda@mail.ugm.ac.id**, [2]rw@ugm.ac.id, [3]a_kartikasari@ugm.ac.id

Abstrak

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. 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 jaccard similarity. Akurasi terbaik adalah 91,78% untuk ketiga metode semantic similarity sedangkan jaccard similarity memiliki akurasi terbaik 84,93%.

Kata kunci— *Clustering pasien, K-means, Semantic Similarity, Sillhoutte Coefficient*

Abstract

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. K-means algorithm, based on distance still has weaknesses in terms of measuring the proximity of meaning or semantics between data. 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, whereas without semantic similarity the best accuracy is 84.93%.

Keywords— *Clustering pasien, K-means, Semantic Similarity, Sillhoutte Coefficient*

Received April 16th,2019; Revised October 10th, 2019; Accepted October 14th, 2019

1.INTRODUCTION

Clustering is a process of grouping data into groups or clusters, where each cluster has data that has high similarities and between clusters has a low similarity [1]. Measure of distance to measure data similarity has a very important role in the performance of the K-means algorithm [2]. The measurement of similarity, in the K-means algorithm based on distance, still has several weaknesses, such as less attention to the semantic meaning between data. To overcome this problem, actually, semantic similarity can be applied to measure the similarity between objects in clustering, so that semantic proximity will be taken into account. Measurement of similarity with semantic similarity can be conducted using ontology, i.e. by measuring the distance between concepts on ontology.

Some researchers have conducted research to address the problem of distance-based proximity measurement with semantics [3.4]. The data used in this study are only types of text-based data and have not been able to accommodate types of categorical data represented by hierarchical model. For data that has types of categorical data with hierarchical model, it can be measured using semantic similarity equation proposed by Girardi et al. [5], Leacock & Chodorow [6], and Rada et al. [7]. An example of data with a category type with a hierarchical model is the ICD-10, i.e. an international standard for classifying diseases and other health problems. In computer science, the ICD can be considered as ontology in a simple form, where the importance is the hierarchy of concept. Ontology in this form is often referred to as terminology.

ICD-10 is used in hospitals as a guideline to determine the code of the patient's disease type. The similarity of the disease from the patient can be seen from the proximity of the patient's disease code on ICD-10. The type of patient's disease is one of the factors that determines the cost of hospitalization from patient. As is known, each patient who will conduct an examination to the hospital, can visit the Emergency Installation Unit (IGD) for patients who are in an emergency, or Polyclinic unit for patients who are not in an emergency. Medical personnel will conduct clinical, laboratory, and supporting examinations to establish the diagnosis, initial planning of patient management, and conclusion whether the patient will be hospitalized or not. If the patient is declared to be hospitalized, the medical staff will provide a financial estimate to the patient's family, so that the patient's family will know the estimated cost needed by the patient. Diagnostic Related Group (DRG) can be simplified by means of payments with unit costs per diagnosis, but not unit costs per type of medical or non-medical services provided to patient [8]. Estimates of patient costs can be conducted by clustering patient data that includes data of disease diagnosis, age, sex, and inpatient class rates.

In this study, patient clustering was carried out, so that patients can be grouped according to similarities in features. The method used in measuring data with centroid is the semantic similarity of Girardi et al., Leacock & Chodorow, and Rada et al. to measure diagnostic features that have been coded with the ICD-10 and Euclidean distance for features of gender, age, and class rates. Clustering patients can help management or medical personnel from the hospital as a consideration in the grouping of DRG (Diagnosis-related Group) to determine the financing of health services.

2.METHODS

2.1 Data Collection

The amount of data used is 244 patient data. Patient data were divided into training data and test data. 171 patient data were used as training data and 73 patient data as testing data. Data used include patient diagnosis that has been coded with international standard for classification of diseases and other health problems, namely ICD-10, class rates, age, and gender.

