Journal of Theoretical and Applied Information Technology
20th June 2015. Vol.76. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

DOCUMENT SIMILARITY DETECTION USING SYNONYMS FACTOR AND PARALLEL PROCESSING

[1]KEMAL ADE SEKARWATI, [2]LINTANG YUNIAR BANOWOSARI,

[3]I WAYAN S. WICAKSANA, [4]ANA KURNIAWATI

[1]Gunadarma University, Department of Information System
[2]Gunadarma University, Department of Information Management
[3]Gunadarma University, Department of Information System
[4]Gunadarma University, Department of Information System
E-mail: [1]Ade@staff.gunadarma.ac.id, [2]lintang@staff.gunadarma.ac.id
iwayan@staff.gunadarma.ac.id ana@staff.gunadarma.ac.id

ABSTRACT

Research regarding Indonesian language document similarity measurement has not yet broad to do. Mostly those researches using Karp Rabin algorithm and string matching method, and the documents used in the form of abstracts and articles which only has one page.
This study focused on measuring similarity detection of Indonesian language documents using synonyms factor for more than one page document and the processing speed measurement. The system developed is to measure the similarity of existing documents with other documents that are stored in an internal database. Similarity calculation results in the form of a percentage of the document similarity comparison. The measurement results of document processing speed in the form of speed detection in processing the documents. The calculation of the similarity detection measurement and its detection speed is performed using the following steps: (i) examine the documents title, (ii) Distribution of work, (iii) the document similarity measurement and (iv) speed measurement of document similarity detection process.
Tests carried out using Indonesian documents that are larger than one page. Documents that have been tested were 15 documents. Test results to calculate document similarities and detection speed conducted on four types of documents. This study has shown that the algorithm used can check the similarity of documents with the maximum number of pages is 56 pages. From the speed of the detection process shows that the speed measurement in detection process algorithm is also said to be successful.

Keywords: *Document, Similarity, Processing Speed, Parallel Processing.*

1. INTRODUCTION

This guide provides details to assist authors in preparing a paper for publication in JATIT so that there is a consistency among papers. These instructions give guidance on layout, style, illustrations and references and serve as a model for authors to emulate. Please follow these specifications closely as papers which do not meet the standards laid down, will not be published.

Document similarity detection approach is the approach measuring documents similarity between suspected documents document with the original document. There exist researches in Indonesian language document similarity measurement but not a lot. Research in Indonesian language document similarity, such as performed by Firdaus 2003 [4] which uses Karp Rabin algorithm. Objects document used by the Hari Bagus Firdaus is Indonesian language documents. Research done by Sinta Agustina 2008 [8] was also using Indonesian language document using Karp Rabin algorithmto detect document similarity. Ana Kurniawati 2010 [2] using string matching method when she designed the new algorithm to determines the sentence structure and calculates the similarity of documents by a factor of synonyms. Object used is Indonesian language documents.

In the study of Ana Kurniawati, the data used was scientific writing abstract of student in computer science and news articles or paper taken from online or internet media. The study of news articles can only be done for a single page, for the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

next page is not successful. The study used data only 1 (one) page, has not been able to measure the similarity when compared to sentences has element expanded or have a child sentence, and have not been able to measure the similarity if there is a passive sentence.

Document similarity detection with large data if done using a single computer takes a long time. To increase the speed of processing large data used parallel processing. Parallel processing can shorten the execution time of a program by dividing a program into parts that are smaller that can be done at each processor simultaneously [3].

In the Central Manager algorithm, load system considered in distribution of the work [5]. This condition makes the algorithm Central Manager good in acceleration program execution time. Load index is used as a determining factor for the performance of this algorithm is the Central Processing Unit (CPU) or hardware I / O, because the CPU and hardware I / O have an important contribution to the execution time. Based on the load distribution, the Central Manager algorithm is using the CPU or memory load index for distributing the work load. For memory load, the Central Manager Algorithm which uses memory load index is the best algorithm that distributes the load evenly. Based on program execution time, Central Manager Algorithms using the CPU or hardware I / O as the load index is the best algorithm that provides the fastest program execution time.

Based on what has been stated above, we need a system development in Indonesian language document similarity detection using synonyms factor in order to overcome the problem in previous studies that have not been able to detect similarities with large documents (more than 1 page). To increase the speed of large data processing, it needs to improve the algorithms and (or) using parallel processing. Central Manager Algorithm is used for speed up execution the program. The tools developed are expected to improve the performance of scientific journal or papers similarity detection speed, most of which have ranged from 10 pages to 15 pages.

