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

DYNAMIC STOPLIST GENERATOR FROM TRADITIONAL INDONESIAN CUISINE WITH STATISTICAL APPROACH

[1]SETYAWAN WIBISONO, [2]MARDI SISWO UTOMO

[1, 2] Faculty of Information Technology, Universitas Stikubank, Semarang, Indonesia
Email : [1]setyasonny@gmail.com, [2]mardi_2@unisbank.ac.id

ABSTRACT

Stoplist is one part of input for information retrieval system that can affect information retrieval quality. The existence of words that are not meaningful can make the retrieval declining. The standart dictionary-based information stoplist also has problem when implemented in a corpus with specific domains. For example the word "recipe" is not a stopword but when using it in domain cuisine, recipes will appear in almost every document. We build dynamic stoplist using Indonesian recipes documents, this documents has non standart dictionary-based stoplist that interesting to study. This paper use three methods to generate stoplist. We use poisson and binomial probability distribution approach and we also use simple frequency distribution approach for classifying candidate stopword. For measuring the result we also employ recently proposed RAKE algorithm. All three of these methods have the same weakness, the stoplist can be generated appropriately if the entire population of the all corpus vocabulary has processed, unlike the stoplist dictionary which can already detect stopword at the stage of pre-processing. The results of the frequency distribution is better than the other methods, but this method requires a longer process than poisson and negative binomial method.

Keywords : *Keyword Extraction, Indonesian Cuisine,Auto Generated Stoplist*

1. INTRODUCTION

Stoplist is one part of input for information retrieval system that can affect information retrieval quality [1]. The existence of words that are not meaningful can make the retrieval declining. In addition to lowering the yield retrieval, improper use of stoplist improper also resulted in the indexing process becomes longer and greater use of resources. In the field of retrieval several researchers tries to make resources saving technique such as Khalaf and Iman [2], they use query tree method to save both cpu resource and storage space.

Stoplist generally using dictionary-based stoplist, usually called information retrieval standard stoplist. This stoplist must have the same language with the corpus language that made stoplist is not suitable for other language. In Indonesian language we commonly used stoplist published by Tala [3]. In a case study Khalaf et al [4], they use the standard Arabic language stoplist to build the data retrieval system index structure. The standart dictionary-based information stoplist also has problem when implemented in a corpus with specific domains. Some words seems that were previously not part of the standard dictionary-based information stopword but potentially became a stopword due to appear in almost every document in corpus. For example the word "recipe" is not a stopword but when using it in domain cuisine, recipes will appear in almost every document.

In this paper we are trying to build stoplist dynamically using three methods, we generate stoplist with poisson and binomial probability distribution approach Jungiewicz, M. & Lopuszyński, M [5] did. The last method we use the frequency distribution for classifying candidate stopword. For measuring the result we also employ recently proposed RAKE algorithm [6]. This Algorithm was designed to extract keywords from individual documents as domain-independent, unsupervised and language-independent method. All those features make this RAKE an excelent candidate tool for extracting keywords from Indonesian Traditional Cuisine.

The corpus consist 597 Indonesian recipes documents. The documents are collected individually from reputable various Indonesian culinary sites such as sajiansedap.co.id, bango.co.id [7]. In our opinion, this set of documents is challenging and interesting. It contains diverse data vocabulary, not only related to food and health issues, but also for technical discussion contracts coming from any different fields (medicine,cultural history, IT, etc.)

2. RELATED WORK

In word frequency study George K. Zip created an empirical law that relates terms frequencies (tf) to rank in a frequency ordered word list [8]. Hans Luhn use statistical words in text information to measure the individual words and phrases [9] significance. With this method Luhn has conclusion that the most discriminant words are those appearing in the middle of the frequency rank. Salton use the document frequency (df) to become a measure of the words discriminatory capacity [10]. Both of them suggested that words can appear in a document collection either in a random manner or concentrated in some paper and they proposed the term frequency times and the inverse document frequency product (tf • idf) as a measure of the significance degree. The words appeared in many documents (df high) or with a low presence (tf low) can be state as stop words.