2.2 Data Normalization

There are two methods for data normalization, namely: range and var methods. Range method is a method that normalizes existing data, so that it has a value between 0 and 1. In this study, the method used for data normalization is the range method with the following formula:

$$X_{baru} = \frac{X_{lama} - X_{min}}{X_{max} - X_{min}} \quad (1)$$

In each initial data column, the data that are searched for have minimum and maximum values. The minimum data were saved to the min data variable, while the maximum data were saved to the data max variable. Data normalization was conducted to normalize the data, so that it has a value between 0 and 1. The parameters that would be normalized are the parameters of the inpatient class and the age of the patient using equation (1).

2.3 Model Design

In order to develop the architectural model and conceptual design that will be developed, it needs the stages of needs analysis, both in the form of data analysis needs and function requirements analysis. The stages of system analysis will provide an understanding of the system that will be developed, and to find the shortcomings of the system to be developed, so as to produce a better system and in accordance with user needs.

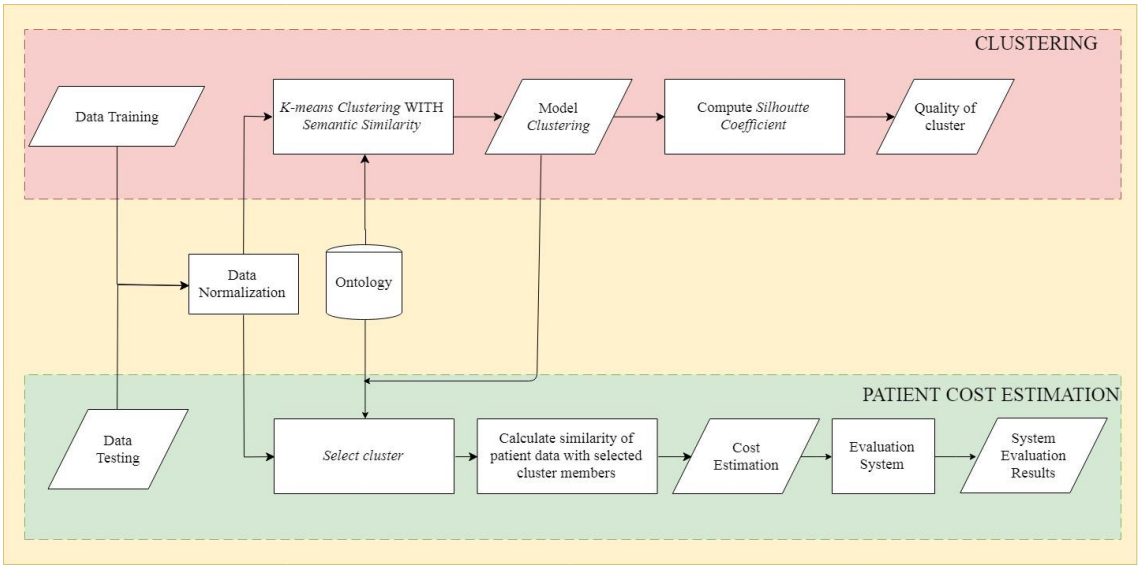

Figure 1 System architecture

Figure 1 is the architecture of the patient clustering system using the K-means algorithm. Broadly speaking, there are two processes in the architecture of clustering systems, namely the process of clustering and the process of estimating patient costs. The clustering process is intended for grouping patient data into clusters, where each cluster has high similarity data.

The first step in the clustering process is that patient data will be carried out in the stages of data normalization first, so that it matches the code format on the ontology for the patient diagnosis feature. For example, there is a patient who has a diagnosis code A01.0 (Typoid Fever), then it will be normalized to A01_0. In addition to the diagnosis data of age data and also patient class rates were also carried out stages of normalization using the range method. The measurement of similarity between data on the K-means algorithm used semantic similarity and euclidean distance. Semantic similarity was used to measure the similarity of

diagnosis features, while euclidean distance was used to measure the similarity of features of gender, age, and insurance class rates.

