Case-Based Reasoning for Stroke Diseases Diagnosis

Nelson Rumui*[1], Agus Harjoko[2], Aina Musdholifah[3]
[1]Master Program of Computer Science; FMIPA UGM, Yogyakarta, Indonesia
[2,3]Departmen of Computer Science and Electronics, FMIPA,UGM, Yogyakarta, Indonesia
e-mail: *[1]**nelsonrumui@gmail.com**, [2]aharjoko@gmail.com, [3]aina_m@ugm.ac.id

Abstrak

    Stroke merupakan jenis penyakit cerebrovascular yang terjadi karena aliran darah ke otak terganggu. Pemeriksaan stroke menggunakan CT-scan, namun alat tersebut tidak selalu tersedia oleh karena itu dapat dilakukan dengan menggunakan Skor Siriraj. Setiap jenis stroke mempunyai gejala yang mirip sehingga dokter harus memeriksa kasus terdahulu yang serupa untuk melakukan diagnosis. Case-based reasoning(CBR) didasarkan pada konsep tersebut, dimana masalah yang mirip mempunyai solusi yang mirip.

    Penelitian ini mengimplementasikan konsep CBR menggunakan skor siriraj dan dense index sebagai indexing dan metode jaccard coeficient untuk perhitungan similaritas antar kasus.

    Pengujian dengan k-fold cross validation 4 fold dengan threshold similaritas (0,65), (0,7), (0,75), (0,8), (0,85), (0,9), dan 0,95. Data uji 45 kasus setiap fold dan basis kasus 135 kasus. Hasil pengujian menunjukkan threshold similaritas 0,7 cocok diterapkan karena menghasilkan sensitivitas (86,95%) dan akurasi (81,67% untuk CBR menggunakan indexing dan 84,44% untuk CBR tanpa indexing) dengan sedikit kasus hasil retrieve yang tidak relevan. Threshold similaritas 0,65 menghasilkan sensitivitas yang tinggi namun banyak kasus retrieve yang tidak relevan. Threshold similaritas (0,75), (0,8), (0,85), (0,9) dan 0,95 menghasilkan sensitivitas (66,07%, 54,76%, 7,14%, 2,97%, dan 0%) dan akurasi CBR menggunakan indexing (62,22%, 51,11%, 6,66%, 2,78%, dan 0%) dan akurasi CBR tanpa indexing (62,78%, 52,22%, 6,66%, 2,78%, dan 0%).

Kata kunci—case-based reasoning, jaccard coefficient, siriraj, stroke, dense index

Abstract

    Stroke is a type of cerebrovascular disease that occurs because blood flow to the brain is disrupted. Examination of stroke using CT-scan, but the tool is not always available, so it can be done by the Siriraj Score. Each type of stroke has similar symptoms so doctors should re-examine similar cases prior to diagnosis. The hypothesis of the Case-based reasoning (CBR) method is similar problems having similar solution.

    This research implements CBR concept using siriraj score, dense index and Jaccard Coeficient method to perform similarity calculation between cases.

    Testing performed by k-fold cross validation(4 fold) and set threshold 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, and 0.95. Using 45 cases of data test and 135 cases of database. The test showed that threshold of 0.7 is suitable to be applied in sensitivity(86,95%) and accuracy(81,67% for CBR using indexing and 84,44% for CBR without indexing). Threshold of 0.65 resulted high sensitivity and accuracy but showed many cases of irrelevant retrieval results. Threshold 0.75, 0.8, 0.85, 0.9 and 0.95 resulted in sensitivity (66.07%, 54.76%, 7.14%, 2.97%, and 0%) and accuracy of CBR using indexing (62.22%, 51.11%, 6.66%, 2.78%, and 0%) and accuracy of CBR without indexing (62.78%, 52.22%, 6.66%, 2.78%, and 0%).

Keywords— case-based reasoning, jaccard coefficient, siriraj, stroke, dense index.

1. INTRODUCTION

1.1 Background

Stroke is a type of cerebrovascular disease (brain blood vessel) that characterized by brain tissue death (cerebral infarction) that occurs due to reduced blood flow and oxygen to the brain. Based on data from Yayasan Stroke Indonesia, number of stroke patients in Indonesia is the most and ranks first in Asia [1]. Data from Riset Kesehatan Dasar (Riskesdas) of Indonesia in 2015, showed that the main cause of death in Indonesia is stroke disease with a percentage of 21.1% [2].