In the 90 Christopher Fox developed a list of stopwords [11] extracted from the Brown Corpus of English literature Based on these frequency descriptions. These stopwords considered as the standard or conventional list and they have been widely used, this method too general to take into specific text colections such as culinary recipies. They not suitable to filtering words that belong to specific research fields or words of recent apparition. Specific stop words can be differ from one corpus domain to another domain as Makrehchi & Kamel [12] suggest. Researcher has been purposed some method recently to generate new stop words lists dynamicly, customized to specific domain.

Poisson distribution is used by Church, K. and Gale, W. [13] to generate stop words list. This method has been used in research to generate stop words list for particular Polish texts [5]. S. Rose [6] has excelent method to exctract keywords from document unsupervised, this method can be used in any domain corpus and any document language. This purposed method called Rapid Automatic Keyword Extraction, or RAKE algorithm. This was a supervised method to identify stop word lists dynamicly based on words adjacent to determine that keywords could be stop words. Jungiewicz, M. & Lopuszyński, M [5] using poisson and negative binomial to generate stoplist and use the stoplist to extract keywords using RAKE algorithm

With stop words list, Candidate keywords extracted from document. RAKE uses this stop word to tokenize the keyword (it can be one word and more). After keywords extracted from document, this words are scored by calculating word co-ocurrences using a metric that belong to long keywords. The top T candidates are chosen as final keywords. That's why the stoplist is the most important thing in the RAKE Algorithm process.

3. AUTOMATIC STOPLIST GENERATION

All the vocabulary that appears in the corpus vocabulary recorded in the database table, cf and df value also recorded in the database table for each vocabulary that appears. df is the number of documents with a given word and cf its frequency in the collection. We don't perform stemming process in this research. The words only through the filtering process separating punctuation and tokenizing.

We add four vocabulary stopword flag attribute for the every word that appears. The first attribute is a dictionary-based flag attribute that will have 1 value if the word is in the dictionary-based stoplist. The second attribute is poisson stopword flag which will have 1 value if the word is a member of poisson distribution stoplist. The third attribute is stopword for binomial that will have 1 value if the word is a member of binomial distribution stoplist. The fourth attribute is stopword flag for frequency distribution stoplist will have 1 value if the word is a member of frequency distribution stoplist. The database table structure vocabulary used in this study shown in Table 1.

Table 1: The Database Table Structure Vocabulary

Field Name	Type	Description
*id	Bigint(20)	Word Id
word	varchar(30)	Word
df	Int(11)	Number of document
cf	Int(11)	Colection frequency
stopd	varchar(1)	Stop word flag (dictionary)
stopp	varchar(1)	Stop word flag (poisson)
stopb	varchar(1)	Stop word flag (binomial)
stopf	varchar(1)	Stop word flag (dist freq)
poisson	float	Poisson value
binomial	float	Binomial value

Dictionary base stopword atribut can be given a value at once when the word presence detected in the corpus. But it can't be done for the other three stopword attribute. The value of stopword flag attribute can be obtained after entire corpus vocabulary detected by the system. The cf and df value need to be calculated first in order to compute those three stopword attributes. cf and df value can be calculated after after all the vocabulary in the corpus population has been recorded.

In Figure 1 is shown a graph plots from the entire corpus vocabular. In Figure 1 are shown the largest vocabulary population has low df and cf value. In figure 2 is shown a graph plots of stopword derived from the dictionary-based stopword. In Figure 2 the same thing happen on the entire population graph vocabulary, which is the largest population has low df and cf value. Low df value indicates that the words that appear in a small portion within document corpus. It can easily be seen that by using dictionary-based stopword only a few words in the corpus with high cf detected as a stopword.

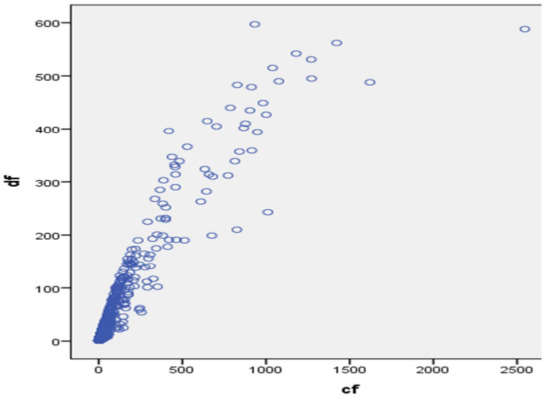

Figure 1: The Number Of Documents With A Given Word (df) vs Its Frequency In The Collection (cf) Plain Plots For The Whole Corpus Vocabular

