JURNAL UPI

Journal of Computers for Society

Journal homepage: https://ejournal.upi.edu/index.php/JCS

Journal of Computers for Society

Stemming Algorithm Modification for Overstemming Cases

*Stephanie Betha R.H**

Teknik Komputer, Universitas Wiralodra, Indonesia
*Correspondence: E-mail: ntephbetha@gmail.com

ABSTRACT

The stemming process plays an important role in the preprocessing of the text. One of the problems that occur in the stemming process is overstemming. Overstemming is an exaggerated word cut causing situations where a word has a very different meaning, but it produces the same stem. Therefore, to overcome these problems, it will be modified on the stemming process. This modification is done by combining two stemming algorithms (hybrid stemming) that is the look-up algorithm of dictionary table and affix removal algorithm using stemming porter. The modification of this stemming algorithm will be tested on title in scientific publication documents. The test results show that stemming process with modification of stemming algorithm can increase the recall value in the title attribute, although not very significant. The recall in an experiment using title attribute is 89,9%.

ARTICLE INFO

Article History:
Submitted/Received 02 Mar 2023
First Revised 09 Apr 2023
Accepted 12 Jun 2023
First Available Online 13 Jun 2023
Publication Date 15 Sep 2023

Keyword:
Stemming modification, Stemming.

1. INTRODUCTION

Joshi et al. (2016) stated that the stemming process functions to return words to their basic form, increase recall and provide relevant results in the word search process (Singh & Gupta, 2017; Bakar et al., 2000; Attar & Fraenkel, 1977; Krallinger *et al*., 2017). Therefore, the stemming process plays an important role in text preprocessing.

One of the problems that occurs in the stemming process is overstemming. Karaa (2013) states that overstemming is excessive cutting of words causing a situation where a word has a very different meaning but produces the same stem. For example, the words "generalized, general, generous, generating" produce the stem "gener". Previous research (Karaa, 2013) has handled overstemming cases by modifying the Porter stemming algorithm. Modification of the Porter stemming algorithm is carried out by creating several word rule tables (Jivani, 2011; Moral *et al*., 2014; Paice, 1996; Abi-Salem *et al*., 1999).

This research also aims to overcome the problem of overstemming cases by modifying the stemming process. This modification was carried out by combining two stemming algorithms (hybrid stemming), namely the word dictionary look-up table algorithm and the affix removal algorithm using Porter stemming (Patel & Shah, 2016; Panograhi & Bele, 2016). This word dictionary look-up table algorithm will adopt a rule table that has been proposed by previous research (Karaa, 2013).

This research uses the title attribute in published documents. The title attribute will undergo a stemming process with modification and a stemming process without modification. These two processes will produce a comparison of the recall values in the title.

2. METHODS

The stemming process is one part of text preprocessing. The stemming process is carried out after the case folding process, tokenization and removal of stop words. Several stages that must be passed in text preprocessing include:

2.1. Case Folding

The case folding process is done by changing uppercase letters to lowercase. Only letters 'a' to 'z' are accepted. Characters other than letters are considered delimiters so they will be removed from the document, for example, punctuation marks (&,#, : ,-, ;) and numbers (0-9).

2.2. Tokenization

The tokenization process is the process of breaking sentences into words (tokens). With this process, the input string will look more concise because it is displayed in the form of each word.

2.3. Removal of Stop Words

Stop words are removed after going through the tokenization process. This process aims to delete words that appear frequently and are not used in processing. Removal of stop words generally follows existing standard lists of stop words, especially in English. Standard stop words in general text, including: a, about, a, able, about, across, after, all, almost, also and so on.

DOI: http://dx.doi.org/10.17509/xxxx.xxx
p- ISSN xxxx-xxxx e- ISSN 2723-4088

2.4. Stemming

Joshi et al. (2016) stated that this process functions to return words to their basic form, increase recall and provide relevant results in the word search process. Recall is the ratio between relevant documents that have been successfully recovered from all relevant documents in the system. Stemming aims to reduce the inflectional forms and derivative forms of a word to a common basic form. For example, the words "cars, car's, car" have the same basic form, namely "car". The main problem in the stemming algorithm is how to get the correct base word from a created word. One of the problems that occurs in the stemming process is overstemming.

Overstemming cases are handled by combining two stemming algorithms (hybrid stemming), namely the word dictionary look-up table algorithm and the affix removal algorithm using Porter stemming. Look up table is basically storing root words in a table in the database along with the derived forms of these words. Words are entered into the table manually. A list of words that have been entered into the database can reduce errors in overstemming. When the user enters an inflected word, the stemmer will search for the presence of the inflected word in the database. Inflection is a change in the form of a word that does not change the meaning of the word (Karaa, 2013).