2. LITERATURE REVIEW

2.1 Indonesia Language Document Similarity Measurement

Research for measuring the similarity of existing Indonesian language documents has already existed, such as performed by Bagus Hari Firdaus 2003 [4] which uses Karp Rabin algorithm. Sinta Agustina 2008 [8] and Ana Kurniawati 2010 [2] who use string matching method when design the new algorithm to determine the sentence structure and calculate the similarity of documents by a synonyms factor. Object used is Indonesian language documents. Table 1 as follows presented comparative studies in document similarity detection.

2.2 Load Balancing

Load balancing is the process of improving system performance through the distribution of inter-processor performance. Load balancing involves the distribution of jobs across the computer network system, thereby increasing the throughput (data transfer successfully from one place to another in a given time period) without having to obtain additional computer hardware faster [6]. Load balancing is a technique that is applied to a parallel system that is used for obtain optimal system conditions, such as the workload of the average distributed among computers, and by implication will reduce program execution time [5]. Load balancing ensures that each processor in the system perform more or less the same amount of work at any point of time. In distributed systems and parallel systems, more than one processor in parallel processing program. The amount of time required to run the entire process that has been assigned to a processor called the processor workload [7].

2.3 Types of Load Balancing Algorithm

Load balancing algorithm is divided into two categories based on the current state of the system is static load balancing algorithms and dynamic load balancing algorithm. [6].

2.3.1. Static Load Balancing Load Balancing Algorithm

A characteristic of static load balancing is the job distribution is performed before the execution of the program begins. In the static load balancing, processor performance is measured before the program is executed. The job distribution on a parallel system is performed by the master processor based on metrics processor performance. Slave processor executes the job that has been given by the master processor and if the job has been completed, the execution results are sent to the master processor. Objective of static load balancing methods is to reduce the execution time and minimize communication time delay. In general, weakness of this static approach is the final selection of the host for the allocation process is

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

carried out when the process is created and cannot be changed during the execution of the process to make changes to the system load. There are four types of static load balancing algorithm that are [6]: Round Robin algorithm, randomized algorithms, Central Manager Algorithm, and Threshold algorithm.

2.3.2. Dynamic Load Balancing Algorithm

In the dynamic load balancing algorithm workload is distributed among the processors at the time the program is run. Master processor determines the new process to the slave processor is based on a collection of new information [1]. In distributed systems, dynamic load balancing can be done in two ways: a distributed and non-distributed. In a distributed manner, dynamic load balancing executed by all nodes in the system and load balancing task is divided into the entire node. Interaction between nodes for load balancing is divided into two forms of cooperative and non-cooperative. In cooperative nodes, nodes work side by side to achieve common goals, such as improving the overall response time, etc. In non-cooperative, each node work alone to achieve local objectives, such as improving the response time of the local task. Dynamic load balancing algorithms generate more messages than non disributed because every node on the system requires communication with each node. The advantage is that even if one or more nodes in a system failure, this does not because the whole process of load balancing stops instead will cause the system to improve its performance. In this type of non distributed, either one or a group of nodes perform load balancing tasks. Non-distributed load balancing algorithm consists of two forms of centralized distributed and semi-distributed. In Centralize forms, load balancing algorithms executed by only one node in the whole system: the central node. This node is only intended for load balancing of the whole system. Other nodes only interact with the central node.

In the semi-distributed form, node on this system is partitioned into a cluster form with the appropriate selection techniques that maintain load balancing inside the cluster. Therefore, the load balancing of the whole system is done through the center of each cluster node. Centralized dynamic load balancing take fewer messages to get a decision because the total number of interactions in the system dramatically decreased compared with the case of semi-distributed. However, centralized algorithms can lead to bottlenecks in the system at the central node and also the process of load balancing is not useless after the central node is damaged. Therefore, this algorithm is suitable for small-sized networks. There are two types of dynamic load balancing algorithm that are: Local Queue algorithms and Central Queue algorithms. [6]

3. RESEARCH METHOD

3.1 Architecture of Speed Measurement and Document Similarity Detection

Architecture of speed detection and document similarity measurement at phase's process consists of four main components:

