Information Retrieval Technique for Indonesian PDF Document with Modified Stemming Porter Method Using PHP

Faizal Riza[1*], Saefulloh Rifai[1], Akmal Dirgantara[1], Sfenrianto[1], Rasenda[1], Syarifudin Herdyansyah[1]

[1]STMIK Nusa Mandiri, Jakarta, Indonesia

*14002266@nusamandiri.ac.id

Abstract. Finding relevant information from a collection of information requires a process of stemming. Stemming is the process of combining or solving each morphological variants of a word into a basic word. Based on the basic structure of the word morphology, Porter's stemming looks appropriate to be applied in conducting basic word searches in Indonesian-language documents, but with a few modifications. For this need, an Information Retrieval Technique for Indonesian PDF Document Application Using PHP from Indonesian documents is made using the Modified Stemming Porter Method. Implementation of the application was carried out using the Php (Hypertext Preprocessor) programming language. Testing was performed on 26 pdf e-book documents are 23,197 basic words out of 28,532 total words. the experiment found 94% as the largest percentage of precision words in the document. And the results obtained 81% as the lowest percentage of the basic words that are precise in the document. The results obtained from the test are that the application can operate well in conducting stemming on e-books in Indonesian.

1. Introduction

Electronic books (e-book) or digital books are electronic versions of books. If the book generally consists of a collection of papers that can contain text or images, then the electronic book contains digital information which can also be in the form of text or images [1]. At this time, it is not new if a student, student, teacher or even the general public keep documents in the form of e-books.

Search facilities using matching keywords with words in documents that are also provided by the operating system still cannot provide relevant search results. The word matching method will count the number of keywords that appear in the document then return the order of documents with the highest number of occurrences to the user. Because users are not looking for documents that contain many words that are the same as keywords, the word matching method is not an ideal solution for information retrieval. In addition, the number of words that have similar meanings (synonyms) and words that have more than one meaning (polysemias) in the use of keywords to express the information needed, will make the search results with the word matching method increasingly far from relevant. This can result in the performance of search engines becoming less good because they do not pay attention to the words semantically.

Stemming has been used extensively in electronic document processing. Stemming is used in several fields such as: information retrieval, question answering (QA), spell checking, machine translation, document clustering, document classification and others. Stemming [2] is a computational procedure

that converts words into their original form (stem) by searching for prefixes, suffixes and deleting them based on the rules of a language. The results of the stemming process are called tokens. One of the advantages of using stemming in the development of information retrieval systems is: efficiency and compressed file indexes. For example, like this: a searcher enters the term stemming as part of the query. That shows that the person is also interested in stemmed and stem. Without the process of stemming, the words "stemming", "stemmed" and "stem" are something different. With the stemming process, each word that has the same root word can still be equated even though it does not have the exact same words.

Stemmer that is done by Tala in [5] is a stemmer used by Indonesian text mining applications. All research has been done to change the form of words into basic words based on the morphological form of a word. This is not appropriate if it is used to look for basic words that are structurally unusual, for example: messing up has a different word structure by complaining, securing. The prefix and affix rules applied to stemmer [6] also experience some errors if the word to be searched for has a basic form ending with proprietary pronouns such as "-mu", "-nya", "-ku". For example the word “paku” becomes "pa" + "ku". This is because the suffixes of proprietary pronouns are executed first and there are no exclusion rules for root words that have suffixes belonging to them. Another weakness in stemmer [2] is that it does not provide a dictionary of basic words that can cause errors in words such as: "peranakan" which is processed into "per" + "ana" + "kan". This is because the algorithm used first sees the suffix rather than the suffix "-an". Research on porter stemmer was also carried out by Indriyono. He did the stemming process by looking at the structure of a word. Research conducted by Indriyono first looked at particle suffixes such as "-lah", "-lah", "-tah" and proprietary pronouns such as "-i", "-mu", "-nya" compared to seeing the prefix of a word [7]. This algorithm is inefficient because it takes 2 tables in each database for particle ending words and proprietary pronouns. In addition, the lack of algorithm in is that this algorithm does not take into account the basic words beginning with the letters "k", "t", "s", "p" correctly.

2. Word structure in Indonesian

2.1. Structure of Indonesian Grammar

Based on its structure, Indonesian words can be attached to 5 different types of affixes, namely: prefix, insert, suffix, personal pronouns and particles. Table affixes for indonesian words shows a list of affixes that can be added to Indonesian words.