The look up table of the word dictionary in this study was carried out by making a list of words that were included in the rules table. Rule tables (Karaa, 2013) are tables that contain rules for cutting words in the stemming process. This word list was created based on a word dictionary obtained from the morewords website. This website is used as a word dictionary of choice because it has words sourced from the Enahanced North American Benchmark Lexicon (ENABLE2K) which has more than 173,528 words. The list of words used, and their derived words are entered into a table in the database manually.

Figure 1 explains the flow of handling overstemming cases that are not handled by the Porter stemming algorithm. If a word is included in the overstemming case in the rules table, then the word will produce a stem according to the word dictionary table. However, if a word is not included in the rule table list, then the word will receive stemming treatment using the existing Porter algorithm.

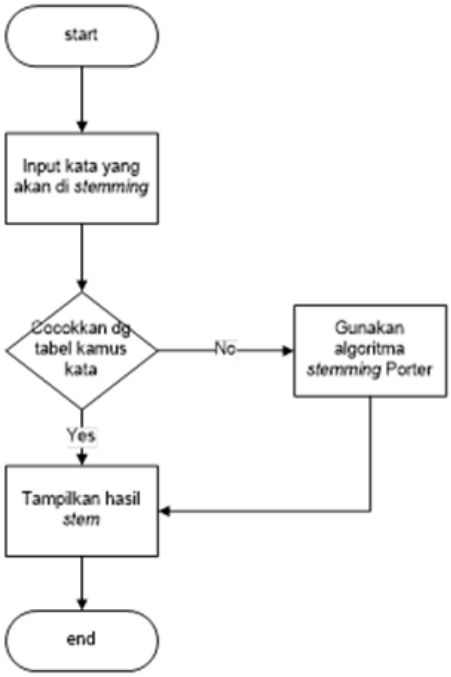

Figure 1. Flow of handling overstemming cases.

DOI: http://dx.doi.org/10.17509/xxxx.xxx
p- ISSN xxxx-xxxx e- ISSN 2774-7891

Overstemming cases will be handled using several rule tables proposed by Karaa (2013) as follows:

(i) First Rule. The first rule handles the case of plural forms in irregular words which are not handled in the Porter stemming algorithm. **Figure 2** shows a list of case categories handled in the first rule:

No	Kategori	*Original Words*	Aturan	Hasil *Stem* Modifikasi	Total Kata
1	Kata yang berakhiran children adalah bentuk jamak dari child	*child/children*	Mengubah children menjadi child	*child/child*	6 kata
2	Kata yang berakhiran -men adalah bentuk jamak dari -man	*dryman/drymen*	mengubah –men menjadi -man	*dryman/dryman*	429 kata
3	Kata yang berakhiran -ci adalah bentuk jamak dari -cus	*abacus/abaci*	mengubah –ci menjadi -cus	*abacu/abacu*	35 kata
4	Kata yang berakhiran –eaux adalah bentuk jamak dari -eau	*plateau/plateaux*	mengubah –eaux menjadi -eau	*plateau/ plateau*	29 kata
5	Kata yang berakhiran –mata adalah bentuk jamak dari -ma	*automa/automata*	mengubah –mata menjadi -ma	*automa/automa*	108 kata
6	Kata yang berakhiran –trices adalah bentuk jamak dari -trix	*matrix/matrices*	mengubah -trices menjadi -trix	*matrix/ matrix*	21 kata
7	Kata yang berakhiran -ses adalah bentuk jamak dari -sis	*analysis/analyses*	mengubah -sis menjadi -s	*analys/analys*	492 kata

Figure 2. List of Case Categories handled in the First Rule.

(ii) Second Rule. The second rule handles words that have the stem "gener" and words that have the prefix "gene". Several words in **Table 1** before undergoing modifications to the stemming algorithm, will experience cases of overstemming. All these words produce the same stem, namely "gener". The following is a list of words that have the stem

DOI: http://dx.doi.org/10.17509/xxxx.xxx
p- ISSN xxxx-xxxx e- ISSN 2723-4088

"gener" with changes in the stem results. **Table 1** shows the cases handled in the second rule.

Table 1. List of Case Categories handled in the Second Rule.

Words	Result of stem before modification	Result of modification stem	Total of words
Generate	Gener	General	76
Generates	Gener		
Generated	Gener		
Generating	Gener		
General	Gener	General	
Generally	Gener		
Generic	Gener	Generic	
Generically	Gener		
Generous	Gener	Generous	
Generously	Gener		

(iii) Rule of Three. The third rule handles words that have the ending –y and do not contain a vowel in the word which is not handled by the Porter stemming algorithm. **Table 2** shows the cases handled in the third rule.

Table 2. List of case categories handled in rule three.