The most optimal subscription of stroke disease is only three hours when a person has a stroke attack, when it not handled quickly can result in permanent disability and even death. To diagnose stroke patients accurately is by examining computerized tomography (CT) scan or magnetic resonance imaging (MRI), but these tools are not always available, as well as with the availability of the specialist doctors, while the diagnosis of stroke-affected patients has to be done imediatlely. Due to the limitations of CT scans and MRIs, stroke examination can be performed by using a simple scoring system. One of which has been validated in many countries is Siriraj score [3]. The Siriraj score distinguishes between ischemic stroke and hemorrhagic stroke, but not specifically, so that a diagnostic tool is needed that can help diagnose stroke patients. The symptoms that appear in one type of stroke are often the same for the same type of stroke, so the doctor should re-examine similar cases prior to diagnosis while stroke patients should be treated promptly. Case-based reasoning (CBR) is based on that concept where the hypothesis of this CBR method is a similar problem having similar solutions [4].

Stroke cases stored in the form of a record can be used as a reference in determining the type of stroke. The CBR method mimic the reasoning of an expert, by remembering the previous cases. Siriraj score can be used as indexing, where the results of the siriraj score are used as keyword in retrieving old cases that are relevant to new problems. After the retrieval results are obtained, a similarity measurement of the new problem with the retrieval cases is compared by each feature of the new case and the old case. In the medical field, when the medical personnel perform the examination of the patient, it usually compares with the old patient who has been examined before and consider the similarity of symptoms experienced, and ignore the similarity in terms of having no particular symptoms (negative matches). Jaccard coefficient is one method to calculate the similarity of binary data by ignoring negative matches [5].

Based on the description, this research intends to develop CBR system for the purpose of diagnosis of stroke by using Siriraj score for indexing using dense index and Jaccard Coeficient method to perform similarity calculation between cases.

1.2 Literature Review

Research in medical especially for stroke diagnosis have been done before. Application of CBR to diagnose nerve dysfunction of stroke patients. The data of the case is collected with a KINARM robot tool. The test was performed with 15 test data and 30 cases of database. Testing yielded sensitivity of 50.97%, specificity 98.06%, and accuracy 82.42% [6].

Research on the CBR domain has been mostly done in the medical field. The application of the CBR method for diagnosing ear and throat (ENT) diseases uses backpropagation method to obtain an index of new problems and cosine coefficient method to calculate the similarity between cases. The test was performed on 111 cases with an accuracy of 91.89% [7]. Furthermore, the application of CBR concept to diagnose heart disease using nearest neighbor similarity method, minkowski distance similarity, and euclidean distance similarity. The test used 58 test data and 134 cases of database with threshold similarity of 0.8. The result shows that the accuracy using nearest neighbor similarity method is 86,21%, using minkowski distance similarity method 100%, and euclidean distance similarity equal to 94,83% [8].

1.3 Stroke Disease

Stroke is a type of cerebrovascular disease characterized by brain tissue death due to reduced blood flow to the brain [1]. Stroke can be divided into two groups: hemorrhagic stroke and Ischemic stroke. The incidence of diseases between ischemic and hemorrhagic strokes is 80% versus 20% [9]. In this study used a stroke classification based on anatomic pathology and its causes consisting of Thrombotic stroke, Embolic stroke, Subarachnoid hemorrhage stroke, and Intraceberal hemorrhage stroke.

1.4 Case-Based Reasoning

Case-based reasoning (CBR) is a problem-solving method, that when given a new problem, it will look for an appropriate solution to the problem [10]. The troubleshooting steps in CBR include: retrieve to find old cases in database similar to new cases. Reuse to reuse the solution from the previous case as a solution for the new case. Revise to improvement the proposed solution for a new case. Retain to save new case into database.

2. METHODS

2.1 System Description

This research implements CBR concept using siriraj score and dense index to perform indexing process and implement jaccard coefficient method to perform similarity measurement. Similarity calculation results are then evaluated to determine the solution of the new problem.