1. Examine Document Title
   The process of checking the document title is done by the master computer. Title documents examined whether it has similarities with the title of the document contained in the papers database.
2. Job Distribution.
   Job distribution process is performed by the master computer.
3. Similarities Measurement.
   Similarity measurement process performed by the slave computer consists of five sub-processes. The first sub-process is a pre-processing subprocess, the second sub-process is a segregation text into sentences sub process, the third sub-process is segregation the sentence into words sub-process, the fourth sub-process is a sub-process analysis of synonyms, and fifth sub-process is the process of document similarity calculation. The results of the documents similarity calculation process are sent to the master computer.
4. Speed of Document Detection Process.
   Output calculation process performed by the master computer is composed of three sub-processes. The first sub-process is the calculation of the documents similarity average. The output of the calculation of the documents similarity average is in the form of a percentage of similarity. The second sub-process is the calculation of the speed of the detection process documents. The output of the calculation document detection process speed is in second. The third sub-process is calculating the total sentence which is similar.

Architecture of the speed detection measurement and of document similarity can be seen in Figure 1 below.

Journal of Theoretical and Applied Information Technology
20th June 2015. Vol.76. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

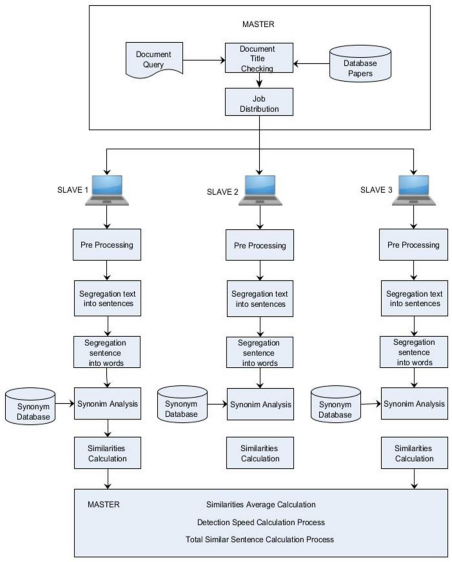


Figure 1 Architecture Of Detection Speed And Document Similarity Measurement

Here is an explanation of the architecture of detection speed and document similarity measurement

1. Document Query
   Queries of this document aims to take the documents to be checked, whether the documents have title similarities with documents titles contained in the database. Documents are checked on the local storage media.
2. Document Title Checking
   The calculation result of document title checking in the form of a percentage of similarity. Calculations to find title similarities use Jaro Winkler formula. In the process of title checking it has not seen its similarities word. To obtain the similarity percentage, then the calculation result multiplied by 100%. The document title similarity calculation formula as follows:

$$dj = \frac{1}{3} X \left(\frac{m}{|S_1|} + \frac{m}{|S_2|} + \frac{m-t}{m} \right) X\ 100\% \ldots\ldots [1]$$

   Formula description:
   d_j = Similarities between 2 (two) sentences, which are S_1 and S_2
   m = the number of words which are equal or similar at the sentence1 and sentence2
   $|S_1|$ = Length of the sentence 1
   $|S_2|$ = Length of the sentence 2
   t = the number of transposition
3. Database Creation
   There are two databases used in this architecture. The first database is a database that contains the papers from student scientific writing that have experienced pre-processing process in advance. The process of pre-processing carried out is to change the document from a *pdf* to *txt* form. If the document contains images, tables or kind, and includes the formula images, tables, and the formula will be deleted. The second database is a synonym database that contains a synonym of verbs and nouns. Synonym database consists of one table.
4. Job Distribution.
   On the job distribution, the existing process is inter-exchanging data. Once the process is complete title check, the master computer waiting for a message from the slave computer does the job. If the slave computer is not doing the job, then the slave computer will send a message about the status of the job. Then the master computer would distribute the job to the slave computers. The process of sending a message from the master to the slave computer is using JSON format.
5. Pre Processing.
   At this stage it does the removal of symbols and punctuation other than the points that are on the content of the document that is being detected.
6. Segregation text into sentences.
   At this stage, the process of separation the sentence contained in the documents examined. Segregation this sentence is using dot (.) as separator between the sentence.
7. Segregation sentence into words.
   At this stage, the decomposition of the sentence into words by using a separator between words such as space or horizontal tabulation.
8. Synonym analysis.
   At this stage, the process of examination sentences synonyms examined. This synonym examination refers to the synonym database. In this process each word in the sentence in a document that is checked compared to the words contained in the synonyms database.
9. Similarities Calculation.
   At this stage, the process of comparing sentences in a document that is checked by the sentence contained in paper documents in the database. Documents examined by the slave computers. Each slave computer will temporarily store the results of the calculation