Original Words	Result of stem before modification	Result of modification stem	Total of words
Cry/cries/cried/crying	cry/cri/cri/cry	cry/cry/cry/cry	20
Dry/dyed/dyed/dying	dye/dyes/dyed/dying	dy/dy/dy/dy	

(iv) Fourth Rule. The fourth rule deals with words ending in –s (not –ss) and the participle form of the word. Porter's stemming algorithm will remove the –s ending in words that have the –s ending. For example, the word "focus" becomes "focu". However, if the word ending in -s is in the past or present participle form, Porter stemming will only delete the ending -ed or -ing. For example, the words "focused", "focusing" become "focus" and "focus". **Table 3** shows the cases handled in the fourth rule.

Table 3. List of Case Categories handled in Fourth Order.

Original Words	Result of stem before modification	Rule	Result of modification stem	Total of words
Focus/Focuses /Focused/Focusing	Focu/focus /focus/focus	Changed -sed to -s and Change -sing to -s	Focu/focu /focu/focu	28

(v) Fifth Rule. Porter's stemming algorithm does not handle present or past participle derivations. For example, the word "studiedly" produces the stem "studiedli" which should be "study". The fifth rule will deal with words in that category. **Figure 3** shows the cases handled in the fifth rule.

DOI: http://dx.doi.org/10.17509/xxxx.xxx
p- ISSN xxxx-xxxx e- ISSN 2774-7891

No	Kategori	*Original Words*	Hasil *stem* sebelum modifikasi	Aturan	Hasil Stem Modifikasi	Total Kata
1	Kata yang berakhiran -iedly atau -iedness berelasi dengan kata yang berakhiran -ied	*study/studied/ studiedness/ studiedly*	*studi/studi/studied/studiedli*	mengubah –ly menjadi -ied mengubah -ss menjadi -ied	*study/study/study/study*	13 kata
2	Kata yang berakhiran -edly atau -edness berelasi dengan kata yang berakhiran -ed	*amaze/amazed /amazedly/ amazedness*	*amaz/amaz/amazedli/ amazed*	mengubah –ly menjadi -ed mengubah -ss menjadi -ed	*amaz/amaz/amaz/amaz*	439 kata

Figure 3. List of Case Categories handled in Fifth Order.

3. RESULTS AND DISCUSSION

Overstemming cases are handled based on the word dictionary contained in the database. This word dictionary was compiled using rules adopted from research by Karaa (2013) and a list of words from www.morewords.com. This case handling is implemented in the title and keyword attributes of publication documents from DBLP data as well as publication documents of the RPLP and Informatics Scientific Group at the School of Electrical Engineering and Informatics (STEI-ITB). The number of publication documents used was 3,012 scientific publications. This publication document has been classified into 7 categories of research fields, namely AI and Image Processing (AI), Computational Theory and Mathematics (CT), Computer Software (CS), Data Format (DF), Distributed Computing (DC), Information System (ISM), Library and Information Science (LIS).

Table 4 shows the number of words that experienced overstemming in the title attribute in published documents.

Table 4. Number of Overstemming Cases handled.

Atribut	Total of Overstemming word	Total of word	Percentage
Keywords	29	15.34	0.19%

Table 4 shows that title words have a greater number of overstemming cases than keywords. This is because keywords contain more phrases (combinations of words) so that many do not change.

The table above shows that it is possible that this type of word rule is still insufficient to handle cases of overstemming in title and keyword attributes. A more varied type of rule wording allows more of both cases to be handled.

Figure 4 shows the recall results for using the title attribute in stemming with modification and stemming without modification.