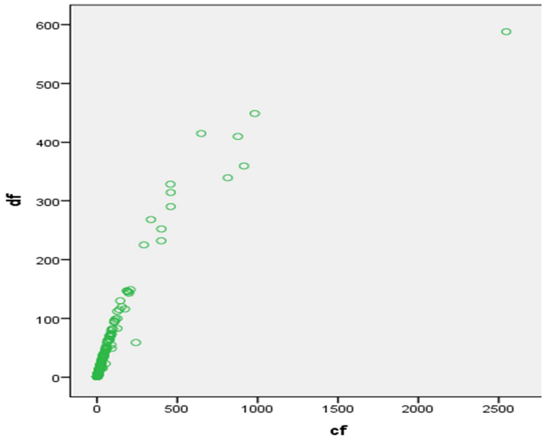

Figure 2: The Model Making Use of The Dictionary Stoplist Approach

Table 2 is the result of using the keyword extraction algorithms with dictionary-based stoplist. In the table we can see 5 keywords with the highest value, the keywords that successfully extracts still has a long sentence that is up to 73 words. With long results of keywords indicates that generated keywords can be strayed from the document meaning. We use rake algorithm php code that published at github [14].

Table 2: Top 5 Keywords Extracted Using Dictionary-Based Stoplist

text	word_count	value
manisnya kecap bango buah tomat cincang kasar buah…	73	4555.33
sambal terasi sdm garam sdm jawa sdm terasi b…	70	4085.30
merebus sendok teh garam sendok the lada putih bub…	69	3633.97
kecap bango manis pedas gurih terimakasih kecap ba…	66	4099.53
nikmat batang daun bawang iris halus butir telor a…	66	3348.33

The next step poisson approach was used to determine the candidate keywords. We use the Poisson distribution that is in use by Jungiewicz, M. et al [5]. We use Jungiewicz, M. simplest method using two variables to deviation detection. The number of documents is the first variable, which a given word is present **df** and the cumulative collection word frequency **cf**. The relation of **df** to **cf** in a large set of documents is defined by the probability theory in [12] the randomly distributed stopwords

$$\mathbf{df}\,(\mathbf{cf}) = N(1 - P(0, \mu = \mathbf{cf}/N)) \qquad (2)$$

where N is the total number of documents, and $P(0, \mu)$ is the probability of the word occurring 0 times, provided its average number of occurrences per document μ (by definition $\mu = \mathbf{cf}/N$). According Jungiewicz, M. dkk [5] For the simplest Poisson model the equation words for stoplist using the separating line **cf/df** = 1.6. it can be reduces to

$$\mathbf{df}\,(\mathbf{cf}) = N(1 - \exp(-\mathbf{cf}/N)) \qquad (3)$$

Figure 3 shows the data plot stopword generated by using the formula (3), resulting in uneven distribution stopword for high df and cf value, but doesn't produce a stopword in low df and cf value.

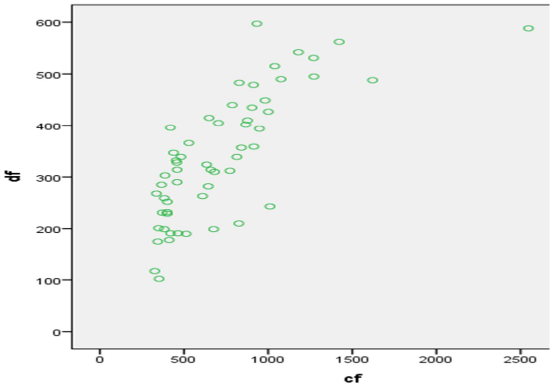

Figure 3: The Model Making Use Of The Poisson Distribution Approach (3)

Table 3 is the result of using the keyword extraction algorithms with poisson approach generated stoplist. In the table we can see 5 keywords with the highest value, the extracted keywords seem much better than dictionary-based stoplist although still has a long word, it contain 27 words. In Table 4 are shown by combining dictionary-based stoplist and poisson generated stoplist, the shorter keywords generated, it has 12 word long.