Journal of Theoretical and Applied Information Technology
20th June 2015. Vol.76. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

of these parts. The calculation results of the similarity in the form of a percentage of documents similarity. This similarity calculation is using the Jaro Winkler formula. To obtain the percentage of similarity, then the calculation result multiplied by 100%. Document similarity calculation formula as follows:

$$dj = \frac{1}{3} X \left(\frac{m}{|S_1|} + \frac{m}{|S_2|} + \frac{m-t}{m}\right) X\ 100\% \text{.....[2]}$$

Formula description:
d_j = Similarities between 2 (two) sentences, which are S_1 and S_2
m = the number of words which are equal or similar at the sentence1 and sentence2
$|S_1|$ = Length of the sentence 1
$|S_2|$ = Length of the sentence 2
t = the number of transposition

10. Similarities average calculation
The process of calculating the similarity average of documents is performed by the master computer. The results obtained in the previous process which are the stages to calculate similarity carried out by the slave computers will be sent to the master computer. Then the results will be divided with five that showing the number of parts contained in a document that is part of an abstract, introduction, methods, results, and conclusions. Similarity average calculation using the following formula:

$$RK = K_{ij} / 5 \quad \text{.....[3]}$$

Formula description:

RP = similarity average
K_{ij} = similarity of documents that are checked against the document contained in the database

11. Detection speed calculation process.
At this stage the calculation of the document detection process speed is performed by parallel processing. The speed calculation process is using the computer clock and the results of calculation in the second denomination.

12. Total similar sentence calculation process
At this stage performed the number of similar sentences calculation from a document that is checked against the documents contained in the papers database.

4. RESULT AND DISCUSSION

4.1. Testing Preparation

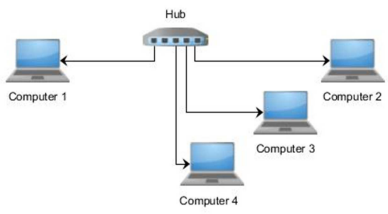
Tests preparation can be divided into three, namely:

1. The data used for the test consists of two kinds of documents. The first data are in the form of a collection of scientific writing of students majoring in Information Engineering Gunadarma University that have been modified. While the second data is a collection of scientific writing of students majoring in Information Engineering Gunadarma University are consisting of 18 writings. Scientific writing is stored in a database.
2. Hardware needed to test this data include: hubs, network cable, and 4 units of computers. Here is the Table 2 are presented the specifications of the hardware used to test the data:

Table 2 Hardware Spesification

No	Hardware	Specification
1	Hub	1 Gb
2	Network Wire	UTP Category 6
3	Computer 1	Prosesor : Intel® Core I5-M520 @2.40 GHz Memori : 4 Gb DDR3 Harddisk : 500 Gb
4	Computer 2	Prosesor : Intel® Core I5-2450M @2.50 GHz Memori : 4 Gb DDR3 Harddisk : 750 Gb
5	Computer 3	Prosesor : Intel® Core I5 @2.30 GHz Memori : 4 GB DDR3 Harddisk : 500 GB
6	Computer 4	Prosesor : Intel® Pentium P6300 @2.27 Ghz Memori : 3 Gb Harddisk : 500 Gb

3. Network Configuration.
Network configuration needed to test the speed of parallel processing carried out during data testing. Figure 2 below is a network configuration to test the data:



ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Figure 2 Network Configuration

Computer 1 acts as a master computer that has the function of providing a data processing job to the slave computers namely computer 2, computer 3, and computer 4.

4.2 Testing Scenario

Testing scenario carried out as follows:

1. Documents used a total of 18 documents. Documents to be tested consist of 15 combinations of pairs of text type and files are in a folder master computer. The size of the documents used between 20-68 kilobytes (Kb). Number of documents pages used between 7-36 pages. The number of sentences used is 139-529 words.
2. Perform document similarity testing. In this test performed comparison between the sentences in a document that is checked by the sentence contained in the document in the paper database. The algorithm calculates the similarity by perform the synonyms analysis process that examine the words contained in a sentence whether there is a word synonymous. A synonym testing is referring to the synonym table. In this process carried out a comparison between the words contained in the sentences in a document that is checked by the words contained in the sentence in the document reside in papers database. The result of the similarity calculation is in the form of a percentage of documents similarity.