2.2 Siriraj Score

Standard examination of stroke disease is by using a CT scan or MRI, when there is no tools then the examination can be done by a simple scoring system, one of which has been validated in many countries is Siriraj score [3]. The Siriraj score distinguishes between ischemic stroke and hemorrhagic stroke. Siriraj score was based on a study of 174 stroke patients admitted to Siriraj hospital during 1984-1985 with the aim of developing a simple, reliable, safe, and can be use as a stroke diagnostic tool in a region without a CT scan. Assessment of the siriraj score can be seen in Table 1 and the calculation is given in equation (1) [11].

Table 1 Siriraj score

Variable	Point
Consciousness	Compos Mentis = 0, Somnolen = 1, Sopor = 2
Vomiting	Yes = 1
Headache	No = 0

Atheroma
  1. Angina Pectoris
  2. Claudicatio Intermitten
  3. Diabetic history
	
	Yes = 1
	No = 0

$$\textit{Siriraj score} = \textbf{(2.5 * Consciousness)} + \textbf{(2 * Vomiting)} + \textbf{(2 * Headache)} +$$
$$\textbf{(0.1 * Diastolic blood pressure)} - \textbf{(3* Atheroma)} - \textbf{12} \qquad (1)$$

If Siriraj score is ≥ 1 : Diagnosis of hemorrhagic
If Siriraj score is ≤ -1 : Diagnosis of ischemic
If Siriraj score is -1 to 1 : Uncertain diagnosis

2.3 Case Representation

Good case representation makes retrieves more efficient. The data cases used as the database in this study is the medical records data of patients with stroke in Dr. DKT Hospital.

Soetarto, Yogyakarta. Case representation consists of problem space and solution space. Problem space includes patient condition, symptoms, risk factors, and disease type as solution space. The case representation in this research is in the frame as show in Table 2.

Table 2 Fitures rating

Fitur	Nilai
1. Age	In units of years
2. Gender	Male = 1, Female = 0
3. Consciousness	Compos Mentis = 0, Somnolen = 1, Sopor = 2
etc..	…

Symptoms
    G1: Speech disorder
    G2: Dizzy
    etc…
	Yes = 1, No = 0

Risk factors
    FR1: Heart disease
    FR2: Hypertension
    Dst…
	Yes = 1, No = 0

Solution Space	Type of stroke disease

2.4 Dense Index

Indexing is the process of grouping cases on the features specified. The indexing process makes time and memory more efficient because the system does not need to calculate the value of similarity to all existing cases, simply calculating the value of similarity to the cases in the same group. This study uses the dense index which is one of ordered indices that can overcome the search records in un-ordered files. The Dense index has an index for each search key in the file that directly points to the indexed record so that all the values of the search key appear in the index file.

2.5 Local Similarity

Local similarity is a measure of similarity at the feature level. Calculating the local similarity should note the data type of each fiture. Types of the feature of the data used include:

1) Numerical Data

Local similarity to numerical data using equation (2) [12].

$$f_i(S_i, T_i) = 1 - \frac{|S_i - T_i|}{R} \tag{2}$$

$f_i(S_i, T_i)$: Local similarities of i-th feature between the source case and the target case.
S_i : i-th fiture of the source case.
T_i : i-th fiture of the target case.
R : Range of the fiture.

2) Boolean Data

Local similarity to boolean data using equation (3) [4].

$$f_i(S_i, T_i) = \begin{cases} 1, & \text{if } S_i = T_i \\ & \qquad\qquad \text{for } S_i, T_i \{\text{true}, \text{false}\} \\ 0, & \text{else} \end{cases} \tag{3}$$

2.6 Global Similarity

A global similarity is a measure of similarity at the case level to calculate the similarity between cases. This research uses jaccard coefficient for global equality calculation. jaccard coefficient is a statistical similarity measurement method used to compare the similarity between a set of data with other data sets [13]. The calculation of global similarities involves the weight of each feature according to the type of disease, but also the level of confidence in the

new problem, where the similarity calculation is based on the similarity of attributes that exist in the new problem and the old case using equation (4).

$$Sim(S,T) = \frac{\sum_{i=1}^{n} f_i(S_i,T_i)*W_i}{\sum_{i=1}^{n} W_{i(N)}} * \frac{J(S_i,T_i)}{J(T_i)} \quad (4)$$

Where:

Sim (S,T)	:	Global similarity between the source case (*S*) and the target case (*T*).
W_i	:	Weight of the *i*-th fiture.
f_i(S_i,T_i)	:	Function of *i*-th local similarity from the source case and the target case.
S_i	:	*i*-th of the Attribute of the source case.
T_i	:	*i*-th of the Attribute of the target case.
$W_{i(N)}$	:	Weight value of the features that appear in the source and in the target case.
$J(S_i,T_i)$	:	The number of features in the target case are identical to the source case.
$J(T_i)$	:	The number of features that appear on the target case.