Table 3: Top 5 Keywords Extracted Using Poisson Stoplist

text	words_count	value
bulat tipis lebih mudah bila ada parutan	27	601.98
memperkaya pilihan legenda kuliner nusantara lonto…	26	568.36
kacang inipun menggunakan iga kambing alih alih ig…	25	513.70
berasal dari pesisir jawa barat khususnya daerah b…	24	513.00
awalnya ketika idul adha kemaren banyak banget dap…	24	520.37

Table 4: Top 5 Keywords Extracted Using Dictionary Stoplist And Poisson Stoplist

text	words_count	value
hati kambing dipotong kotak batang serai diambil p…	12	92.34
kaya citarasa rempah rempah khas kuliner indonesia	12	132.50
batang korek api bahan isian sediakan lembar kulit…	12	119.33
pelengkap sambal seledri cincang serai memarkan so…	12	129.50
ruas jari jahe haluskan ruas jari kunyit haluskan…	11	96.00

We use negative binomial model distribution to compare the Poisson distribution. It is related to the Poisson variable, but larger variance can be used in binomial distribution. It can be also represented as infinite combination of Poisson distributions with different μ. After substituting the negative binomial probability distribution function for $P(0, \mu)$ in (2), we get

$$\overline{\mathrm{df}}\,(\mathrm{cf}) = N\left(1-\left(1+\frac{\mathrm{cf}}{Nr}\right)^{-r}\right), \qquad (4)$$

where $r > 0$ is the negative binomial distribution additional parameter. the negative binomial variable converges to the Poisson model, In the case of r ! 1 with fixed μ. Fig 4 shows a plot stoplist resulting from the negative binomial distribution

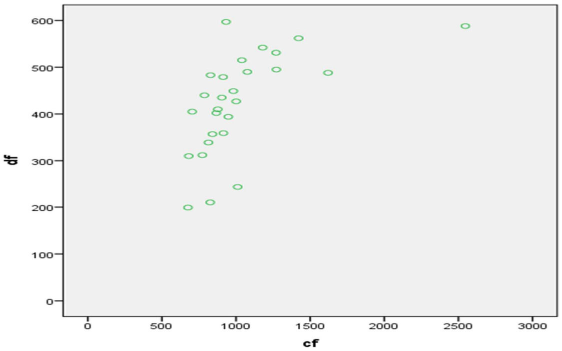

Figure 4: The Model Making Use Of The Negative Binomial Approach (4)

Table 3 is the result of using the keyword extraction algorithms with negative binomial approach generated stoplist. In the table we can see 5 keywords with the highest value, the extracted keywords looks similar with poisson generated stoplist keyword and it contain 27 words. In Table 4 are shown by combining dictionary-based stoplist and negative binomial generated stoplist, the shorter keywords generated, it has 12 word long same as poisson generated stoplist but with different rake value.

Table 5: Top 5 Keywords Extracted Using Negative Binomial Stoplist

text	words_count	value
bulat tipis lebih mudah bila ada parutan	27	601.98
memperkaya pilihan legenda kuliner nusantara lonto…	26	568.36
kacang inipun menggunakan iga kambing alih alih ig…	25	513.70
berasal dari pesisir jawa barat khususnya daerah b…	24	513.00
awalnya ketika idul adha kemaren banyak banget dap…	24	520.37

Table 6: Top 5 Keywords Extracted Using Dictionary Stoplist And Negative Binomial Stoplist

text	words_count	value
hati kambing dipotong kotak batang serai diambil p…	12	92.34
kaya citarasa rempah rempah khas kuliner indonesia	12	133.50
batang korek api bahan isian sediakan lembar kulit…	12	119.33
pelengkap sambal seledri cincang serai memarkan so…	12	129.50
ruas jari jahe haluskan ruas jari kunyit haluskan…	11	96.00

For the third method we use frequency distribution approach to generate the stoplist. The frequency distribution is made for the following reasons: large data sets can be summarized, can obtain some picture of the characteristics of the data, and is the basis for graphing.