The results of the clustering process will form patient clusters that have high features similarity, and have a low similarity between data on different clusters. To find out the quality of the clusters produced, testing will be carried out using the silhouette coefficient method. Silhouette coefficient is able to measure the quality and strength of clusters, so it can be seen how well the data are placed in a cluster.

The testing process is used to determine the patient's cost estimate and also to evaluate the results of the patient's cost estimate. The process of determining patient cost estimate using the normalization method and the similarity calculation that is the same as the clustering process. The normalized patient data will only be measured to each centroid of each cluster for cluster selection. The selected cluster is a cluster that has a centroid with a high similarity value. Hence, to determine the estimated cost of a new patient, only a single cluster is matched with the new patient data. The results of cost estimate displayed by the system are in the form of a range of minimum and maximum costs obtained from patient data that have similarity values above or equal to the set threshold value.

2.4 Semantic Similarity

In measuring the similarity between data in clustering algorithms, this study used a measure of semantic similarity. Semantic similarity was chosen to measure the similarity between data due to limitations of distance calculation algorithms such as euclidiean distance, which cannot measure semantic proximity between data. Semantic similarity is obtained by calculating the distance between concepts on ontology.

2.4.1 Semantic similarity between concepts

The measurement of semantic similarity of two concepts was measured using the equation proposed by Girardi et al. [5], Leacock & Chodorow [6], and Rada et al. [7].

1. Semantic similarity of Girardi et al.
   The two nodes (concept) *x* and *y* that have been represented in the form of hierarchical trees can be calculated similarity to the equation:

$$d(x,y) = \frac{p_{min}(x,y)}{l(x) + l(y)} \quad (2)$$

   Where:
   $d(x,y)$ = the value of the distance between nodes *x* and *y*.
   $p_{min}(x,y)$ = the minimum number of edges between nodes *x* and *y*.
   $l(x)$ = the depth level of the node *x*.
   $l(y)$ = the depth level of the node *y*.

2. Semantic similarity of Leacock & Chodorow
   Leacock & Chodorow used the path length between the two nodes to measure semantic similarity. The equation of the semantic similarity method of Leacock Chodorow can be seen as follows:

$$sim_{L\&C}(u,r) = -\log \frac{length(u,r)}{2 * D} \quad (3)$$

   Where:
   length (u, r) = shortest distance from *node u* with *node r*.
   D = maximum depth from the node to the root between node *u* with node *r*.

3. Semantic similarity of Rada et al.
   The semantic similarity proposed by Rada et al. use the shortest path distance and depth level to measure the similarity between the concepts on ontology. The distance between the two concepts C1, C2 is calculated as the shortest path that connects the concept. The similarity between the two concepts C1 and C2 can be calculated as follows [7]:

$$\mathrm{Sim}(\mathrm{C1},\mathrm{C2}) = 2 \times Max\text{-}length(C1, C2) \tag{4}$$

Where:

Max = maximum depth from node to root between node C1 and C2.
length(C1,C2) = shortest distance from node C1 with node C2.

2.4.2 Semantic similarity between sets of concepts

To calculate the similarity of the collection of concepts using equation (5) [5].

$$d_H(X,Y) = \frac{1}{|X \cup Y|}\left(\sum_{x \in X \setminus Y} \frac{1}{|Y|} \sum_{y \in Y} d(x,y) + \sum_{y \in Y \setminus X} \frac{1}{|X|} \sum_{x \in X} d(x,y)\right) \tag{5}$$

Where:

$d_H(X,Y)$ is a similarity value of X with Y
X or Y is a collection of concepts

2.5 Jaccard Similarity

Jaccard similarity is used to calculate the similarity between two objects of patient diagnosis. The value of Jaccard similarity is obtained from intersection divided by union from two sets of compilations. Jaccard distance is a measurement that is not similar between data sets. This can be determined by the inverse of the Jaccard coefficient obtained by removing the Jaccard similarity from the value of Jaccard similarity [9]. The equation for calculating Jaccard similarity is as follows:

$$Jaccard\ (A,B) = \frac{A \cap B}{A \cup B} \tag{6}$$

2.6 K-means Similarity

K-means clustering algorithm aims to classify patients. In K-means, each data must be included in a particular cluster, but it is possible for each data to be included in a particular cluster at a stage of the process, in the next step, move to another cluster [10].