Tabel 1. Affixes for Indonesian words

Category	Morfem	Example
Prefix	me-, di-, be-, pe-, ter-, se-	memasak, dibawa, berair
Insertions	-em-, -el-, -er-	gerigi, telunjuk
Sufix	-kan, -an, -i, -isme, -isasi	minuman, aktualisasi
Possessive pronoun	-ku, -mu, -nya	miliknya, mejaku
Particle	-lah, -kah, -pun	sudahkah, masuklah

The structure of an Indonesian word is formulated [2]:

[prefix1] + [prefix2] + root + [suffix] + [possessive pronoun] + [particle]

The structure shows that a word in Indonesian constructed from a base word using various morphological operations including combining, adding affixes and repetition [5]. Form of repetition of a word can be divided into 2 types, namely: full repetition and partial repetition. Not all combinations of prefixes and suffixes can be used together. There are some prefix and suffix combinations that are not permitted in Indonesian grammar as shown in table combination of prefix and suffix is not allowed.

Tabel 2. Combination of prefix and suffix is not allowed

Prefix	Sufix
ber	I
di	An
ke	i | kan
meng	An
peng	i | kan
ter	an

In general, morphology is divided into 2 processes, namely inflectional process and derivational process [5]. The inflectional process is the process of changing a word that does not change its basic word type. For example, the word "memukuli" comes from the basic word "pukul". The word "pukul" is a verb and the word "memukuli" is also a verb so that there is no change in the type of words in the process.

2.2. Second Levels of Morphology

Not all word formation processes from basic words can be completed with one level of morphology. Examples of word-formation by adding affixes to basic words with one morphological level are "mem" + "baca" to "membaca", "men" + "cari" to "mencari". The addition of affixes to the basic words to form new words by changing the phonemes of the basic words cannot be solved by one level of morphology. To solve this problem, the second levels of morphology are needed to solve this problem. The use of second levels of morphology is another way to describe phonemes in finite-state terms [5].

3. Research Methods

3.1. Data Collection

Data collection This research basically prioritizes aspects of the implementation and analysis of a theory, about stemming using the modified Porter algorithm approach. As in researches about the analysis or proof of a theory that many apply the method of library study, so it is in this study, the authors do more library studies by taking references from the internet, ebooks, papers, journals, theses, and books. books related to programming algorithms, stemming, especially in Indonesian, word formation in Indonesian, and various other scientific literature.

In this study, the authors used a basic word dictionary downloaded from the internet with a total of 28,532 words. These basic words are then stored in a database that will later be used in the stemming process. A collection of punctuation marks and numbers (48 unit) are stored in a file that will be used as a reference to eliminate punctuation and numbers in the preprocessing stage. The next step is deleting the stopword list. Stopword list is a general word that is considered not to provide important information, so that its existence can be ignored, for example the words: which, and, that is, are and so forth. In this study the authors used the Tala stopword list published in Asian research [8] consisting of 758 words. Like punctuation, stopword lists are stored in the form of text files which will be used as references to eliminate words that are included in stopwords.

The study was conducted on 26 test documents in Portable Document Files (PDF) format. The document is destemming using a php library called PHP simple PDF DOM parser, then saving it as an index document stored in the mySQL database.

3.2. System analysis

The stemming process begins by changing all capital letters into lowercase letters, then proceed with eliminating punctuation and numbers by referring to the punctuation file and numbers created. The next step is to change the series of words into a collection of terms (corpus), followed by checking whether words are found in stopwords. If found, the word will be automatically removed. Likewise, if words are found that appear more than once (double words), then the word will also be omitted and only one word will be processed.

