Hybrid Keyword Extraction Algorithm and Cosine Similarity for Improving Sentences Cohesion in Text Summarization

Rizki Darmawan
Informatics Engineering Graduate Program, STMIK Eresha
rizkidmw@gmail.com

Romi Satria Wahono
Faculty of Computer Science, Dian Nuswantoro University
romi@romisatriawahono.net

Abstract: As the amount of online information increases, systems that can automatically summarize text in a document become increasingly desirable. The main goal of a text summarization is to present the main ideas in a document in less space. In the create text summarization, there are two procedures i.e. extraction and abstraction procedure. One of extraction procedure is using keyword extraction algorithm which is easier and common but has problem in the lack of cohesion or correlation between sentences. The cohesion between sentences can be applied by using a cosine similarity method. In this study, a hybrid keyword extraction algorithm and cosine similarity for improving sentences cohesion in text summarization has been proposed. The proposed method is using compression 50%, 30% and 20% to create candidate of the summary. The result shows that proposed method affect significant increasing cohesion degree after evaluated in the t-Test. The result also shows that 50% compression ratio obtains the best result with Recall, Precision, and F-Measure are 0.761, 0.43 and 0.54 respectively; since summary with compression ratio 50% has higher intersection with human summary than another compression ratio.

Keywords: text summarization, keyword extraction, cosine similarity, cohesion

1 INTRODUCTION

As the amount of online information increases, systems that can automatically summarize one or more documents become increasingly desirable. Recent research has investigated types of summaries, method to create them, and methods to evaluate them (Hovy & Lin, 1999). It is necessary that the end user can access the information in summary form and without losing the most important aspects presented therein. Some of the application areas of the generation of extractive summaries from a single document are the summaries of web pages presented on the search engines (Porselvi & Gunasundari, 2013). Frequent workshop and symposia in text summarization reflect the ongoing interest of the researchers around the world.

The main goal of a summary is to present the main ideas in a document in less space. If all sentences in a text document were of equal importance, producing a summary would not be very effective, as any reduction in the size of a document would carry a proportional decrease in its informative of document (Hovy & Mckeown, 2001). Luckily, information content in document appears in bursts, and one can therefore distinguish between more and less informative segments.

The method for creating the summary can be divided into two ways: manually and automatically. Text summarization is a method to automatically summarize the text. In the create text summarization, there are two procedures i.e.: extraction and abstraction (Das, 2007). Extraction is a procedure used to create a summary by taking important sentences word by word that comes from the text, while abstraction is a procedure that is used to create a summary by information fusion, sentence compression and reformulation (Aliguliyev, 2009).

Text summarization with extraction procedure called extract summarization is easier to create than using abstraction Extractive procedure are usually performed in three step create an intermediate representation of the original text, sentence scoring and select high scores sentences to summary. There are several method that use in extractive procedure such as Keyword Extraction, Naïve-Bayes, Hidden Markov Models, Graph Method, Latent Sematic Indexing (Das, 2007).

Keyword extraction is an important technique for document retrieval, web page retrieval, document clustering, summarization, text mining, and so on (Rajman, 1998). By extracting appropriate keywords, we can easily choose which document to read to learn the relationship among documents. A popular algorithm for indexing is the TF/IDF measure, which extracts keywords that appear frequently in a document, but that don't appear frequently in the remainder of the corpus. The term "keyword extraction" is used in the context of text mining, for example (Rajman, 1998). A comparable research topic is called "automatic term recognition" in the context of computational linguistics and "automatic indexing" or "automatic keyword extraction" in information retrieval research. Recently, numerous documents have been made available electronically. Domain independent keyword extraction, which does not require a large corpus, has many (Ishizuka, 2003).

The first step creates a representation of the document. Usually, it divides the text into paragraphs, sentences, and tokens. Sometimes some preprocessing, such as stop word removal is also performed. The second step tries to determine which sentences are important to the document or to which extent it combines information about different topics, by sentence scoring (Ferreira et al., 2013). Usually, abstractive summarization requires heavy machinery for language generation and is difficult to replicate or extends to broader domain (Das, 2007).