Figure 2 shows the flow diagram of the K-means algorithm with semantic similarity.

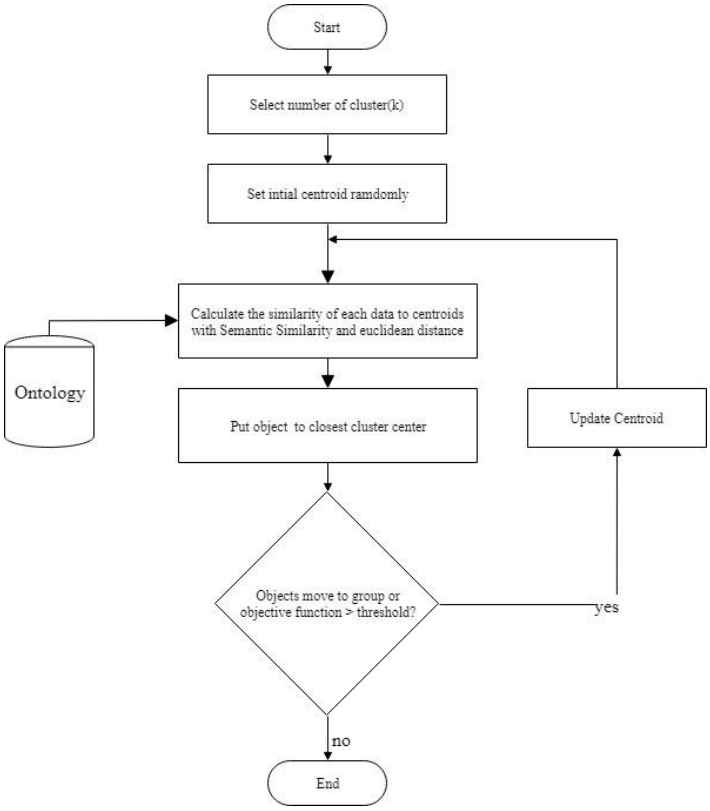

Figure 2 Flowchart of K-means using semantic similarity

The first step in the K-means algorithm is to determine the number of clusters of patients to be formed. The determination of the number of clusters affects the determination of the number of centroids. If the number of clusters to be formed is three, it will choose as many as three data centroids. The initial centroid value in the first iteration was given randomly. When the initial centroid has been selected, then it is to calculate the similarity for each patient data to each centroid.

Data on patients who have high similarity values with centroids in a particular cluster, the patient data were categorized and allocated to the cluster. The process for calculating the similarity value of patient data with centroid using the semantic approach and euclidean distance. Sematic similarity between concepts is calculated using equation (2), (3), (4) and the semantic similarity between sets of concepts is calculated using equation (5). Jaccard similarity is calculated using equation (6). After all patient data were allocated to a cluster, the next step is to check the convergence of the patient cluster results by comparing the cluster results in the previous iteration with the cluster results in the iteration that are running or using the specified objective function value. If the results are the same or if the change in the objective function value is below the specified threshold value, then the clustering data results will be converged, but if different or if the objective function value changes are above the specified threshold value, then it has not been converged. It is necessary to do the next iteration and re-determine the new centroid based on the data from each cluster. The new centroid determination is conducted by looking for the average similarity value of all members in each cluster.