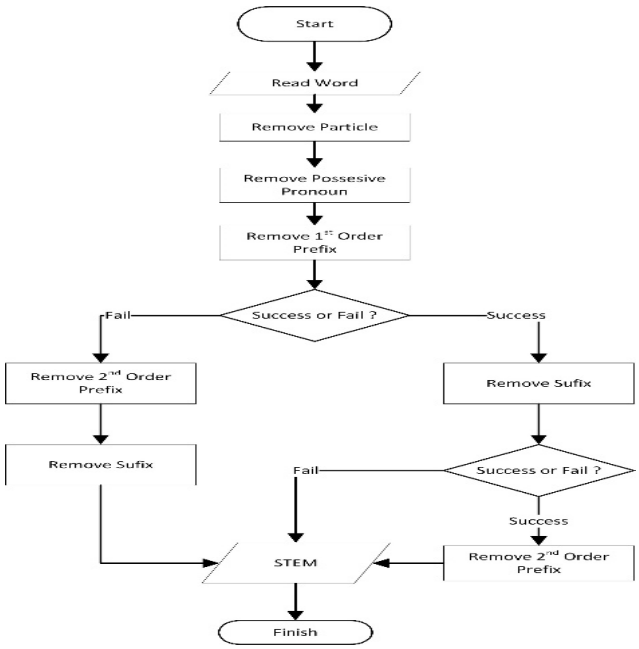

Figure 1. The algorithm used to improve the stemming process

As shown in figure of the algorithm used to improve the stemming process, there are deficiencies in the algorithm, which in this study will be upgraded are:

1) The algorithm in the figure 1 first removes particles and proprietary pronouns compared to prefix 1 and prefix 2. This algorithm uses a database to store words that do not need to be processed because they are the exception words for each process. For example, in the process of removing particles. This process removes the "-kah", "-pun", "-lah" particles. In this process, there are some words that do not need to be processed because they are basic words that contain these particles such as the word "menikah". The phonemes or "-kah" particles in the word "menikah" need not be removed because they form part of a basic word and the word "menikah" is entered into the database.
2) The algorithm in figure abaout the algorithm used to improve the stemming process does not take into account the second-level morphology that applies to basic words beginning with the letters "k", "t", "s", "p" correctly if given the prefix "pe-" and "me". The process of the word "menikah" cannot be applied to the word "menulis" because it can produce the wrong basic words. Therefore, an appropriate second-level morphology is needed to solve this problem

In general the process of stemming is divided into 5 parts, namely: eliminating the first prefix ("meng", "peng-", "mem", "pem-", "meny-", "peny-", "men-", " pen- ", etc.), removes the second prefix ("ber-"," per- "," ter- "," se- "," pel- ", etc.), removes particles (" -kah ", "-lah", "-tah"), removes personal pronouns ("-ku", "-mu", "-nya"), removes suffixes ("-kan", "-an", "-i", "-isme", "-isasi", "-onal").

3.3. Application Design

In this study a web-based application was built using the PHP programming language and MySQL database to facilitate the implementation and analysis of the modified Porter stemming algorithm. This application consists of 5 main pages, namely Documents, Stemming, Results, Basic Words, Stopword. Design application.

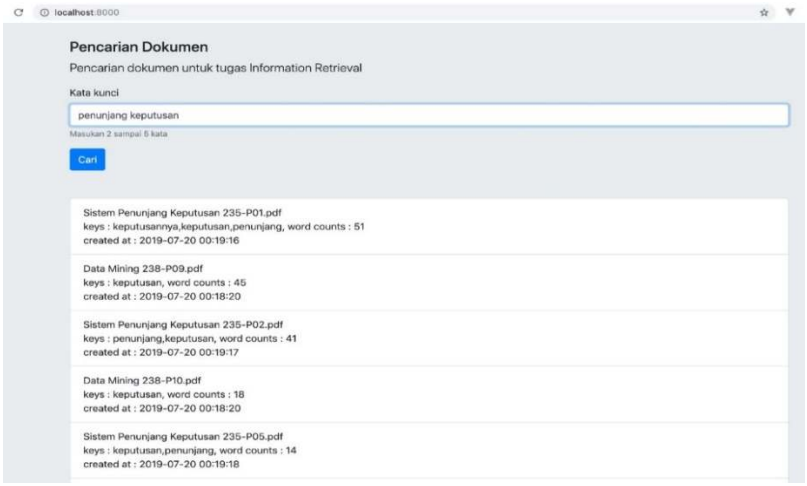

Figure 2. Aplication interface of searching method

4. Result Analysis

In assessing the implemented system, the researcher runs several tests in the form of search keywords. The study tested by applying a modified stemming potter algorithm. Evaluation is done by accepting valid files and rejecting invalid files, test cases are valid and invalid from the specified algorithm. from the implementation of the test obtained the results shown in table of result of searching test cases, which presents the results of test cases. The analysis column displays the results of the test results in which the basic form of the Indonesian word is provided as input, and the system performs stemming on the affix structure.

