Prosiding
ANNUAL RESEARCH SEMINAR 2016
6 Desember 2016, Vol 2 No. 1

ISBN : 979-587-626-0 | UNSRI http://ars.ilkom.unsri.ac.id

Implementation of Document Classification using Naïve Bayes Classifier for the Performance and Level of Accuracy of Document Searching using Boyer-Moore Algorithm

Muhammad Arief Algiffary[1], Muhammad Fachrurrozi[2], Novi Yusliani[3]
Faculty of Computer Science, Sriwijaya University
Jl. Srijaya Negara, Bukit Besar, Palembang, Indonesia
Email: ariefalgiffary@gmail.com[1], mfachrz@unsri.ac.id[2], novi_yusliani@unsri.ac.id[3]

***Abstrak*—Search engine is a program that searches a data in a database based on keywords entered by the user. There are several algorithms that have been used, one of them is Boyer-Moore. However, the use of the method found problems such as slow speed of search and low accuracy of search results when used on a large scale. In this study, is used classification of documents using the Naive Bayes Classifier to overcome these problems. Based on the results of research using 1000 documents, it was found that the speed of searching software with a classified documents is better than the speed of searching software without classified documents. However, the result of accuracy level have the same great value.**

***Keywords*—search engine, document's classification, naive bayes classifier, boyer-moore algorithm.**

I. INTRODUCTION

Search engine is a program that searches a data in a database based on keywords entered by the user (Oxford, 2012). Data sought in this context is a document, which can be texts, images, audios, videos, and other multimedia objects. One document search's method that has been used in the document searching is String Matching. In String Matching method, there are several algorithms that have been used, such as Brute Force, Knuth-Morris-Pratt and Boyer-Moore. From some of the String Matching algorithm, (Wibowo, 2011) performed a comparison against the performance of the algorithm of Knuth-Morris-Pratt and Boyer-Moore. This research produced that Boyer-Moore algorithm has better performance than Knuth-Morris-Pratt with a percentage of 32% faster.

However, the use of the document searching's method can only be used to search on a small scale. The using of the methods such as string matching, pattern matching, sequential search and binary search to search in a massive scale encountered several obstacles, namely slow speed of search and low accuracy of search's results (Juniawan, 2009). These constraints resulted in a statement for better performance and results can apply the use of documents classified by the classification method (Februariyanti and Aliarso, 2012).

Classification of documents is a process of categorization performed on a set of documents and assist the process of finding a certain document quickly and accurately (Distiawan, 2009). Document classification determines the group of documents in accordance with the existing categories. If the document you want to find are known to have a particular category, the documents are searched only in the category.

(Liliana, Hardianto, and Ridok, 2011) performs classification document for Indonesian news language using Support Vector Machine. This classification method achieve an accuracy rate of 85%. (Asy'arie and Pribadi, 2009) also made the classification of news document in Indonesian language by using Naïve Bayes classifier. The research proves that Naïve Bayes Classifier shown to be effectively used to classify news automatically with an accuracy up to 90%.

Based on these researches, this research will compare a document searching software using String Matching - Boyer-Moore algorithm with the document classified by Naïve Bayes Classifier method against a document searching software without the document classified.

II. RElATED WORKS

Research related to the classification in Indonesian language using a Naive Bayes Classifier algorithm has been done, one of them by (Asy'arie and Pribadi, 2009) with his research Automatic Classification News Articles in Indonesian Language by Using Naive Bayes Classifier Method. In their research, (Asy'arie and Pribadi, 2009) discusses the classification of news document in Indonesian language into pre-defined categories based on keywords in the document. Classification methods used in this study is the Naive Bayes Classifier. The stages are carried out in the classification process in the form casefolding, parsing, stopwords elimination, stemming, words weighting, and the final document classification using Naive Bayes Classifier. This study uses 250 documents are divided into five categories of documents. The test methods used are looking for value of Recall and value of Precission. The results obtained from this research is a level of accuracy of 92.87% in the Recall and 91.16% in Precision.

Another research conducted by (Wibisono, 2005) is Indonesian News Classification Using Naïve Bayes Classifier. Text documents used for the experiment was taken from Kompas newspaper pages online with the number of 582 text documents. The text document is divided into five categories, namely the metro, health, sports, technology, and lifestyle. A text document is divided into two parts, namely learning documents and document testing. The results of this research has an accuracy of 89.47%. Value of accuracy remains high, especially if the documents are learning to use a large (greater than or equal to 400). The conclusion of this research is the Naïve Bayes Classifier proven method can be used effectively to automatically classify news.