The step will repeat again until there is no change in membership of each cluster or changes in the value of the objective function used below the threshold value, so that the data can be converged, the threshold value used is 0.1. The objective function is used to check data convergence in a cluster, namely Sum Square Error (SSE). SSE is the sum of all distances of each data with the cluster center point [1]. So that, the final result of this method is grouping patients who have high similarity between data in a cluster.

2.7 The Process Of Estimating Patient Costs

Estimated patient costs generated from the system are the minimum and maximum cost ranges obtained from patients who have a similarity value above or equal to the threshold value.

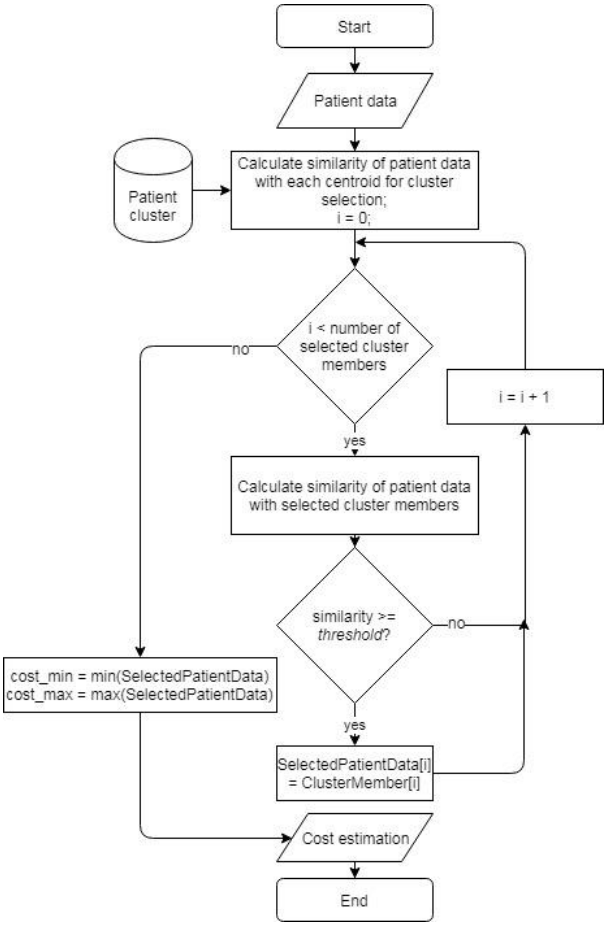

Figure 3 Flowchart determining patient cost estimates

Figure 3 is a sequence of processes to determine patient cost estimate. Each patient's data will be estimated at cost, the first step is to measure the similarity of patient data that will be predicted with each centroid of the cluster. The calculation of the similarity of data with centroid aims to narrow the search space in estimating patient costs, so that the process of calculating similarities to obtain estimates of patient costs will only be carried out in one particular cluster, namely clusters with centroids which have high similarity values with predictable patient data. Estimated patient costs are obtained from cluster members who have similarity values above or equal to the threshold value. Estimated patient costs displayed by the system in the form of a range of costs, namely minimum and maximum costs.

3.RESULTS AND DISCUSSION

3.1 Testing Scheme

Tests are conducted to compare methods of measuring data using semantic and Jaccard similarity on clustering and cost estimates. The semantic data measurement method used the semantic similarity of Girardi et al., Leacock & Chodorow, and Rada et al., while the non-semantic data measurement method used Jaccard. In this study, the amount of data used was

244 patient data. Patient data were divided into training data and test data. 171 patient data were used as training data and 73 patient data as test data. Tests carried out in this study include:

1. Looking for the optimal number of clusters.
2. Measuring the accuracy of the proposed method.
3. Measuring the computational time of the clustering process

3.2 Determining the Number of Clusters

Table 1 is the result of testing the determination of the number of clusters using the silhouette coefficient method. Based on the results of the tests conducted, the highest average value of silhouette coefficient by measuring the similarity of data using the semantic similarity of Girardi et al., which is 0.72 with the number of clusters k=10. The average value of silhouette coefficient by measuring the similarity of data using the semantic similarity of Leacock & Chodorow is 0.73 with the number k=10. The average value of the silhouette coefficient used the measurements of the semantic similarity of Rada et al. is 0.77 with the number k=10. While, the average value of the silhouette coefficient using the Jaccard Similarity measurement is 0.69 with the number k=10. From the three semantic similarity measurement methods used, the best number of clusters is 10 clusters.