   Tests implementation conducted on four combinations types of documents. The combination of documents being tested as follows:
   1. Documents that are not modified.
   2. The modified document that is by move some sentences positions.
   3. The document is modified by removing some sentences positions and replaces a few words with synonyms.
   4. The document is modified by removing some sentence positions, replace a few words with synonyms, and move the location of words in a sentence.
3. Conduct document processing testing in parallel. The test aims to determine the parallel processing speed of document detection in each computer configuration used. The result is a speed within seconds.

   Testing implementation is using varying computer configuration to see processing speed detection on each computer configuration. Computer configuration is used as follows:
   1. Configuration 1 consists of 3 computers is computer 1, computer 2, and the computer 3. Computer 1 acts as the master computer while the computer 2 and the computer 3 acts as a slave computer.
   2. Configuration 2 consists of 4 computers that is computer 1, computer 2, computer 3, and computer 4. Computer 1 acts as a master computer while the computer 2, computer 3, and 4 computer acts as a slave computer

4.3 Testing Result

The first test performed on documents 1 and 2 documents is not modified. The first test results are shown in Table 3.

Table 3 First Test Result

No	Doc 1	Similar Sentence	Similarity %	Doc2	Similar Sentence	Similarity %	Number of Page	Total	
								2L (sec)	3L (sec)
1	102	495	100	103	393	100	43	131	74
2	103	393	100	104	502	100	42	198	88
3	104	502	100	105	382	100	40	82	73
4	105	382	100	108	437	100	42	158	97
5	105	382	100	109	391	100	38	512	152
6	108	437	100	109	391	100	42	283	157
7	109	391	100	110	473	100	41	366	241
8	111	529	100	112	478	100	50	450	163
9	112	478	100	113	362	100	44	310	210
10	113	362	100	114	274	100	42	111	77
11	114	274	100	115	139	100	28	108	44
12	115	139	100	116	449	100	39	42	28
13	118	160	100	119	452	100	49	216	210
14	121	365	100	123	263	100	44	38	25
15	123	263	100	124	348	100	56	304	206

The testing results that have been performed on 15 document pairs are the following:

Based on the similarity percentage column either similarity percentage of documents 1 and document 2 shows the percentage similarity average of 100%. Therefore it is proven that the algorithm can indicate documents 1 and documents 2 that were examined have similarity with the documents contained in the papers database. The algorithm is also proved that by using a number of different slave computers and work in parallel, the result is faster with three slave computers compared to two slave computers.

The second test performed on documents 1 and 2 which are modified document. Modifications made are to change the position of the sentence contained in the documents 1 and 2. The results of the second test document can be seen in Table 4.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 4 Second Test Result

No	Doc1	Similar Sentence	Similarity %	Doc 2	Similar Sentence	Similarity %	Number of Page	Total 2L (dec)	Total 3L (dec)
1	102	493	100	103	393	100	43	263	174
2	103	393	100	104	499	100	42	183	170
3	104	499	100	105	379	100	40	160	79
4	105	379	100	108	435	100	42	195	110
5	105	379	100	109	391	100	38	216	149
6	108	435	100	109	391	100	42	305	155
7	109	391	100	110	470	100	41	368	141
8	111	526	98	112	478	100	50	420	417
9	112	478	100	113	362	100	44	308	124
10	113	362	100	114	274	100	42	111	110
11	114	274	100	115	85	81	28	154	107
12	115	85	81	116	447	100	39	186	105
13	118	157	97	119	452	100	49	670	204
14	121	365	100	123	263	100	44	42	37
15	123	263	100	124	348	100	56	363	120

The result of the testing that has been performed on 15 document pairs is the following:

Based on the similarity percentage column either similarity percentage of document 1 and document 2 shows the percentage of the lowest similarity was 81% or the highest similarity percentage is 100%. If seen from the number of initial sentence compared to the number of sentences that are similar, the number of modified sentences ranging from 2 to 54 sentences. Changes in sentences ranging from 2 to 3 sentences, produces a similarity average percentage ranged between 97% - 100%. While the modification of the original sentence of 54 sentences, produces an average of 81% the similarity percentage. Therefore it is proven that the algorithm can indicate that documents 1 and document 2 that were examined, they have similarity to documents contained in the papers database. The algorithm is also proved that by using a number of different slave computers and work in parallel, the result is faster with three slave computers compared to two slave computers.