Research on the document search, (Boyer and Moore, 1977) in A Fast Algorithm String Searching finding algorithm which is currently known as the Boyer-Moore Algorithm. This algorithm runs with matching characters from the right pattern, unlike other algorithms that starts from the left. This is because a lot of information that will be obtained in this way. If the rightmost character pattern not found a match with the characters in the string, it can be seen not finding a pattern and a shift directly to do as much as the length of the pattern. In this study used two tables, tables delta1 and Delta2. By definition, the same value as long delta1 pattern if the characters in the string is not found in the pattern, while Delta2 is the distance that can be shifted so that the characters in the string mismatch but contained in the pattern may be misaligned. In this way, the algorithm has great character leap if found unsuitable character thus reducing the amount of the matching process. This speeds up the search process and also make the process more efficient.

Research (Boyer and Moore, 1977) reinforced by research (Wibowo, 2011) comparing algorithm Knuth-Morris-Pratt with Boyer-Moore in Comparative Knuth-Morris-Pratt algorithm and algorithm Boyer-Moore in the Search Text in Indonesian and English. In this study, conducted experiments on a few words or sentences in Indonesian and English of an article. The result of this comparison is that the algorithm Boyer-Moore has a search speed is better 41% in the search text in English and preferably 23% in search of Indonesian-language text. So it was concluded that the Boyer-Moore algorithm has a search speed 32% better than the algorithm Knuth-Morris-Pratt.

III. PROPOSED METHOD

In this research, there are two softwares we developed, the document classification software and the document searching software. The document classification software is used to afford the classified documents, while the documents searching software is used to obtain a scientifical value on speed of search and level of accuracy. We used the final project documents of Informatics students as the data set.

2.8. *Preprocessing*

The purpose of preprocessing is to transform document input into document that can be read and processed by the system. Preprocessing is used in both document classification software and document searching software. The steps are:

- Casefolding is a process to convert all alphabetical letters in documents into small letters. The process also eliminates the characters besides alphabetical letters.
- Tokenizing is a process to break down all of the sentences in document into words.
- Stopwords Removal is a process to filter words that are not useful and frequently appears, and eliminate it. The purpose of the process is to eliminate words that have low value in information retrieval.
- Stemming is a process to cut affixes in a word and convert it into a rootword. The kind of affixes are prefixes, suffixes, infixes, and confixes (the combination of prefixes and suffixes).

- Words Weighting is a process to deliver value or weight to a word based on its appearance in a text document or category. The process is only used in the documents classification software in the classification phase.

$$\text{TF-IDF} = \text{TF} . \log (\frac{n}{DF})$$

2.9. *Naïve Bayes Classifier*

We propose a method called Naïve Bayes Classifier to classify document based on the classification made. The document is classified into 5 categories such as Image Processing, Natural Language Processing, Expert Systems, Information Systems, and Decision Support Systems based on the contents of the document.

There are two phases in the document classification software, learning phase and classification phase.

1. Learning phase is used to obtain an information that representing a category, represent by a probability value. Data used in learning phase as much as 750 documents

$$Probability\ value = \frac{Word's\ frequency}{Total\ document(s)}$$

2. Classification phase is the phase to classify the document. During classification, the Naïve Bayes Classifier's method is used after obtaining a probability value from data traning. Data used in classification phase as much as 250 documents.

$$P(Class|Doc) = \frac{P(Class)}{P(Doc)} P(Doc|Class)$$

After all of the documents's classified, the document's searching software perform.

2.10. *Boyer-Moore Algorithm*

At this stage, we do a document searching by matching the keyword entered by the user to the words in the document. The algorithm used in the searching process is Boyer-Moore algorithm. If an example keyword is "bayes", the first process is calculate the shift table on each character in the word "bayes". There are two tables that are used in making the shift. These are bad-character table and good-suffix table.

Bad-Character Table

Create a table with a large shift equal to the length of pattern "bayes" which in this case is worth 5. Then the contents of the table with a long of pattern.