Table 1 Testing results determine the number of clusters with the silhouette coefficient

Number *cluster*	**Average *Silhoutte Coefficient***			
	Girardi dkk.	***Leacock & Chodorow***	***Rada dkk.***	***Jaccard Similarity***
2	0.27	0.26	0.31	0.25
3	0.22	0.23	0.38	0.2
4	0.47	0.45	0.56	0.44
5	0.51	0.55	0.63	0.49
6	0.62	0.61	0.67	0.54
7	0.57	0.6	0.65	0.56
8	0.68	0.68	0.72	0.6
9	0.69	0.7	0.74	0.65
10	**0.72**	**0.73**	**0.77**	**0.69**
11	0.65	0.66	0.7	0.6
12	0.44	0.46	0.48	0.41
13	0.43	0.39	0.38	0.39
14	0.21	0.21	0.15	0.19
15	0.17	0.22	0.17	0.15

3.3 Comparison of accuracy of the similarity method

Comparison of similarity measurement methods using semantic similarity of Girardi et al., semantic similarity of Leacock & Chodorow, semantic similarity of Rada et al., and Jaccard Similarity are presented in Table 2. Measurement of system accuracy was conducted by comparing the estimated range of costs incurred by the system with the actual costs incurred by patient. Estimates of the costs displayed by the system are in the form of a range of minimum and maximum costs obtained from patient data that have similarity values above or equal to the set threshold value. If the actual costs incurred by the patient fall into the estimated range of costs estimated by the system, then it is true. The best accuracy is 91.78% for the three semantic similarity methods, whereas without semantic similarity the best accuracy is 84.93%.

Table 2 Comparison of accuracy in each measurement method

Similarity method	***Threshold***			
	>=60%	**>=70%**	**>=80%**	**>=90%**
Jaccard Similarity	84.93%	75.34%	63.01%	61.64%
Girardi dkk.	91.78%	83.56%	68.49%	67.12%
Leacock & Chodorow	91.78%	80.82%	68.49%	67.12%
Rada dkk.	91.78%	84.93%	69.86%	67.12%

3.4 Evaluation of computing time

In Table 3, it can be seen the comparison of execution times of each data measurement method in clustering represented in seconds.

Table 3 Comparison of execution time of clustering process

Number *cluster*	**Execution time (second)**			
	Girardi	***Leacock Chodorow***	***Rada***	**Jaccard Similarity**
2	69.61	71.65	75.31	7.73
3	56.32	73.57	78.42	4.14
4	65.72	70.57	71.18	10.12
5	78.16	49.26	47.93	9.27
6	72.72	53.79	47.48	18.06
7	48.67	42.63	42.02	17.94
8	28.84	31.36	29.59	9.91
9	30.63	29.04	34.29	11.10
10	39.55	38.81	33.20	16.00
11	38.36	31.86	36.27	11.94
12	32.86	30.81	31.26	12.84
13	34.65	32.46	34.19	7.90
14	40.23	35.10	35.83	13.24
15	110.98	82.87	100.90	15.48
Average	**53.38**	**48.13**	**49.85**	**11.83**

Based on the results of the comparison of execution time presented in Table 3, the method of measuring Jaccard similarity data has an average execution time less than the method of measuring semantic similarity data. The non-semantic method has a less execution time because the method has a simpler formula, which is only looking for equations without taking into account semantic proximity.