2.7 System Architecture Design

The system is designed to perform a diagnosis of stroke disease based on data of the patients. System users are divided into three groups: admin, paramedic, and expert. The built-in system architecture is shown in Figure 1. The paramedics input new problem into the system. The new problem such as patient's condition, symptoms, and risk factors that appear or felt by patient. System will calculates the siriraj score based on new problem, then the result of the interpretation of siriraj score is used as the keyword to retrieve the cases with dense index, then system performs the local similarity calculation based on the type of feature data and global similarity using jaccard coefficient between the retrieve cases and the problem new. Furthermore, system choose case that have the highest similarity value to be evaluated, if it meet the target similarity then the solution of the case that has the highest similarity value is used as the solution of the new problem. If it do not meet the target similarity then the new problem will be saved without a solution to be revised by an expert.
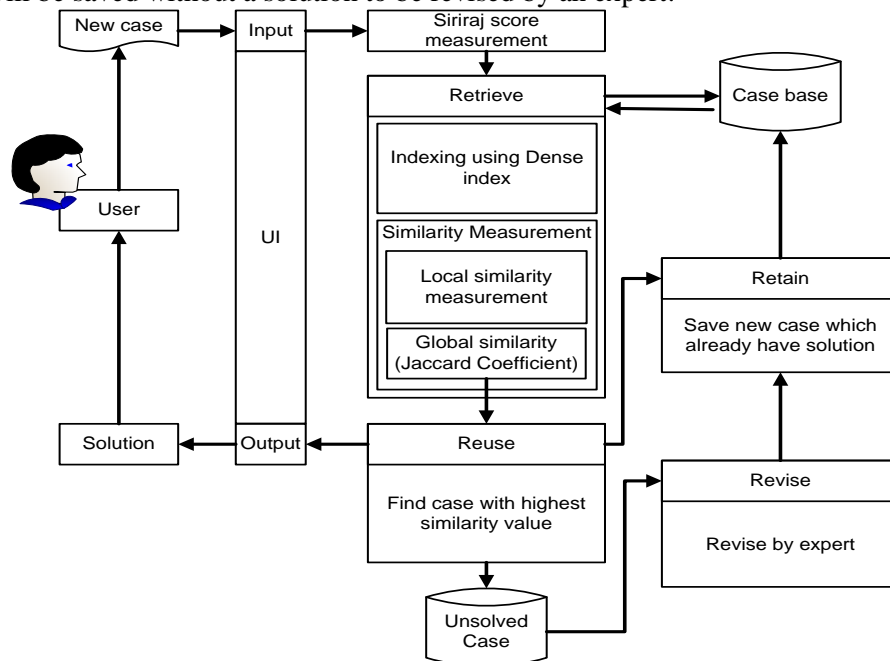

Figure 1 CBR system architecture design for stroke diagnosis

2.8   Data Cases and Testing Method

The cases that used in this study is the medical record data of stroke patients obtained from DKT Dr. Soetarto, Hospital, Yogyakarta. Total cases are 180 cases consisting of 112 cases of ischemic stroke (56 cases of Thrombotic and 56 embolic cases), 56 cases of hemmorhage stroke (28 subarachnoid cases and 28 intracellular cases), and 12 non-stroke cases. The test perform using k-fold cross validation method with 180 cases where divided into 4-folds. Details of the data can be seen in Table 3

Table 3 Data of the stroke patient divide into 4-folds

	Ischemic		Hemorrhage		Non	Total
	Thrombotic	Embolic	Intracerebral	Subarachnoid	Stroke	
Subset 1	14	14	7	7	3	45
Subset 2	14	14	7	7	3	45
Subset 3	14	14	7	7	3	45
Subset 4	14	14	7	7	3	45
Total	56	56	28	28	12	180