Table 1. Shifting with Initiation Value Long Pattern "bayes"

a	b	a	y	e	s
i	0	1	2	3	4
Bmbc[a]	5	5	5	5	5

Perform as many repetitions m - 1 starting from the leftmost character. Then calculate the value of bmBc [a] with m - i - 1. For the first step, the character and the value of i = 0, the value of bmBc [b] is 5-0 - 1 = 4. So the contents of all bmBc [b] with a value of 4 ,

Table 2. Table Shifts Rated i = 0

a	b	a	y	e	s
i	0	1	2	3	4
Bmbc[a]	4	5	5	5	5

Add the value of i by 1 so that i = 1. Repeat steps 2 to get the value bmBc [a] is a 5 - 1 - 1 = 3.

Table 3. Table Shifts Rated i = 1

a	b	a	y	e	s
i	0	1	2	3	4
Bmbc[a]	4	3	5	5	5

Add the value of i by 1 so i = 2. Repeat step 2 to get the value bmBc [y] is 5 - 2 - 1 = 2.

Table 4. Table Shifts Rated i = 2

a	b	a	y	e	s
i	0	1	2	3	4
Bmbc[a]	4	3	2	5	5

Add the value of i by 1 so i = 3. Repeat step 2 to get the value bmBc [e] is 5 - 3 - 1 = 1.

Table 5. Table Shifts Rated i = 3

a	b	a	y	e	s
i	0	1	2	3	4
Bmbc[a]	4	3	2	1	5

Add the value of i by 1 so that i = 4. Repeat steps 2 to get the value bmBc [s] is a 5 - 4 - 1 = 0.

Table 6. Table Shifts Rated i = 4

a	b	a	y	e	s
i	0	1	2	3	4

Bmbc[a]	4	3	2	1	0

For all the characters other than the characters on "bayes", the value of a shift table is 5.

Good-Suffix Table

The shift based on the good-suffix table can not be calculated because there are none suffix in the pattern so that the shift value table based on good-suffix is considered 0.

After obtained the value of the shift, then do pattern matching "bayes" in the document. Matching process will continue until the last word in the document. Step - matching measures can be seen as follows:

Make a matching of the rightmost character pattern. If there is a mismatch, do shift corresponding character values in the table shift. In the table below, the character "s" and "d" is not suitable, then the shifting is done by shifting the value of "d". In the table of bad-character shift value "d" of 5, while the good-suffix table worth 0. So taken the biggest value between the two tables is 5.

Table 7. Table Match Character “s” and “d”

m	e	t	o	d	e		n	a	i	v	e		b	a	y	e	s
b	a	y	e	s													

After shifting as much as five times, the last character in the string and pattern matched back. In the table below, the character "s" and "i" does not match, then the shifting is done by shifting the value of "i". In the table of bad-character shift value "i" of 5, while the good-suffix table worth 0. So taken the largest value between the two tables is 5.

Table 8. Table Match character "s" and "i"

m	e	t	o	d	e		n	a	i	v	e		b	a	y	e	s
					b	a	y	e	s								

After shifting as much as five times, the last character in the string and pattern matched back. In the table below, the character "s" and "a" does not match, then the shifting is done by shifting the value of "a". In the table of bad-character shift value "a" of 3 while the good-suffix table worth 0. So taken the largest value between the two tables is 3.

Table 9. Table Match character "s" and "a"

m	e	t	o	d	e		n	a	i	v	e		b	a	y	e	s
										b	a	y	e	s			

After shifting as much as three times, the last character in the string and pattern matched back. Once matched character by character, it turns out the word "bayes" found.

Table 10. Table Match Character appropriate

m	e	t	o	d	e		n	a	i	v	e		b	a	y	e	s
													b	a	y	e	s

IV. Experimental Result

The experiments are conducted using a computing platform based on Processor Intel Core i5 CPU @2.30 GHz, Random Access Memory 4 GB 1333 MHz DDR3, 250 GB Solid State Drive. The development environment is Operation System MAC OS X El Capitan and Java programming language with Netbeans 8.0.2 as a framework.

Accuracy testing of document classification software is perform using data test as much as 250 final project documents that are divided into 5 categories such as Image Processing, Natural Language Processing, Expert Systems, Information Systems, and Decision Support Systems.