Keyword Extraction Algorithm is easier and common in extract summarization. Yet the keyword extraction algorithm has problem in the lack of cohesion or correlation between sentences (Nandhini & Balasundaram, 2013) (Mendoza, Bonilla, Noguera, Cobos, & León, 2014) (Ishizuka, 2003). The correlation between sentences can be seen from the relationship between sentences and extent to which the ideas in the text are expressed clearly and relate to one another in a

systematic fashion by avoiding a confusing jumble of information (Nandhini & Balasundaram, 2013).

One way to resolve the problem of cohesion between sentences in extract summary is with determine the optimal combination between sentences (Fattah & Ren, 2009). The determination and cohesion optimization can be applied by using a cosine similarity method (Bestgen & Universit, 2006). The function for similarity measure should be easy to compute, it should implicitly capture the relatedness of the documents, and it should also be explainable (Rafi & Shaikh, 2010). The similarity between two sentences, according to the vector representation described is calculated as the cosine similarity (Manning, Raghavan, & Schlutze, 2009).

The objective of this work is to improve cohesion in text summarization by keyword extraction algorithm using cosine similarity method. Finally, our work of this paper is summarized in the last section.

2 RELATED WORKS

Many studies have been published in cohesion problem for text summarization in some approach like using optimal combination for the summarization (Mendoza, 2013; Nandhini, 2013) and another technique that concern with cohesion in text summarization.

Mendoza et al. (2013) proposed is combined the population based global search with a local search heuristic (memetic approach). The local search heuristic exploits the problem knowledge for redirect the search toward best solution. The objective function for this method is defined formed by the features like cohesion which proved effective in selecting relevant sentences from a document. The best results of MA-SingleDocSum evaluated with ROUGE-1 and ROUGE-2 is 8.59% with DUC 2001.

Nandhini et al. (2013) work to extract the optimal combination of sentences that increase readability through sentence cohesion using genetic algorithm. The results show that the summary extraction using their proposed approach performs better in *F*-measure, readability, and cohesion than the baseline approach (lead) and the corpus-based approach. In the case of 10% compression rate the F-measure is 0.284, 20% compression is 0.466 and 30% compression is 0.502. The best F-measure is 30% compression

Smith et al. (2011) work to measure cohesion is automatically through the amount of co-references in the text and how intact the text is after summarization. They compare four different types of techniques (Every3, 100First, CogSum, PrevSum) were used to create the summaries. The results proved that the summary produced by a traditional vector space-based summarizer is not less cohesive than a summary created by taking the most important sentences from the summarizer. Comparing the cohesion there are significances, for instance, for broken references the 100First is significantly better than all the other ($p < 0.001$) is 0.459.

Silber et al. (2002) present a linear time algorithm for lexical chain computation. The algorithm makes lexical chains computationally feasible candidate as an intermediate representation for automatic text summarization By using lexical chains, they can find statistically the most important concepts by looking at the structure in the document rather than the deep semantic meaning. Lexical chains appropriately represent the nouns in the summary is 79,12%.

3 PROPOSED METHOD

The proposed model using keyword extraction algorithm with compression ratio parameter and combining with cosine similarity for conducting this experiment. Cosine similarity is used to re-arrange sentence extraction from the result of keyword extraction algorithm process.

The keyword extraction algorithm using calculation based on TF/IDF, weight a given term to determine how well the term describes an individual document within a corpus. It does this by weighting the term positively for the number of times the term occurs within the specific document, while also weighting the term negatively relative to the number of documents which contain the term. Consider term t and document d, where t appears in n of N documents in D. The TF-IDF function is of the form as follows:

$$TFIDF\ (t,d,n,N) = TF\ (t,d)\ x\ IDF\ (n,N)$$

When the TF-IDF function is run against all terms in all documents in the document corpus, the words can be ranked by their scores. A higher TF-IDF score indicates that a word is both important to the document, as well as relatively uncommon across the document corpus. This is often interpreted to mean that the word is significant to the document, and could be used to accurately summarize the document. TF-IDF provides a good heuristic for determining likely candidate keywords, and it (as well as various modifications of it) has been shown to be effective after several decades of research.