The tests were performed by comparing the diagnostic results between CBR system which using indexing and CBR system without indexing by assigning different similarity threshold values, that is 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, and 0.95. Then measures the accuracy, sensitivity, and specificity of each subset. Tests in the medical domain talk about sensitivity and specificity. Sensitivity is a measure that measures how well a test classifies a person who is ill correctly, while the specificity is the proportion of people who are not sick and not sick as well when identified [14]. The sensitivity of the system to determine the extent to which the system is able to diagnose stroke disease correctly, and the specificity to know the extent to which the system is able to diagnose cases not stroke correctly. To calculate the accuracy, sensitivity, and specificity of the system, a predictive classification table called the confusion matrix is used shown in Table 4.

Table 4 Confusion matrix

		Predicted condition	
		Predicted condition positve	Predicted condition negative
True conddition	Condition positive	T_P (True Positive)	F_N (False Negative)
	Condition negative	F_P (False Positive)	T_N (True Negative)

Based on the confusion matrix then accuracy can be calculated with equation (5), sensitivity can be calculated with equation (6), and specificity can be calculated with equation (7).

$$Accuracy = \frac{T_P + T_N}{T_P + F_P + T_N + F_N} * 100\% \tag{5}$$

$$Sensitivity = \frac{T_P}{T_P + F_N} * 100\% \tag{6}$$

$$Specificity = \frac{T_N}{F_P + T_N} * 100\% \tag{7}$$

3. RESULTS AND DISCUSSION

3.1 Test Results with K-Fold Cross Validation

Testing result using k-fold cross validation method with k = 4 for CBR system using indexing and CBR system without indexing can be seen in Table 5.

Table 5 Testing results

	Threshold	CBR system using indexing	CBR system without indexing
Accuracy	0.65	85.55%	90.56%
	0.7	81.67%	84.44%
	0.75	62.22%	62.78%
	0.8	51.11%	52.22%
	0.85	6.66%	6.66%
	0.9	2.78%	2.78%
	0.95	0%	0%
Sensitivity	0.65	91.08%	91.08%
	0.7	86.95%	86.95%
	0.75	66.07%	66.07%
	0.8	54.76%	54.76%
	0.85	7.14%	7.14%
	0.9	2,97%	2,97%
	0.95	0%	0%
Spesificity	0.65	8.32%	83.34%
	0.7	8.32%	50%
	0.75	8.32%	16.66%
	0.8	0%	16.66%
	0.85	0%	0%
	0.9	0%	0%
	0.95	0%	0%
Number of irrelevant retrieval cases	0.65	16	14
	0.7	7	4
	0.75	5	3
	0.8	5	2
	0.85	4	2
	0.9	4	2
	0.95	4	2

3.2 Comparison of Accuracy, Sensitivity, and Specificity Results

The results of the system accuracy calculation in table 5 shows that the accuracy of CBR system without indexing is higher than CBR system using indexing for threshold similarity of 0.65, 0.7, 0.75, and 0.8. The difference in the results of the system accuracy are shown in the graph in Figure 2.
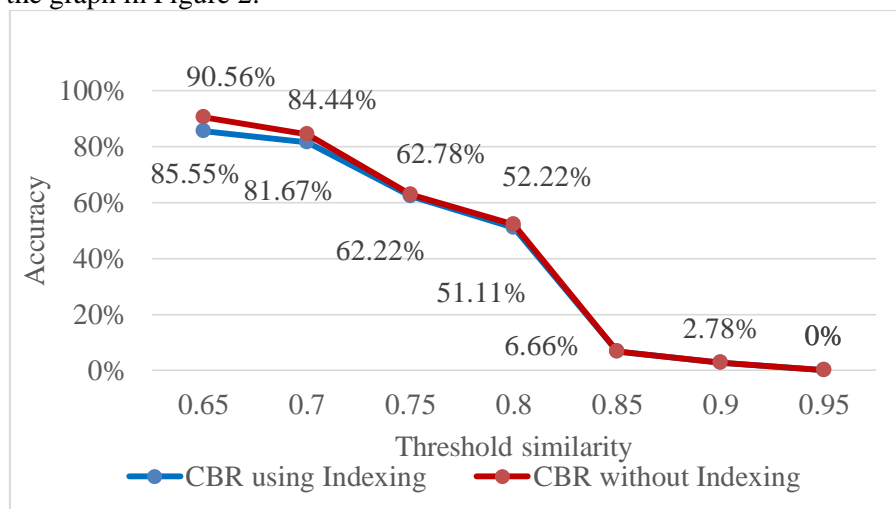