DOI: http://dx.doi.org/10.17509/xxxx.xxx
p- ISSN xxxx-xxxx e- ISSN 2723-4088

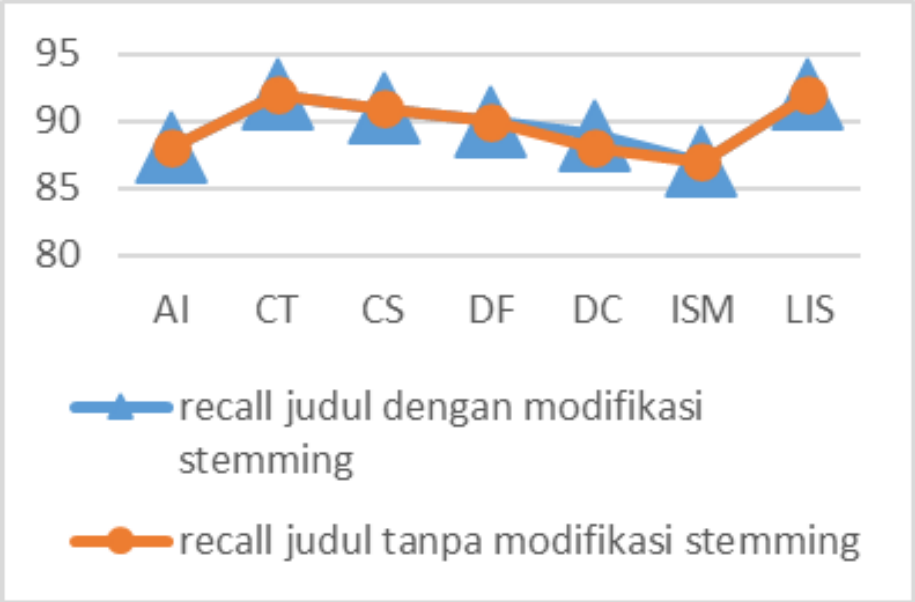

Figure 4. Comparison graph of recall values in titles.

Figure 4 shows that the recall generated by stemming with modification has a value that is almost the same as stemming without modification. These two processes produced average recall values of 89.9% and 89.7%. Differences in the recall value of title words occur in the DC category. The recall value for the DC category in stemming with modification has a higher value than the recall for the DC category in stemming without modification.

Stemming with modifications to the title attribute does not really result in an increase in recall results from the stemming process without modification. This is because the title attribute only has a slight variation in the words contained in the stemming rule with modification. Apart from that, words that undergo stemming with modification may include words that are general (have a small IDF value) in the title so that they do not really affect the increase in recall. General words mean words that have a frequency of occurrence in several categories, resulting in a small IDF value. Apart from that, there are not many variations in the types of word rules, so they are less able to handle cases of overstemming in the data used. This does not really affect the increase in recall.

4. CONCLUSION

Overstemming cases that are handled on the data used can increase recall results on the use of the title attribute, but the results are not very significant. This is caused by the large number of words that are not handled by the rule table proposed by Karaa (2013). The title attribute only has a slight variation of the words contained in the stemming rule with modification. In addition, words that undergo stemming with modification may include words that are general (have small value) in the title so that they do not really affect the increase in recall.

5. AUTHORS' NOTE

The authors declare that there is no conflict of interest regarding the publication of this article. The authors confirmed that the paper was free of plagiarism.

DOI: http://dx.doi.org/10.17509/xxxx.xxx
p- ISSN xxxx-xxxx e- ISSN 2774-7891

6. REFERENCES

Abu-Salem, H., Al-Omari, M., and Evens, M. W. (1999). Stemming methodologies over individual query words for an Arabic information retrieval system. *Journal of the American Society for Information Science*, *50*(6), 524-529.

Attar, R., and Fraenkel, A. S. (1977). Local feedback in full-text retrieval systems. *Journal of the ACM (JACM)*, *24*(3), 397-417.

Bakar, Z. A., Sembok, T. M. T., and Yusoff, M. (2000). An evaluation of retrieval effectiveness using spelling-correction and string-similarity matching methods on Malay texts. *Journal of the American Society for Information Science*, *51*(8), 691-706.

Jivani, A. G. (2011). A comparative study of stemming algorithms. *Int. J. Comp. Tech. Appl*, *2*(6), 1930-1938.

Joshi, A., Thomas, N., and Dabhade, M. (2016). Modified porter stemming algorithm. *International Journal of Computer Science and Information Technologies*, *7*(1), 266-269.

Karaa, W. B. A. (2013). A new stemmer to improve information retrieval. *International Journal of Network Security and Its Applications*, *5*(4), 143.

Krallinger, M., Rabal, O., Lourenco, A., Oyarzabal, J., and Valencia, A. (2017). Information retrieval and text mining technologies for chemistry. *Chemical reviews*, *117*(12), 7673-7761.

Moral, C., de Antonio, A., Imbert, R., and Ramírez, J. (2014). A survey of stemming algorithms in information retrieval. *Information Research: An International Electronic Journal*, *19*(1), n1.

Paice, C. D. (1996). Method for evaluation of stemming algorithms based on error counting. *Journal of the American Society for Information Science*, *47*(8), 632-649.

Panigrahi, P. K., and Bele, N. (2016). A review of recent advances in text mining of Indian languages. *International Journal of Business Information Systems*, *23*(2), 175-193.

Patel, M., and Shah, A. (2016). An unsupervised stemming: A review. *International Journal of Computer Science and Information Security*, *14*(7), 476.

Singh, J., and Gupta, V. (2017). A systematic review of text stemming techniques. *Artificial Intelligence Review*, *48*, 157-217.

DOI: http://dx.doi.org/10.17509/xxxx.xxx
p- ISSN xxxx-xxxx e- ISSN 2723-4088