Cosine similarity is a measure of similarity between two vectors of n dimensions by finding the cosine of the angle between them, often used to compare documents in text mining (Satya & Murthy, 2012). Given two vectors of attributes, A and B, the cosine similarity, θ, is represented using a dot product and magnitude as:

$$Similarity = Cos\ \theta = \frac{A.B}{|A||B|}$$

The resulting similarity ranges from 0 with usually indicating independence, and 1 with usually indicating exactly the same and in between those values indicating intermediate similarity and dissimilarity. For the text matching, the attribute vector A and B are usually the term frequency vectors of the documents. In the case of information retrieval the cosine similarity of two documents will range 0 to 1, since the term frequencies (TF-IDF weights) cannot be negative. The angle between two term frequency vectors cannot be greater than 90^0.

In Figure 1 can be explained that after the data from UCI Reuters- 21578 completed prepared then the data will be tested into summarization stage.

The summarization stage consists of three component i.e. keyword extraction algorithm, compression ratio selector and cosine similarity method. These three component will summarize the text were feeding as the result final text were summarized.

The first pre-processed document is tokenized by keyword extraction algorithm and then calculates TF/IDF for each term. Then sum all of TF/IDF term for each sentence and get sum of each sentence the next process is rank all of sentence based on sum of TF/IDF. The compression ratio determine the position of sentence rank. In this study using a compression of 50% that means the sentence summary shrinkage 50% from the original text. After sentence is selected then perform calculation of their similarity with cosine similarity method. After the calculation of cosine similarity, the next process is re-arranging all of

sentence based on cosine similarity from the highest to the lowest similarity. This new text with new sentence arrangement will be the final summarized text.

Extractive summary can be evaluated using various characteristic such as F-measure and cohesion (Nandhini & Balasundaram, 2013b). F-Measure is measuring how far the technique is capable of predicting of correct sentence. Evaluation can be classified into intrinsic and extrinsic evaluation (Nandhini, 2013). Intrinsic evaluation judges the summary quality by its coverage between machine-generated summary and human generated summary. Extrinsic evaluation focuses mainly on the quality by its effect on other tasks. In intrinsic evaluation, Precision (P), recall (R), and F-measure (F) are used to judge the coverage between the manual and the machine generated summary:

$$P = \frac{|S \cap T|}{|S|}$$

$$R = \frac{|S \cap T|}{|T|}$$

$$\mathrm{F} = \frac{|2 * \mathrm{P} * \mathrm{R}|}{|\mathrm{R} + \mathrm{P}|}$$

Where S is the machine generated summary and T is the manual summary (Nandhini & Balasundaram, 2013b). For the cohesion evaluation, we can measure with the formula as follows:

$$CoH = \frac{Log(C_s * 9 + 1)}{Log\ (M * 9 + 1)} \qquad Ns = \frac{(o) * (o-1)}{2}$$

$$C_s = \frac{\Sigma_{\forall S_i, S_j \in Summary}\ Sim_{cos}\ (S_i, S_j)}{N_s}$$

$$M = \max Sim_{cos}\ (i,j)\ ,\ i,j\ \leq N$$

$$Ns = \frac{(o) * (o-1)}{2}$$

Where CoH corresponds to the cohesion of a summary, Cs is the average similarity of all sentences in the summary S, Sim_{cos}(Si,Sj) is the cosine similarity between sentences Si and Sj, Ns is the number of nonzero similarity relationships in the summary, O is the number of sentences in the summary, M corresponds to the maximum similarity of the sentences in the document and N is the number of sentences in the document. In this way, CoH tends to zero when the summary sentences are too different among them, while that CoH tends to one when these sentences are too similar among them. Thus, this feature tends to favor the summaries that contain sentences about the same topic (Mendoza et al., 2014).