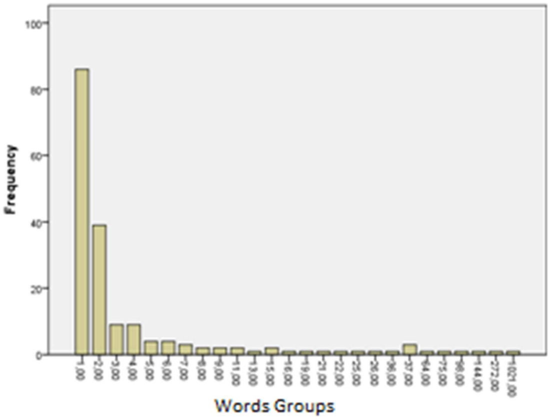

Figure 5: Word Classification Graph

When preparing Frequency Distribution Table, it must be confirmed that the classes do not overlap so that each observation values should be entered exactly into one class. Make sure also that there would be no observational data is left behind. Using the same width for all classes, although it is sometimes not possible to avoid an open interval. The calculation resulting from the class as much as 9 classes (5) and the class interval of 67 (6). The resulting frequency distribution table are shown in Table 7. In Figure 6is a plot of stoplist data with frequency distribution approach

$$1+(3.3 * \log(179)) = 8,43$$
$$= 9 \quad \textbf{(5)}$$

$$598 / 9 = 66,33$$
$$= 67 \quad \textbf{6)}$$

Table 7: Tabel distribusi frekuensi

Class	Range	Value
1	1-67	2183
2	68-135	65
3	136-203	32
4	204-271	10
5	272-339	13
6	340-407	8
7	408-475	6
8	476-543	9
9	544-611	3

After frequency distribution table of the existing corpus has obtained, we tried to establish a group stopword is a word that is contained in 0.5 * total class number, in this case we have 0.5 * 9 = 4.5 class with the highest value. So filtering using the words that are included in the top class (4 classes past the greatest value). That class 340-407,408-475, 476-543, 544-611. All said four members of the class designated as stopword.

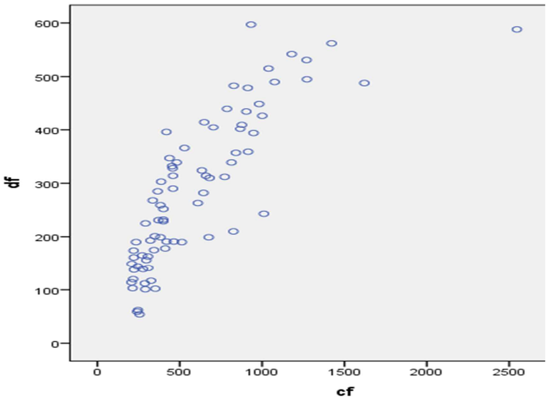

Figure 6: The Model Making Use Of The Frequency Distribution Approach (4)

Table 8 is the result of using the keyword extraction algorithms with frequency distribution approach generated stoplist. In the table we can see 5 keywords with the highest value, the extracted keywords shorter compared by poisson and binomial stoplist generated keyword. In Table 9 are shown by combining the dictionary-based stoplist and frequency distribution stoplist. Its produce shorter keyword compared with other methods.

Table 8: Top 5 Keywords Extracted Using Frequency Distribution Stoplist

text	words_count	value
memperkaya pilihan legenda kuliner nusantara lonto…	26	563.94
lebih mudah bila ada parutan singkong cuci bersih…	25	498.13
diwariskan turun temurun dari mendiang nenek edi s…	22	440.97
istilah balapan maka dari keunikan seperti itu lon…	23	427.67
khasnya kota semarang dimasaknya menggunakan arang	21	397.17

Table 9: Top 5 Keywords Extracted Using Dictionary Stoplist And Frequency Distribution Stoplist

text	words_count	value
kaldu perendam kupas kulit ari cumi buang kantong	11	106.89
memasukkan citarasa belimbing wuluh nan segar data…	10	84.20
warung sate klatak mak adi terletak jalan imogiri…	10	82.13
irisan cabai rawit ditaburi nori rumput laut bungk…	10	74.13
mantab banget deh coba	9	72.70

aja kalo percaya ikat kangk…		

4. RESULT

In the end we did stopword mapping generated from the 3 methods above statistics using Venn diagrams. In Figure 7 show stoplist slices produced by these three methods. Frequency distribution and Poisson distributions have the same number of words in the stoplist results, while negative binomial have fewer number of words for the stoplist is generated. All three of these methods share the same 56 words from their list of stoplist.

