Analysis And Comparison Stemming Indonesian Text Using Vega Algorithm And Confix-stripping Algorithm

Helmi Agustian[1], Tati Harihayati[2]

Teknik Informatika – Universitas Komputer Indonesia
Jl. Dipatiukur 112-114 Bandung

Email : helmi_mail@yahoo.com[1], tharihayati@yahoo.com[2]

ABSTRAK

Stemming algorithm is an algorithm that can transform word to its root word. In the document searching, stemming expected to increase recall and index compression. Calculating the stemmer strength can predict how well a stemming algorithm in the abillity to recall and index compression[2]. Based on the results to calculate six measures of stemmer strength which is mean number of word per conflation class, index compression factor, the number of words and stems that differ, the mean number of characters removed in forming stems, and the last median and mean modified hamming distance between words and their stems. Vega algorithm tend to often transform word rather than confix-stripping algorithm. The results also show that confix-stripping algorithm tend to often transform words that correspond to the same stemm for a corpus.

Kata Kunci : *Stemming, Vega, Confix-stripping*

1. INTRODUCTION

Stemming is a process of finding the root word of a term by eliminating affix, prefix, confix, suffix and combination of both prefix suffix on the word derivative. Stemming is used to transform word to its root word that accordance with Indonesian morphological structure. Stemming algorithms are used in many types of language processing and text analysis, and can make process of text retrival efficiently and effectively

Indonesian affixes is more complex when compared with the English affixes. Because affixes in Indonesian language consists of a prefix, confix, suffix, and combination of both prefix-suffix. The affixes that attached to a word must be removed to change to its root word

Stemming process can also becalled conflation, which are used in search engines for indexing and query expression and also for natural language processing problems. Conflation process can be done manually or automatically. The program that can automatic conflate word is called 'stemmer'. Stemmer used in information retrieval system to reduce the size of index file. The term can be stemmed at the time indexing and saarching happened. The advantage of stemming at the process of indexing is compression and efficiency of index file. Which mean that stemming can improve the performance of information retrieval system.

There are several type of stemming process, including : Table look-up, affix removal stemmers, successor variety dan n-gram stemmer. In this research about analysis and comparison of vega algorithm and confix-stripping algorithm, both of them are affix removal stemmers type.

The tests in this study are using metrics that describe in the journal by Frakes W.B. and Fox C.J. [2]. These metrics are used to measure the stemmer strength, which is important in the case of a particle removal algorithms, because it can predict the ability of recall and precision as well as index compression of the stemmer algorithm. Stemmer strength is a measure of how well stemmer algorithm can change word into a stem. The parameters that used to measure stemmer strength are, mean number of words per conflation class, index compression factor, the number of words and stems that differ, the mean number of characters removed in forming stems, and the last median and mean modified hamming distance between words and their stems

1.1. Vega Algorithm

Vega stemming algorithm is an algorithm that was introduced by Vinsensius Berlian Vega in 2001. Vega algorithm does not use the dictionary to check the root words.

Consecutive steps of vega algorithm as follow, the removal of particles, removal of the possessive pronouns, circumfix disposal, removal of the prefix and finally suffix disposal.

1.2. Confix-stripping Algorithm

Confix-stripping algorithm (CS) is a stemming algorithm for Indonesian Language introduced by Jelita Asian. Confix-stripping algorithm is algorithm that developed from stemming algorithm by Nazief and Adriani [1]. This algorithm uses the dictionary to check the root words. Confix-stripping algorithm stages can be described as Figure 1.

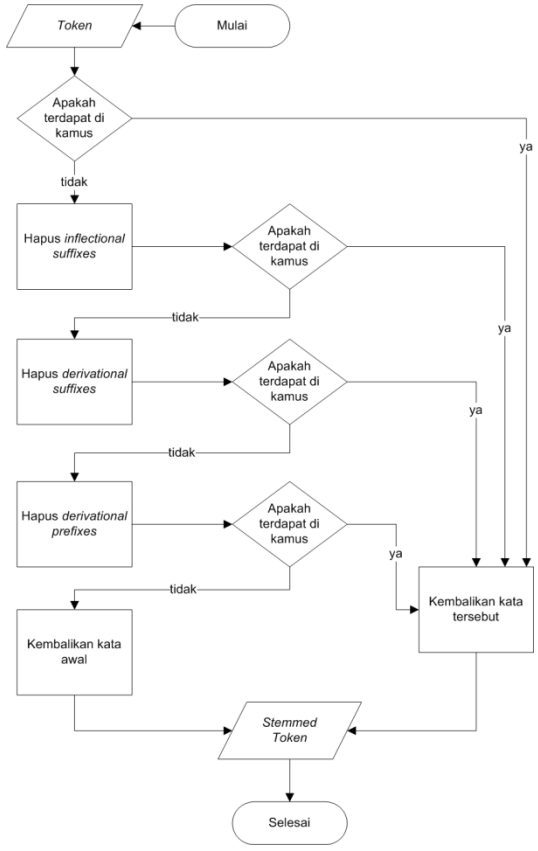


Figure 1 Confix-stripping algorithm steps

1.3. Metrics measure

Tests measure in this study using the parameters described in the journal by Frakes W.B. and Fox C..J. [2]. The first parameter is mean word per conflation class, a measure of the average number of words that correspond to the same stem for a corpus, with the following formula:

$$wcc = \frac{n}{s}$$

where :

wcc = word per conflation class

n = total of unique corpus

s = total of uniqe stems

Stronger stemming algorithm tend to have word per conflation class value greater, to show that algorithm better