Figure 2 Graph of the comparison of system accuracy

The sensitivity of the system in Table 5 shows the same results between CBR system using indexing and CBR system without indexing. The results of the system sensitivity are shown in the graph in Figure 3.
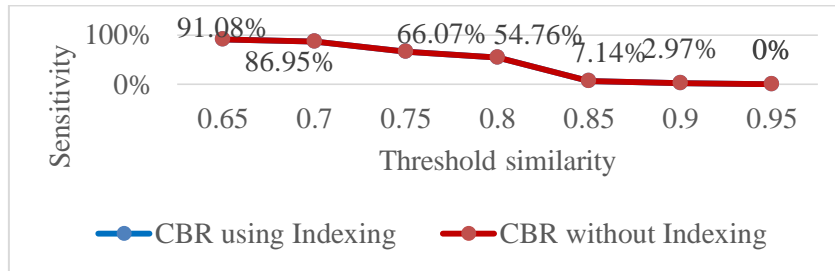

Figure 3 Graph of the comparison of system sensitivity

The specificity of the system in Table 5 shows that the specificity of CBR system without indexing is higher than CBR systems using indexing for the application of threshold similarity 0.65, 0.7, 0.75, and 0.8. The application of threshold similarity 0.85, 0.9, and 0.95 have the same specificity between CBR system using indexing and CBR system without indexing. The difference in the results of the system specificity are shown in the graph in Figure 4.
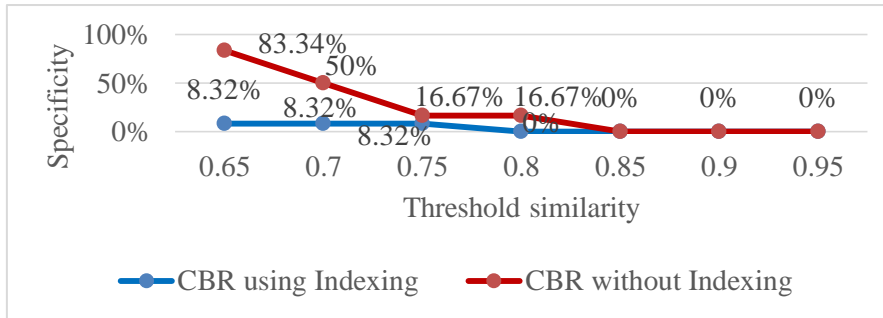

Figure 4 Graph of the comparison of system specificity

3.3 Comparison of the Irrelevant Cases Retrieval

The average number of the irrelevant retrieval cases in Table 5 shows that the CBR system without indexing resulting in fewer irrelevant cases than the CBR system using indexing for the application of threshold similarity 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, and 0.95. The difference in retrieval results of the irrelevant system are shown in the graph in Figure 5.
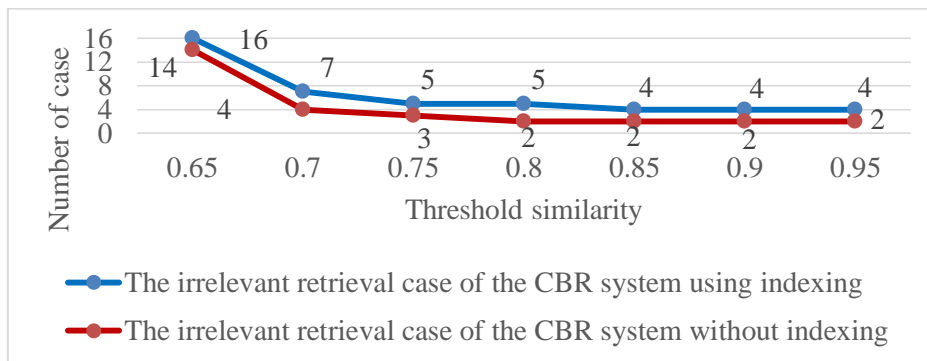

Figure 5 Graph of the comparison of the Irrelevant cases retrieval

3.4 Retrieval Time Calculation

The retrieve time of the cases was based on 135 cases consisting of 84 ischemic cases, 42 cases of hemorrhagic, and 9 non stroke cases. The average calculation result of case retrieve time can be seen in Table 6.