The third testing performed on documents 1 and 2 which are modified document. Modifications made are to change the position of the sentence and replace a few words with synonyms of the document to be compared. The third test results can be seen in Table 5.

Table 5 Third Testing Result

No	Doc 1	Similar Sentence	Similarity %	Doc 2	Similar Sentence	Similarity %	Number of Page	Total 2L (dec)	Total 3L (dec)
1	102	490	100	103	393	100	43	178	168
2	103	392	100	104	499	100	42	186	175
3	104	499	100	105	379	100	40	169	79
4	105	379	100	108	435	100	42	199	105
5	105	379	100	109	391	100	38	210	205
6	108	435	100	109	391	100	42	323	154
7	109	391	100	110	470	100	41	473	142
8	111	525	96	112	478	100	50	520	432
9	112	478	100	113	362	100	44	370	123
10	113	362	100	114	274	100	42	132	49
11	114	274	100	115	84	80	28	156	112
12	115	84	80	116	447	100	39	165	107
13	118	156	97	119	452	100	49	573	128
14	121	365	100	123	263	100	44	39	36
15	123	263	100	124	348	100	56	306	120

The results of the testing that has been performed on 15 documents pairs are the following:

Based on the similarity percentage column either similarity percentage of document 1 and document 2 shows the similarity percentage of the lowest was 80% and the highest similarity percentage is 100%. If seen from the number of initial sentence compared to the number of sentences that are similar, the number of modified sentences ranged from 1 to 55 sentences. Modification of the original sentence of 1 to 5 sentences produces an average similarity percentage ranged between 96% - 100%. Modification of the original sentence by 55 sentences produces a similarity percentage average of 80%. Therefore it is proven that the algorithm can indicate that documents 1 and document 2 that were examined have similarity to documents contained in the papers database. The algorithm is also proved that by using a number of different slave computers and work in parallel, the result is faster with three slave computers compared to two slave computers.

Fourth Tests performed on documents 1 and document.2 which are modified. Modifications made are to change the position of the sentence, replace a few words with synonyms, and changing the location of the contents of the sentence of the document to be checked. The fourth test results can be seen in Table 6.

Table 6 Fourth Testing Result

No	Doc 1	Similar Sentence	Similarity %	Doc 2	Similar Sentence	Similarity %	Number of Page	Total 2L (det)	Total 3L (det)
1	102	490	100	103	393	100	43	179	173
2	103	393	100	104	499	100	42	271	161
3	104	499	100	105	379	100	40	159	80
4	105	379	100	108	435	100	42	158	110
5	105	379	100	109	391	100	38	220	156
6	108	435	100	109	391	98	42	261	158
7	109	391	100	110	470	100	41	363	144
8	111	526	98	112	478	100	50	488	432
9	112	478	100	113	362	100	44	361	122
10	113	362	100	114	274	100	42	129	48
11	114	274	100	115	85	81	28	179	110
12	115	85	81	116	447	100	39	156	107
13	118	157	97	119	452	100	49	642	204
14	121	365	100	123	263	100	44	43	28
15	123	263	100	124	348	100	56	163	119

The results of the testing that has been performed on 15 document pairs are the following:

Based on the similarity percentage column either similarity percentage of document 1 and document 2 shows the similarity percentage of the lowest was 81% or the highest similarity percentage is 100%. If seen from the number of initial sentence compared to the number of sentences that are similar, the number of modified sentences ranging from 2 to 54 sentences. Modification of the original sentence of 2 to 5 sentences produces an average percentage of similarity ranged between 97% - 100%. Modification of the original sentence of 54 sentences produces an average of 81% the percentage of similarity. Therefore it is proven that

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

the algorithm can indicate that documents 1 and document 2 that were examined, they have similarity to documents contained in the papers database. The algorithm is also proved that by using a number of different slave computers and work in parallel, the result is faster with three slave computers compared to two slave computers.