Tabel 3. Basic word search test results with the modified stemming porter method

No	Name of file	Words	Basic Words	Not Basic Words	Precision
1	PDF File 1	799	674	125	84%
2	PDF File 2	893	791	102	89%
3	PDF File 3	674	544	130	81%
4	PDF File 4	637	518	119	81%
5	PDF File 5	631	522	109	83%
6	PDF File 6	2.104	1.775	329	84%
7	PDF File 7	1.009	863	146	86%
8	PDF File 8	1.360	1.152	208	85%
9	PDF File 9	1.108	972	136	88%
10	PDF File 10	830	697	133	84%
11	PDF File 11	911	774	137	85%
12	PDF File 12	2.082	1.732	350	83%
13	PDF File 13	878	756	122	86%
14	PDF File 14	1.022	945	77	92%
15	PDF File 15	1.804	1.532	272	85%
16	PDF File 16	1.023	959	64	94%
17	PDF File 17	711	658	53	93%
18	PDF File 18	843	765	78	91%
19	PDF File 19	799	683	116	85%
20	PDF File 20	1.573	1.356	217	86%
21	PDF File 21	1.318	1.146	172	87%
22	PDF File 22	1.124	974	150	87%
23	PDF File 23	861	807	54	94%
24	PDF File 24	950	841	109	89%
25	PDF File 25	1.073	890	183	83%
26	PDF File 26	1.515	1.318	197	87%
	Total	28.532	24.644	3.888	
	Average				87%

Based on testing in table basic word search test results with the modified stemming porter method from 26 test documents in the Portable Document Files (PDF) format are 24,644 basic words out of 28,532 total words. The highest percentage of precision base words found in PDF File 24 and PDF File 23 is 94%. While the lowest percentage was found in PDF File 3 and PDF File 4 with a percentage of 81%. Information Retrieval Technique for Indonesian PDF Document with Modified Stemming Porter Method Using PHP has been successfully created. On the application stemming results page there is a link that can display the stemming results document which contains a collection of basic words of the Indonesian language. Words that have errors in the stemming process are not in the dictionary database, nor are parts of a foreign language dictionary or other dictionary enclosed in parentheses () in the stemming documents.

5. Conclusion

Based on the results of basic word search test results with the modified stemming porter method, it can be concluded that the application of Information Retrieval Technique for Indonesian PDF Document with Modified Stemming Porter Method Using PHP has been successfully created. Basic word search results found in 26 pdf e-book documents are 23,197 basic words out of 28,532 total words. the experiment found 94% as the largest percentage of precision base words in the document. And the results obtained 81 % as the lowest percentage of the basic words that are precise in the document. The difference in percentage results is obtained from the number of basic words found by the application. The word contained in the column is not a base word on the stemming results page is a word that did not succeed in becoming an Indonesian basic word due to an error in the stemming process, not in the dictionary database, nor is it part of a foreign language dictionary or other dictionary, in stemming documents, words that don't work became the basic Indonesian word enclosed in parentheses ().

References

[1] M. Vassiliou and J. Rowley 2008 *Progressing the definition of 'e-book'* (Libr. Hi Tech)
[2] F. Z. Tala 2003 *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia*, (M.Sc. Thesis, Append. D).
[3] S. D. Larasati, V. Kuboň, and D. Zeman 2011 *Indonesian morphology tool (MorphInd): Towards an Indonesian corpus* (Communications in Computer and Information Science)
[4] M. Adriani, J. Asian, B. Nazief, S. M. M. Tahaghoghi, and H. E. Williams 2007 *Stemming Indonesian: A confix-stripping approach* (ACM Trans. Asian Lang. Inf. Process)
[5] F. Pisceldo, R. Mahendra, R. Manurung, and I. W. Arka 2005 *A Two-Level Morphological Analyser for the Indonesian Language* (Proc. 2008 Australas. Lang. Technol. Assoc. Work. (ALTA 2008))
[6] M. Abdi and T. Corresponding 2015 *Links between Bloom's Taxonomy and Gardener's Multiple Intelligences: The issue of Textbook Analysis* (Adv. Lang. Lit. Stud., vol. 6, no. 1,)
[7] B. V. Indriyono, E. Utami, and A. Sunyoto 2015 *Pemanfaatan Algoritma Porter Stemmer Untuk Bahasa Indonesia Dalam Proses Klasifikasi Jenis Buku* (J. Buana Inform)
[8] J. Asian 2007 *Effective Techniques for Indonesian Text Retrieval* (ph.D. Thesis)