Table 6 Test results of retrieval time

	Average time (seconds)	
	CBR system using indexing	CBR system without indexing
Ischemic cases	10.845	16.913
Hemorrhagic cases	5.095	16.845
Non-stroke cases	11.485	17.903
Data test (-1<siriraj score<1)	17.036	16.893
Total average	11.115	17.139

The retrieval time testing performed yielded average retrieve rates for ischemic cases of the CBR system using indexing requires an average of retrieve time is 10.845 seconds while for CBR system without indexing takes longer retrieve time which is 16.913 seconds. The average retrieve time rate for hemorrhagic test cases for CBR system using indexing requires an average retrieve time of 5.095 seconds while for CBR system without indexing requires a longer retrieve time that is 16.845 seconds. The average retrieve time rate for non-stroke cases for CBR system using indexing requires an average retrieve time of 11.485 seconds while for CBR system without indexing requires a longer retrieve time that is 17.903 seconds. The average retrieve time rate for case test data that has a siriraj score between -1 and 1 for CBR system using indexing requires an average retrieve time of 17.036 seconds while for CBR system without indexing requires a much faster retrieve time that is 16.893 seconds. Total average retrieve time for CBR system using indexing require an average retrieve time of 11.115 seconds whereas for CBR system without indexing requires a longer retrieve time that is 17.139 seconds. The results of the average retrieve time testing between CBR system using indexing and CBR system without indexing are shown in the graph in Figure 6.
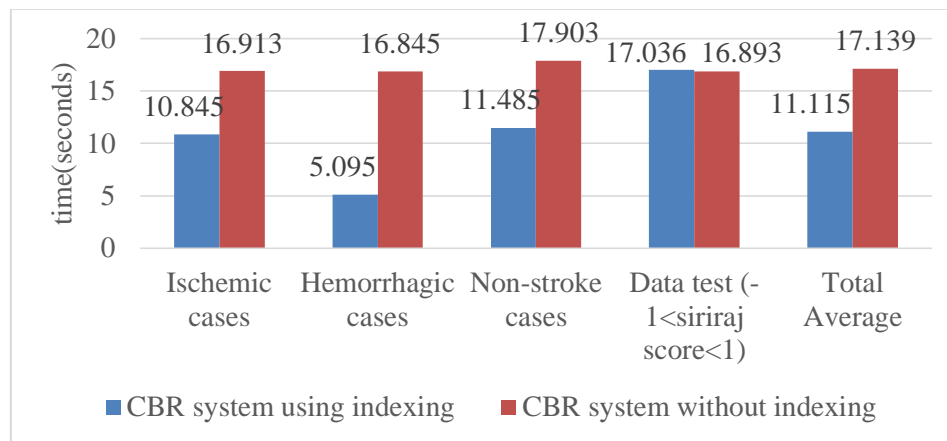

Figure 6 Graph of retrieval time comparison

4. CONCLUSIONS

This study resulted in a CBR system for stroke diagnosis by applying the jaccard coefficient method to calculate similarity between cases, and the siriraj score for distinguishing types of ischemic and hemorrhagic strokes, and also using the dense index as the indexing to perform retrieve cases from database. The test results using k-fold cross validation method indicates that CBR system without indexing gives better specificity and accuracy than CBR system using indexing for application of threshold similaritas value 0.65, 0.7, 0.75 and 0.8.

The test results show that the threshold similarity of 0.7 is suitable to be applied because it produces sensitivity (86.95%) and accuracy (81.67% for CBR system using indexing and 84.44% for CBR system without indexing) with few cases of irrelevant retrieve result. Threshold similarity of 0.65 yields great sensitivity and accuracy but results in many irrelevant retrieval cases. Threshold similarity of 0.75 and 0.8 results in a small number of irrelevant retrieve cases but results in low accuracy and sensitivity. Threshold similarity of 0.85 and 0.9

resulted in accuracy (6.66% and 2.78%) and sensitivity (7.14% and 2.97%), while the applied value of threshold similarity of 0.95 has a degree of accuracy and sensitivity 0%.