The second metrics is the index compression factor, shows how much the level of index compression an algorithms stemming can get, or fractional reduction in index size achieved through stemming. If the number of words in the corpus is n, then after stemming process resulting numbers of stem and removing duplicate word that is s. So to calculate the index compression factor (icf) are:

$$icf = \frac{(n-s)}{n} x\ 100\%$$

where :

icf = index compression factor

n = total of unique corpus

s = total of uniqe stems

Stronger stemming algorithm tend to have index compression factor value greater, to show that algorithm is better.

The third parameter is the number of different words between term and stem. A stemming algorithm often leave words unchanged. Stronger stemmers will change words more often than weaker stemmers.

The fourth parameter is the mean number of character removed in forming stems. Stronger stemming algorithm remove more character from word to form stems

The fifth and sixth metrics respectively are the mean and median modified hamming distance between word and their stems. The hamming distance between two strings of equal length is defined as the number of character in the two strings that are different at the same position. For strings of unequal length add the difference in length to the hamming distance function.

For example, there are three words “membaca”, “dibaca” and “membackan”. If a stemming algorithm can convert all three to stem “baca”, then:

1. membaca → have 3 difference character with ‘baca’
2. dibaca→ have 2 difference character with ‘baca’
3. membacakan → have 6 difference character with ‘baca’

So the mean modified hamming distance can be calculated as the average (mean) value as follows:

(3+2+6) / 3 = 3,66

We get the value of mean modified hamming distance which is 3,66 character

The value of median modified hamming distance can also be obtained as the median of these values after sorted. So the median modified hamming distance obtained as 3.

2. RESEARCH CONTENT

Result of testing vega algorithm and confix-stripping algorithm obtained from calculated the six metrics measure. The sample documents taken from sarihusada digital library at the total of 10 sample documents. Every document have through preprocessing that resulting corpus that ready to calculate the metrics, as follows

2.1. Mean number of words per conflation class

Table 1 Mean number of words per conflation class

Doc No	n	s		wcc	
		Vega	CS	Vega	CS
1	417	407	400	1,02 character	1,04 character
2	78	77	77	1.013 character	1.013 character
3	86	85	85	1.012 character	1.012 character
4	184	182	181	1.01 character	1.016 character
5	80	79	78	1.01 character	1.016 character
6	462	451	450	1.024 character	1.027 character
7	58	57	57	1.017 character	1.017 character
8	85	84	83	1.012 character	1.024 character
9	115	114	114	1.0087 character	1.008 character
10	128	126	124	1,015 character	1,032 character
Total	1693	1662	1649	1,018 character	1,026 character

Stemming algorithm that good in terms of stemmer strength want a mean word per conflation class larger than the other stemming algorithm. Based on the Table 1 the mean word per conflation class of confix-stripping algorithm is better with an average of 1,026 words.

2.2. Index Compression Factor

Table 2 Index compression factor

Doc No	n	s		icf	
		Vega	CS	Vega	CS
1	417	407	400	2.398 %	4.076 %
2	78	77	77	1.282 %	1.28 %
3	86	85	85	1.162 %	1.162 %
4	184	182	181	1.086 %	1.63 %
5	80	79	78	1.25 %	2.5 %
6	462	451	450	2.381 %	2.597%
7	58	57	57	1.724 %	1.724 %
8	85	84	83	1.176 %	2.353 %
9	115	114	114	0.869 %	0.869 %
10	128	126	124	1,562 %	3,125 %
Total	1693	1662	1649	1,831 %	2,598 %

Stemming algorithm which is good in terms of stemmer strength want index compression factor value greater than the other stemming algorithms. Based on testing at Table 2 index compression factor parameter confix-stripping algorithm is better at 2,598%

2.3. The number of words and stems that differ

Table 3 The number of words and stems that differ

Doc No	Vega	CS
1	409 character	365 character
2	76 character	65 character
3	83 character	72 character
4	179 character	72 character
5	74 character	61 character
6	448 character	413 character
7	54 character	49 character
8	82 character	65 character
9	108 character	88 character
10	123character	101 character
Total	1636 character	1351 character

Stemming algorithm which is good in terms of stemmer strength is more often change the words than the other stemming algorithms, resulting in different words. Based on testing at Table 3 vega algorithm is better with total difference 1636 words from 10 sample documents.