The dataset used in this research is collected from UCI Dataset containing documents of Reuters-21578 that has collection appeared on the Reuters newswire in 1987. The documents were assembled and indexed with categories by personnel from Reuters Ltd. (Sam Dobbins, Mike Topliss, and Steve Weinstein) and Carnegie Group, Inc. (Peggy Andersen, Monica Cellio, Phil Hayes, Laura Knecht, Irene Nirenburg) in 1987. The detail dataset can be downloaded at https://archive.ics.uci.edu/ml/datasets/Reuters21578+Text+Categorization+Collection.

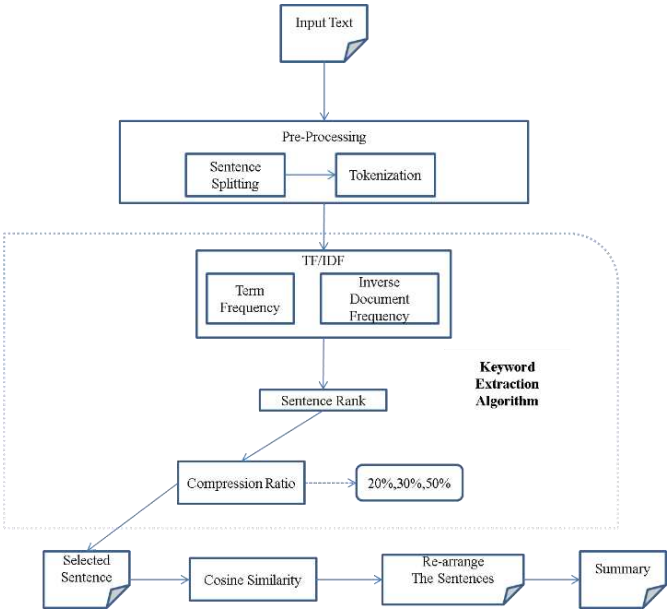

Figure 1. Block Diagram Proposed Model

4 EXPERIMENTAL RESULTS

The research using computer platform with specification based on Intel Core i3 2.30 Ghz CPU, 2 GB RAM, and Microsoft Windows 7 Ultimate 32 Byte. The software is using Java with Netbeans IDE 7.3.1.

Evaluation of the results is the calculation of recall, precision and F-measure. It can be seen that the lowest recall at dataset 6 that is equal to 0.484 and the highest recall on dataset 10 is equal to 0.909. The lowest precision is dataset 6 is equal to 0.284 and the highest precision on dataset 2 is equal to of 0.685. While the lowest F-measure at dataset 6 that equal to 0.358 and highest F-measure at dataset 2 that equal to 0.748. It is shown in Table 1.

Table 1. Recall-Precision of Summary with Compression 50%

Dataset	Recall	Precision	F-Measure
Dataset 1	0.771	0.492	0.600
Dataset 2	0.824	0.685	0.748
Dataset 3	0.908	0.478	0.626
Dataset 4	0.565	0.565	0.565
Dataset 5	0.635	0.328	0.433
Dataset 6	0.484	0.284	0.358
Dataset 7	0.888	0.381	0.532
Dataset 8	0.861	0.331	0.478
Dataset 9	0.772	0.392	0.520
Dataset 10	0.909	0.454	0.606

In compression summary 30% can be seen that the lowest recall at dataset 5 that is equal to 0.418 and the highest recall on dataset 8 is equal to 0.907. The lowest precision is dataset 5 is equal to 0.295 and the highest precision on dataset 2 is equal to of 0.666. While the lowest F-measure at dataset 5 that equal to 0.346 and highest F-measure at dataset 10 that equal to 0.690 as shown in Table 2.