From the test results explanation above, with modifications made to either change the position of the sentence, replace the word with a synonym, or change the location of a word in a sentence, the modified document remain detectable similar. The above results also show that the algorithm in this study can check the number of pages of documents similarity is more than one page. The maximum number of pages examined in this study is 56 pages. From the speed of the detection process shows that the speed measurement algorithm detection process is also said to be successful. This success can be seen from the speed of the process by using 3 computers slave faster than using 2 slaves computer.

5. CONCLUSION AND FUTURE WORK

To build an Indonesian language document similarity measurement system using synonyms factor with large data and measure the speed of detection processing of documents to go through the stages of sentence separation, analyze synonyms, compute the similarity of documents, and calculate the speed of document similarity detection process. Here are the conclusions and suggestions for further research.

5.1 Conclusion

Based on the research objectives, the experiments performed and the achievement of the objectives of the research can be said to be successful with the establishment of a system used to measure the similarity of documents in Indonesian language for large-sized documents. The test results of 15 pairs of documents are not modified, the success similarity percentage average of 100%. For testing of the modified document changes in the position of the sentence contained in the document 1 and 2 document with the changes between 2-3 sentences, yield the similarity percentage average between 97% - 100%, while the modification of the original sentence of 54 sentences, produces similarity percentage average of 81%. For testing of the modified document changes in the position of the sentence and replace a few words with synonyms of document 1 and document 2, with modifications to the original sentence of 1 to 5 sentences produce a similarity percentage average of of 96% - 100%, while the modification of the original sentence by 55 sentences produce a similarity percentage average of 80%. For testing of the modified document changes in the position of the sentence, replace a few words with synonyms, and changing the location of the contents of the document sentences 1 and 2 with a modified document original sentence of 2 to 5 sentences, produces a similarity percentage average of ranged between 97% - 100 %, while the modification of the original sentence of 54 sentences, produces a similarity percentage average of 81%.

From these results it can be concluded that by changing the position of the sentence, replacing words with synonyms, and changing the position of the word in a sentence, the examined sentence is considered to have the similarity percentage. In terms of the number of pages examined, the algorithm in this study indicates that the document can be checked more than one page. On testing also showed that the speed measurement algorithm detection process is also said to be successful. This success can be seen from the speed of the process by using 3 slave computers faster than using 2 slave computers.

5.2 Future Work

With this study it can be shown that the Indonesian document similarity measurement for large-sized documents can performed by using the architecture, algorithms and systems has been developed. Here is a challenge that is still open in this field, including:

1. Adding stemming process to Indonesian document similarity measurement for large documents.
2. Regarding parallel processing, adding with more than 3 slave computers or with a flexible number of slave computers to Indonesian document similarity measurement for large documents

REFERENCES:

[1] Alakeel, Ali M., 2010, *Load Balancing in Distributed Computer Systems*, International Journal of Computer Science and Security, Vol : 10, Issue 6, pages 153-160.

[2] Ana Kurniawati, 2010, *Algoritma Mengukur Kemiripan Dokumen Berbahasa Indonesia Dengan Faktor Sinonim*, Disertasi, Universitas Gunadarma.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[3] Demus Yabu Maundima, Suyono Balla Hamanai, dan Didimus Dundu Tay., 2011, *Perbedaan Paralel Processing dan Load Balancing*. http://www.scribd.com/doc/155977460/Perbedaan-Paralel-Processing, tanggal akses : 12 Juni 2014.

[4] Hari Bagus Firdaus, 2003, *Deteksi Plagiat Dokumen Menggunakan Algoritma Rabin-Karp*, Jurnal Ilmu Komputer Dan Teknologi Informasi, Vol III No.2.

[5] Hendra Rahmawan dan Yudi Satria Gondokaryono, 2009, *The Simulation of Static Load Balancing Algorithms*, International Conference on Electrical Engineering and Informatics, Malaysia.

[6] Rajguru, Abhijit A., 2012, *A Comparative Performance Analysis of Load Balancing Algorithms in Distributed System using Qualitative Parameters*, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-1, Issue-3.

[7] Sharma, Sandeep., Sarabjit Singh, and Meenakshi Sharma., 2008, *Performance Analysis of Load Balancing Algorithms*, World Academy of Science, Engineering and Technology.

[8] Sinta Agustina, 2008, *Aplikasi Anti Plagiarisme Dengan Algoritma Karp-Rabin, Skripsi, Universitas Gunadarma*, Jakarta.