2.4. The mean number of characters removed in forming stems

Table 4 The mean number of characters removed in forming stems

Dok No	m		n	Vega	CS
	Vega	CS			
1	1662	1450	417	3.985 word	3.477 word
2	305	265	78	3.91 word	3.397 word
3	317	289	86	3.686 word	3.36 word
4	699	604	184	3.798 word	3.282 word
5	279	227	80	3.487 word	2.837 word
6	1763	1567	462	3.816 word	3.391 word
7	206	172	58	3.551 word	2.965 word
8	295	225	85	3.4705 word	2.647 word
9	408	319	115	3.547 word	2.773word
10	486	360	128	3,796 word	2,812 word
Total	6420	5478	1693	3,792 word	3,235 word

Stemming algorithm which is good in terms of stemmer strength remove more often the character to form a stem. Based on testing at Table 4 vega algorithm is better because the average erase 3,792 characters from 10 test documents.

2.5. Mean Modified Hamming Distance

Table 5 Mean modified hamming distance

No	h		n	Vega	CS
	Vega	Cs			
1	1662	1528	417	3.985 character	3.664 character
2	305	280	78	3.91 character	3.589 character
3	317	304	86	3.686 character	3.534 character
4	699	646	184	3.798 character	3.51 character
5	279	239	80	3.487 character	2.987 character
6	1763	1641	462	3.816 character	3.552 character
7	206	180	58	3.551 character	3.103 character
8	295	240	85	3.47 character	2.823 character
9	408	333	115	3.547 character	2.895 character
10	486	382	128	3,796 character	2,984 character
Total	6420	5773	1693	3,792 character	3,409 character

Stemming algorithm which is good in terms of stemmer strength have hamming distance value greater than the other stemming algorithms, or have mean modified hamming distance value greater for numbers of corpus in the test document. Based on testing at Table 5 vega algorithm is better because it has an average hamming distance as much as 3,792 words.

2.6. Median Modified Hamming Distance

Table 6 Median modified hamming distance

Dok No	Vega	CS
1	4	4
2	4	3
3	3	3
4	4	3
5	3	3
6	4	3
7	3	3
8	3	3
9	3	3
10	4	3
Average total	3,5	3,1

Stemming algorithm which is good in terms of stemmer strength have hamming distance value greater than the other stemming algorithms, that mean also for numbers of curpus in the document have middle value (median) greater. Based on testing Table 6 from 10 documents vega algorithm have better average median value of 3.5.

3. CONCLUDING

3.1. Conclusion

Based on the results of testing that has been done in the research content about analysis and comparison vega algorithms and confix-stripping algorithms, it can be concluded that all six test parameters did not show consistency in deciding which one is better algorithms. The test results from 10 documents show that vega algorithm better on four test parameters, which is the number of words and stems that differ, mean number of characters removed in forming stems, the mean modified hamming distance, and the median modified hamming distance. Confix-stripping algorithm better on two other test parameters, which is the mean number of words per conflation class and index compression factor.

The paragraph above means that the two algorithms each has their own advantages for specific cases. Vega algorithm has advantages in changing the term into different words or new, while confix-stripping algorithm has advantages in reducing the number of

terms, such as reducing the size index in the indexing process.

3.2. Suggestions

Berdasarkan hasil pengujian yang telah dilakukan, maka saran yang diharapkan setelah dilakukannya analisis dan perbandingan algoritma vega dan algoritma confix-stripping pada teks bahasa indonesia adalah sebagai berikut :

1. Melakukan pengujian terhadap stem yang dihasilkan algoritma vega apakah dapat mempengaruhi proses indexing.

2. Melakukan analisis performansi algoritma confix-stripping dalam mereduksi ukuran index size.

3. Melakukan implementasi terhadap proses pencarian pada information retrieval.

Based on the results of testing that has been done, then the suggestions are expected after doing an analysis and comparison vega algorithms and confix-stripping algorithms the Indonesian text is as follows:

1. Perform testing to stems result from vega algorithm whether it can affect the indexing process.

2. Analyze the performance of confix-stripping algorithm in reducing the size of the index size.

3. Conduct an implementation for searching process of an information retrieval.

4. BIBLIOGRAPHY

[1] Asian, J., Williams, H.E. and Tahaghoghi, S.M.M. “Stemming Indonesian”. RMIT University, Melbourne, Australia.

[2] Frakes, W.B., Fox, C.J. “Strength and Similarity of Affix Removal Stemming Algorithms”. Computer Science Department. Virginia Tech and James Madison University.

[3] Vega, B. V., dan Bressan, S. 2001. “Indexing the Indonesian web: Language Identification and Miscellaneous Issues”. Tenth International World Wide Web Conference, Hong Kong.