Table 2. Recall-Precision of Summary with Compression 30%

Dataset	Recall	Precision	F-Measure
Dataset 1	0.554	0.464	0.505
Dataset 2	0.702	0.666	0.684
Dataset 3	0.653	0.444	0.528
Dataset 4	0.526	0.412	0.462
Dataset 5	0.418	0.295	0.346
Dataset 6	0.453	0.397	0.423
Dataset 7	0.688	0.428	0.525
Dataset 8	0.907	0.561	0.694
Dataset 9	0.555	0.458	0.478
Dataset 10	0.863	0.575	0.690

In compression summary 20% can be seen that the lowest recall at dataset 2 that is equal to 0.148 and the highest recall on dataset 10 is equal to 0.863. The lowest precision is dataset 2 is equal to 0.215 and the highest precision on dataset 10 is equal to of 0.647. While the lowest F-measure at dataset 5 that equal to 0.176 and highest F-measure at dataset 10 that equal to 0.740 as shown at Table 3.

Table 3 Recall-Precision of Summary with Compression 20%

Dataset	Recall	Precision	F-Measure
Dataset 1	0.253	0.538	0.344
Dataset 2	0.148	0.215	0.176
Dataset 3	0.306	0.329	0.317
Dataset 4	0.434	0.412	0.423
Dataset 5	0.459	0.459	0.459
Dataset 6	0.406	0.522	0.456
Dataset 7	0.666	0.424	0.497
Dataset 8	0.907	0.678	0.776
Dataset 9	0.469	0.584	0.521
Dataset 10	0.863	0.647	0.740

For the 50 % compression the highest recall in summary of dataset 10 and lowest recall in dataset 6, while the highest precision in summary of dataset 2 and lowest precision in summary of dataset 6. The highest F-measure of 50 % compression in summary of dataset 2 and the lowest F measure in summary of dataset 6.

For the 30 % compression the highest recall in summary of dataset 8 and lowest recall in summary of dataset 5, while the highest precision in summary of dataset 2 and lowest precision in summary of dataset 5. The highest F-measure of 50 % compression in summary of dataset 10 and the lowest F measure in summary of dataset 5. For the 20 % compression the highest recall in summary of dataset 8 and lowest recall in summary of dataset 2, while the highest precision in summary of dataset 8 and lowest precision in summary of dataset 2. The highest F-measure of 50 % compression in summary of dataset 10 and the lowest F measure in summary of dataset 5. Overall of that analysis is shown in Table 4.

Table 4. Overall Analyses of Recall, Precision And F-Measure

	Recall		Precision		F-Measure	
	Highest	Lowest	Highest	Lowest	Highest	Lowest
Compression 50 %	Data set10	Data set 6	Data set 2	Data set 6	Data set 2	Data set 6
Compression 30 %	Data set 8	Data set 5	Data set 2	Data set 5	Data set 10	Data set 5
Compression 20 %	Data set10	Data set 2	Data set10	Data set 2	Data set10	Data set 5

The main factor of that performance is how much the intersection against human summary because it related to the equation of recall and precision. If intersection is high, automatically make the high result, although length of word in machine and human has big influence contribution to the result. This study result also confirm some studies that intersection between human summary and machine play big influence for evaluation measurement such as recall, precision and F-measure (Conroy, 2001).The comparison of average recall, precision and F-measure is shown in Table 5 and Figure 2.

Table 5. Comparison of Average Recall, Precision and F-Measure

Compression	Recall	Precision	F-Measure
50%	0.761	0.439	0.547
30%	0.625	0.470	0.533
20%	0.484	0.481	0.471