Retrieve time testing with database of 135 cases indicates CBR system using indexing takes faster time in retrieve case than CBR system without indexing, where the CBR system using indexing requires 9.099 seconds of total average time while CBR system without indexing need 16,958 seconds of total time in retrieving.

For further research, need to handled on the reuse process if there are 2 or more cases of retrieval cases results that have the same similarity value which highest similarity value. Beside that, the weighting process used in this research is weighting done only by one expert and it is subjective, so for further research it is suggested for the weighting process done by some experts and taken average in order to get better weighting.

REFERENCES

[1] Indonesia Stroke Foundation, Sekilas tentang stroke, 2016 [Online]. Available: http://www.yastroki.or.id/read.php?id=340, [Accesed: 23-Jun-2016].

[2] Ministry of Health of the Republic of Indonesia, 2016 [Online]. Available: http://www.depkes.go.id/article/print/15082800001/indonesia-tuan-rumah-pertemuan-penyakit-tidak-menular-regional-asean.html, [Accesed: 23-Jun-2016].

[3] A. Sherin, A. Khan, S. Rehman, N. H. Shah, G. Shabbier, and M. Zarif, "Comparability and validity of Siriraj stroke score and Allen stroke score in differentiation of acute ischemic and haemorrhagic stroke", *JPMI (Journal of Post Graduate Medical Institute),* vol 3, pp. 206-216, 2011.

[4] Hans and D. Burkhard, "Case Completion and Similarity in Case-Based Reasoning", *ComSIS*, vol. 1, no. 2, pp. 28, 2004.

[5] Clustan, Computing Proximities, 2017 [Online]. Available: http://www.clustan.talktalk.net/computing_proximities.html, [accesed: 06-Jul-2017].

[6] Baig and Mariam, Case-Based Reasoning - An Effective Paradigm For Providing Diagnostic Support For Stroke Patients, *Thesis*, School of Computing, Queen's University, Canada, 2008.

[7] T. Rismawan and S. Hartati, "Case Based Reasoning untuk Diagnosa Penyakit THT (Telinga Hidung dan Tenggorokan)", *IJCCS (Indonesian J. Compt. Cybern. Syst.)*, vol. 6, no. 2, pp. 67–78, Jul. 2012 [Online]. Available: https://jurnal.ugm.ac.id/ijccs/article/view/2154/1934, [Accesed: 10-Jun-2017].

[8] E. Wahyudi and S. Hartati, "Case-Based Reasoning untuk Diagnosis Penyakit Jantung", *IJCCS (Indonesian J. Compt. Cybern. Syst.)*, vol. 11, no. 1, pp. 1–10, Jan. 2017 [Online]. Available: https://jurnal.ugm.ac.id/ijccs/article/view/15523/11717, [Accesed: 10-Jun-2017].

[9] G. Dewanto, Suwono, B. Riyanto, and Y. Turana, *Panduan Praktis Diagnosis dan tata Laksana Penyakit Saraf*, Jakarta: Buku Kedokteran EGC, 2007.

[10] P. K. Sankar and S. C. K. Simon, *Foundations of Soft Case-Based Reasoning*, New Jersey: John Wiley & Sons Inc, 2004.

[11] P. Widiastuti and A. Nuartha, "Sistem Skoring Diagnostik untuk Stroke: Skor Siriraj", *CDK (Cermin Dunia Kedokteran)* Ed. 233, pp. 10, 2015.

[12] K. M. Jha, D. Pakhira, and B. Chakraborty, "Diabetes Detection and Care Applying CBR Techniques", *IJSCE (International Journal of Soft Computing and Engineering)*, vol. 6, no. 2, pp. 132-137, 2013.

[13] J. Phillips, "Jaccard Similarity and Shingling", 2013 [Online]. Available: https://www.cs.utah.edu/~jeffp/teaching/cs5955/L4-Jaccard+Shingle.pdf, [Accesed: 19-Jun-2016].

[14] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques Third Edition,* San Francisco: Elsevier Inc, 2012.