4.CONCLUSIONS

Based on testing with silhouette coefficient, the method of measuring semantic similarity data on K-Means algorithm is able to produce better quality clustering results compared to Jaccard similarity. The quality of clustering results generated from the method of measuring semantic similarity data belongs to the strong structure. From the three semantic similarity methods used, the semantic similarity method of Rada et al. produce clustering

quality that is better than semantic similarity of Girardi et al. and semantic similarity of Leacock & Chodorow. The best accuracy is 91.78% for the three semantic similarity methods, whereas without semantic similarity the best accuracy is 84.93%.

5.FUTURE WORKS

From the results, there are several things that need to be added and developed for further research, namely the need to use methods other than K-Means to conduct clustering, so that the best clustering method can be obtained with optimal results and estimates of patient costs can be closer to actual costs.

REFERENCES

[1] J. Han, and M. Kamber, "Data Mining: Concepts, Models and Techniques," *Intelligent Systems Reference Library*, 2006.

[2] D. J. Bora and D. A. K. Gupta, "Effect of Different Distance Measures on the Performance of K-Means Algorithm: An Experimental Study in Matlab," *International Journal of Computer Science and Information Technologies*, vol. 5, p. 6, 2014 [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1405/1405.7471.pdf. [Accessed: 4-Feb-2019]

[3] S. S. Desai and J. A. Laxminarayana, "WordNet and Semantic Similarity Based Approach for Document Clustering," in *2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, Bengaluru, India, 2016, pp. 312–317 [Online]. Available: https://ieeexplore.ieee.org/document/7779377. [Accessed: 28-Jan-2019]

[4] Ahmed, M. Malki, and S. M. Benslimane, "Ontology Partitioning: Clustering Based Approach," *International Journal of Information Technology and Computer Science*, vol. 7, no. 6, pp. 1–11, May 2015 [Online]. Available: http://www.mecs-press.org/ijitcs/ijitcs-v7-n6/IJITCS-V7-N6-1.pdf. [Accessed: 2-Feb-2019]

[5] D. Girardi, S. Wartner, G. Halmerbauer, M. Ehrenmüller, H. Kosorus, and S. Dreiseitl, "Using Concept Hierarchies to Improve Calculation of Patient Similarity," *Journal of Biomedical Informatics*, vol. 63, pp. 66–73, Oct. 2016 [Online]. Availbale: https://www.sciencedirect.com/science/article/pii/S1532046416300752. [Accessed: 28-Jan-2019]

[6] I. Fahrurozi, "Sistem Rekomendasi Berbasis Kombinasi Semantic Similarity dan Collaborative Filtering (Studi Kasus pada Toko *Accessories* Handphone Besseling Cell)," Thesis, Universitas Gadjah Mada, Yogyakarta, 2017.

[7] A. F. S. Althobaiti, "Comparison of Ontology-Based Semantic-Similarity Measures in the Biomedical Text," *Journal of Computer and Communications*, vol. 05, no. 02, pp. 17–27, 2017 [Online]. Available: https://file.scirp.org/pdf/JCC_2017020917284790.pdf. [Accessed: 28-Jan-2019]

[8] G. R. Hatta, "Pedoman Manajemen Informasi Kesehatan Disarana Pelayanan Kesehatan (Revisi 3)," Jakarta: Universitas Indonesia, 2017.

[9] S. Niwattanakul, J. Singthongchai, E. Naenudorn, and S. Wanapu, "Using of Jaccard Coefficient for Keywords Similarity," *Proceedings of the International MultiConference of Enginers and Computer Scientists*, p. 5, 2013 [Online]. Available: http://www.iaeng.org/publication/IMECS2013/IMECS2013_pp380-384.pdf. [Accessed: 6-Feb-2019]

[10] I. Riadi, "Framework Untuk Forensik Internet Menggunakan K-Means Clustering dan Horizontal Partitioning," Desertasi, Universitas Gadjah Mada, Yogyakarta, 2014.