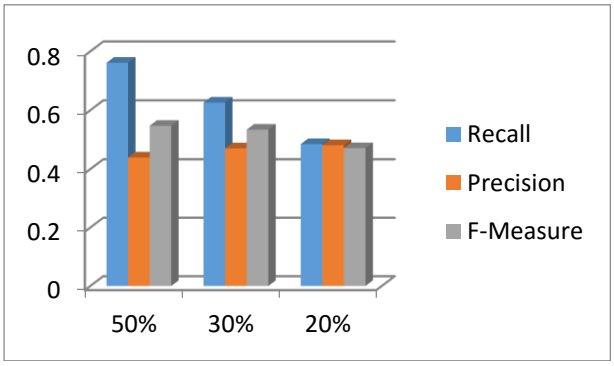

Figure 2. Average Recall, Precision and F-Measure Diagram

From the data that shown in Table 5, it's shown that the best F-measure is 50% compression that has value is 0.547. It's because they have highest intersection than the other compression that compare with human summary. The results also reflect that summary with 50% compression is the better summary than the others. Another study also reflect that higher compression has higher average of recall, precision and F-measure (Nandhini & Balasundaram, 2013b) and this result also confirmed by Ferreira et al (2014) that the best result summary is 50 % compression (Ferreira et al., 2014)

To prove whether there are differences in the degree of cohesion after using the cosine similarity method is using t-test models. A significant difference in performance is considered when the results of t-test showed that (P <= t) < alpha (0.05). T-test of the statistical test on the summary results that using the cosine similarity method and without using the cosine similarity method is shown in the Table 6.

Table 6. T-Test: Paired Two Sample for Means
Of Cohesion Degree

	Without Cosine Similarity	*Cosine Similarity*
Mean	31.92255942	**35.42168762**
Variance	33.59318702	40.794073
Observations	10	10
Pearson Correlation	0.968118831	
Hypothesized Mean Difference	0	
df	9	
t Stat	-6.721927271	
P(T<=t) one-tail	4.3178E-05	
t Critical one-tail	1.833112933	
P(T<=t) two-tail	**8.63559E-05**	
t Critical two-tail	2.262157163	

From Table 6, it shows the average of cohesion degree of summary that using the cosine similarity method is higher than without using cosine similarity that has value is 35.42168762 with P value = 8.63559E-05 . The significance level is set to be 0.05. It means that cohesion degree in summary using cosine similarity and without using cosine similarity have significant differences (P value < 0.05). Therefore, it can be concluded that summary with cosine similarity method makes an improvement when compared with summary without using cosine similarity in cohesion degree.

The best average F-measure of summary in three compressions is 50% compression. According to another study that using compression ratio to get the result, also reflect that highest compression ratio has best F-measure (Nandhini & Balasundaram, 2014). One reason to explain about this phenomena is intersection human summary and machine summary is higher according to compression ratio. Intersection means that how many words in machine summary have same similarity with number of word in human summary. If intersection is high, automatically make the high result, although length of word in machine and human has big influence contribution to the result. This study result also confirm some studies that intersection between human summary and machine play big influence for evaluation measurement such as recall, precision and F-measure (Conroy, 2001).

From t-test result, summary that using cosine similarity has increased significantly in cohesion degree compared with the summary without using cosine similarity. The results of these experiments also show that the highest F-measure is compression of 50%. The result can be compared with another research like Nandhini & Balasundaram (Nandhini & Balasundaram, 2013b) and Aliguliyev (Aliguliyev, 2009) that increase of compression in order to increase of F-measure.

5 CONCLUSION

Recent research has investigated types of summaries, method to create them, and methods to evaluate them. It is necessary that the end user can access the information in summary form and without losing the most important aspects presented therein. Some of the application areas of the generation of extractive summaries from a single document are the summaries of web pages presented on the search engines. The main goal of a summary is to present the main ideas in a document in less space. If all sentences in a text document were of equal importance, producing a summary would not be very effective, as any reduction in the size of a document would carry a proportional decrease in its informative

In this research is used keyword extraction algorithm model with cosine similarity method that combined in some compression ratio. In the experiment is tested that keyword extraction algorithm using compression ratio of 20%, 30% and 50%. The best compression ratio from the extraction of keyword extraction algorithm is 50% with the F-measure is 0.761. In this research also shows there is different between summary with cosine similarity and without cosine similarity related to cohesion between sentences after tested with t-test, where summary with cosine is the best performance.