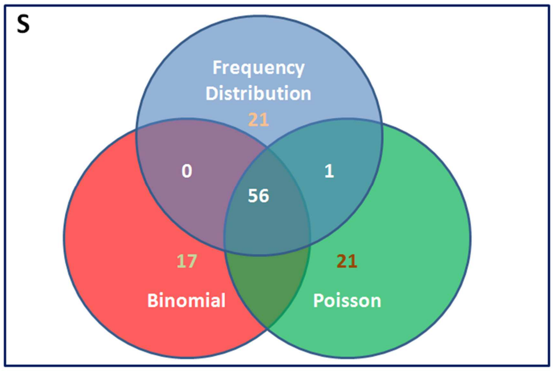

Figure 7: The Model Making Use Of The Poisson Distribution (3) With The Negative Binomial Approach (4)

In Table 10 are shown a summary of the results of the third algorithm used to generate the stoplist. The table 10 shows that extracted keywords using frequency distribution generated stoplist has better result than the Poisson and negative binomial approach.

Table 10: Stoplist Generator Summary Result

Generator	Num of Stopword Found	Longest Keyword Extracted	Highest Rake Value
Dictionary	248	73	4555.33
Poisson	78	27	601.98
Negative Binomial	73	27	601.98
Frequency Distribution	78	26	563.94

5. CONCLUSION

The methods that used are capable of producing stoplist, and seems has better result when used for the extraction of keywords using RAKE algorithm. Judging from the length of keywords generated by frequency distribution that has a better result but this method needs to be tested further in complete information retrieval system. We recommend using precission and recall measurement for further analysis.

All three of these methods have the same weakness, the stoplist can be generated appropriately if the entire population of the all corpus vocabulary has processed, unlike the stoplist dictionary which can already detect stopword at the stage of pre-processing. The results of the frequency distribution is better than the other methods, but this method requires a longer process than poisson and negative binomial method.

For further research we will try on different domains and with a larger corpus, as well as doing more recall and precission measurements on results.

REFERENCES:

[1] C.D. Manning, and H. Schütze. "Foundations of Statistical Natural Language Processing", *MIT Press*, Cambridge, Mass., 1999.

[2] K. Khatatneh and I. Hussein, "Information Retrievals Tries Tree Vs Inverted File Word Method for Arabic Language", *Journal of Theoretical and Applied Information Technology*, 2010.

[3] F. Z. Tala, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia", *M.S. Thesis. M.Sc. Thesis*. Master of Logic Project. Institute for Logic, Language and Computation. University van Amsterdam The Netherlands, 2003.

[4] K. Khatatneh, M. Wedyan, M. Alham, and B. Alrifai. "Using New Data Structure to Implement Documents Vectors in Vector Space Model in Information Retrieval System", *Journal of Theoretical and Applied Information Technology*, 2010.

[5] M. Jungiewicz and M. Lopuszyński, "Unsupervised Keyword Extraction from Polish legal Texts. Advances in Natural Language Processing", *Springer LNCS*, 65–70, 2014.

[6] S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic Keyword Extraction from Individual Documents Text Mining: Applications and Theory", *John Wiley & Sons*, Ltd., 2010.

[7] S. Wibisono and T. Aryanto, "Aplikasi Web Scraping Untuk Koleksi Konten Resep Masakan Jawa Berbasis XML". *Universitas Stikubank Research*, 2014.

[8] K. Zipf, "Selective Studies and the Principle of Relative Frequency in Language", *MIT Press*, Cambridge, 1932.

[9] P. Luhn, "The Automatic Creation of Literature Abstracts", *IBM Journal of Research and Development*, 2(2) 159–165, 1958.

[10] G. Salton and S. Yang, "On The Specification of Term Values in Automatic Indexing", *Journal of Documentation*, 29(4), 351–372, 1973.

[11] C. Fox, "A stop list for general text". *ACM-SIGIR Forum*, *24*, 19-35, 1990.

[12] M. Makrehchi and M. Kamel, "Automatic Extraction of Domain-Specific Stopwords from Labeled Documents". *Berlin/ Heidelberg: Springer*, 2008.

[13] K. Church and W. Gale, "Poisson Mixtures", *Journal of Natural Language Engineering*, 2, 163-190, 1995.

[14] https://github.com/Richdark/RAKE-PHP.