Table 11. Document Classification’s Accuracy Test Results

No	Category	Data Train	Data Test	Accuracy	Percentage
1	Image Processing	150	50	48	96%
2	Natural Language Processing	150	50	48	96%
3	Expert Systems	150	50	41	82%
4	Information Systems	150	50	46	92%
5	Decision Support Systems	150	50	44	88%

Percentage of the accuracy of test results can be calculated by:

$$\text{Accuracy} = \frac{Document-right-classified}{Total\ Data\ Test}\ x\ 100\%$$

$$= \frac{227}{250}\ x\ 100\% = 90{,}8\ \%$$

Accuracy testing of document searching software is done using 8 keywords that consist of 3 keywords of one-word, 3 keywords of two-words and 2 keywords of three-words. Before do a search on the document, keyword combinations will be searched first. Testing is done by comparing the number of keywords found by the system and the number of keywords that should exist in the document.

Table 12. Accuracy's Level in Classified Document

No	Keyword	Combination	Classified Doc	Testing (Error)	Result
1	1 word	1 keyword	20	20(0)	100%
2	2 word	3 keyword	20	60(1)	99.166%
3	3 word	7 keyword	14	98(3)	99.235%
Rate					99.467%

Table 13. Accuracy's Level in Unclassified Document

No	Keyword	Combination	Unclassified Doc	Testing (Error)	Result
1	1 word	1 keyword	20	20(0)	100%
2	2 word	3 keyword	20	60(1)	99.166%
3	3 word	7 keyword	14	98(3)	99.235%
Rate					99.467%

For the speed of search test, keywords used are the same keywords in the testing accuracy. On each keywords, calculated the time required methods in the search for eywords on the entire document. Results obtained in the form of milisecond.

Table 14. Result of Speed of Search (in milisecond)

No	Comb	Keyword	Time (Classified Doc)	Time (Unclassified Doc)
1	1 Word	enkripsi	35,8	93,6
2		klasterisasi	29,8	90,8
3		biometrik	28,2	91,6
4	2 Word	brute force	45,3	101,3
5		bayesian probability	46,8	95,3
6		sms gateway	44,7	94,8
7	3 Word	latent semantic indexing	70,4	107,9
8		maximum marginal relevance	74,0	123,5

II. CONCLUSION

The conclusions obtained from this study can be summarized as follows:

1. Use of the classification of documents make the search speed of classified documents better than the search speed of documents that are not classified
2. The accuracy's level of the result of two kinds of searching document had the same value of results. Not affected by the classification of documents.
3. The accuracy's level of the document classification software affect documents that will be generated by the searching classified documents software. The better document classification software to classify documents, the better documents produced by the searching classified documents software.
4. The result of the class determining by the searching classified documents was very influential by the classification program. The more precise probability value attached to the classification program, the better the searching classified documents determine the category of keywords entered.

REFERENSI

[1] Asy'arie, A. D., & Pribadi, A. W. 2009. Automatic News Articles Classification in Indonesian Language by Using Naive Bayes Classifier Method. Proceedings of the 11th International Conference on Information Integration and Web-based Apps & Services ACM.

[2] Boyer, R. S. and J. S. Moore. 1977. A Fast String Searching Algorithm. Magazine Communications of the ACM 20 (10) : 762-772.

[3] Februariyanti, H., & Zuliarso, E. 2012. Klasifikasi Dokumen Berita Teks Bahasa Indonesia menggunakan Ontologi. Jurnal Teknologi Informasi DINAMIK Volume 17, No.1, Januari 2012. Fakultas Teknologi Informasi, Universitas Stikubank. Semarang.

[4] Juniawan, Indra. 2009. Klasifikasi Dokumen Teks Berbahasa Indonesia Menggunakan Minor Component Analysis. Departemen Ilmu Komputer, FMIPA, Institut Pertanian Bogor. Bogor.

[5] Liliana, D. Y., Hardianto, A., & Ridok, M. 2011. Indonesian News Classification using Support Vector Machine. World Academy of Science, Engineering and Technology, 57, 767-770.

[6] Wibisono,Y. 2005. Klasifikasi Berita Menggunakan Naive Bayes Classifier. Jurusan Pendidikan Matematika, FPMIPA, Universitas Pendidikan Indonesia. Bandung.

[7] Wibowo, Kevin. 2011. Perbandingan Algoritma Knuth-Morris-Pratt dan Algoritma Boyer-Moore dalam Pencarian Teks di Bahasa Indonesia dan Inggris. STEI, Institut Teknologi Bandung. Bandung.