REFERENCES

Aliguliyev, R. M. (2009). Expert Systems with Applications A new sentence similarity measure and sentence based extractive technique for automatic text summarization. Expert Systems With Applications, 36(4), 7764–7772. doi:10.1016/j.eswa.2008.11.022

Bestgen, Y., & Universit, F. (2006). Improving Text Segmentation Using Latent Semantic Analysis. Association for Computational Linguistic, (2001).

Conroy, J. (2001). Matrix Decomposition 1 Introduction. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. (pp. 1–20). ACM.

Das, D. (2007). A Survey on Automatic Text Summarization Single-Document Summarization. Carnegie Mellon University, 1–31.

Fattah, M. A., & Ren, F. (2009). GA, MR, FFNN, PNN and GMM based models for automatic text summarization. Computer Speech & Language, 23(1), 126–144. doi:10.1016/j.csl.2008.04.002

Hovy, E., & Lin, C. (1999). Automated Text Summarization in Summarist. Asscociation for Computer Linguistic.

Hovy, E., & Mckeown, K. (2001). Summarization. Asscociation for Computer Linguistic, 28.

Ishizuka, M. (2003). Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information. International Journal on Artificial Intelligence Tools.

Manning, C., Raghavan, P., & Schlutze, H. (2009). Introduction to Information Retrieval (p. 581). Cambridge University.

Mendoza, M., Bonilla, S., Noguera, C., Cobos, C., & León, E. (2014). Expert Systems with Applications Extractive single-document summarization based on genetic operators and guided local search. Expert Systems With Appilications, 41(9), 4158–4169. doi:10.1016/j.eswa.2013.12.042

Miller, G. A., Beckwith, R., Fellbaum, C., & August, R. (1993). Introduction to WordNet : An On-line Lexical Database, (August).

Nandhini, K., & Balasundaram, S. R. (2013). Improving readability through extractive summarization for learners with reading difficulties. Egyptian Informatics Journal, 14(3), 195–204. doi:10.1016/j.eij.2013.09.001

Nandhini, K., & Balasundaram, S. R. (2014). Extracting easy to understand summary using differential evolution algorithm. Swarm and Evolutionary Computation, 1–9. doi:10.1016/j.swevo.2013.12.004

Porselvi, A., & Gunasundari, S. (2013). Survey on web page visual summarization. International Journal of Emerging Technology and Advanced Engineering, 3(1), 26–32.

Rafi, M., & Shaikh, M. S. (2010). An improved semantic similarity measure for document clustering based on topic maps. Computer Science Department Karachi Pakistan.

Rajman, M. (1998). Text mining – knowledge extraction from unstructured textual data. In In Proceedings of the 6th

Conference of International Federation of Classification Societies.
Satya, K. P. N. V, & Murthy, J. V. R. (2012). Clustering Based On Cosine Similarity Measure. International Journal of Engineering Science & Advanced Technology, 2(3), 508–512.
Silber, H. G. (2002). an Intermediate Representation for Automatic Text Summarization. Association for Computational Linguistic, 28, 1–11.
Smith, C., Danielsson, H., & Arne, J. (2011). Cohesion in Automatically Created Summaries. Santa Anna IT Research

BIOGRAPHY OF AUTHORS

Rizki Darmawan. Received M.Kom from STMIK ERESHA, Jakarta. He is an IT professional. His current research interests include information retrieval and machine learning.

Romi Satria Wahono. Received B.Eng and M.Eng degrees in Computer Science respectively from Saitama University, Japan, and Ph.D in Software Engineering from Universiti Teknikal Malaysia Melaka. He is a lecturer at the Graduate School of Computer Science, Dian Nuswantoro University, Indonesia. He is also a founder and chief executive officer of Brainmatics, Inc., a software development company in Indonesia. His current research interests include software engineering and machine learning. Professional member of the ACM and IEEE Computer Society.