R. M. Hadi
Al-Mustansiriy university
Baghdad, Iraq
raghad_alrudieny@yahoo.com

S. H. Hashem
University of Technology
Baghdad, Iraq
soukaena.hassen@yahoo.com

A. T. Maolood
University of Technology
Baghdad, Iraq
abeer282003@yahoo.com

Received on: 28/09/2016
Accepted on: 23/02/2017

An Effective Preprocessing Step Algorithm in Text Mining Application

Abstract*- Text mining was a process of mining the significant information from the text documents. Any text mining system was created its process by preprocessing step; which involve tokenization, stop words removal, stemming and finally creating term frequency and inverse document frequency matrix (TF-IDF matrix). These steps provide the highest time consuming stage in knowledge discovery. The proposed method tries to build effective preprocessing step to even win area of memory space and time requirements. That by proposed a method for improved stop words removal algorithm and improved stemming algorithm based porter stemming algorithm. The proposed method is tested in two levels, first level uses only vector space model which based on used traditional stop words removal and with traditional porter stemming and the second level uses vector space model with combined features of improved stop words removal algorithm and improved stemming algorithm. The results show that using second level as effective preprocessing step for text mining application achieves good performance from reducing storage space used in memory about 10% and the processing time become faster which achieves good performance to build the final TF-IDF matrix*.

Keywords*- Document clustering, stop word, stemming, tokenization, TF-IDF, similarity*

How to cite this article: R. M. Hadi, S. H. Hashem and A. T. Maolood, "An Effective Preprocessing Step Algorithm in Text Mining Application" *Engineering & Technology Journal,* Vol. 35, Part B. No. 2, 2017, pp. 126-131, 2017.

1. Introduction

Data preprocessing was used for extracting useful and stimulating knowledge from text data [1]. Pre-processing steps contain the pool of documents which must be collected and transmitted it to the next step which represented by word tokenization step in which the collection of words was tokenized [2]. The pool of data has a feeble analytical value. For this purpose the idea of knowledge discovery was produced, such that data assembly, data pre-processing and data alteration was complicated in Knowledge discovery. The knowledge discovery is described with extensive series of variables and data bases [3]. The pool of words is the most central component in text document which characterize by words vector. From these components are stem words which collected in glossary or expert document gathering. This outcomes in drawbacks such as great dimension vector (the words vector contains large number of unique words) so text document must be preprocessed first [4]. By tokenization, stop words removal, and stemming.

- Tokenization phase, it is the process of identifying the token and their count [6]. The identification of token for all input documents represents most often important process, by the tokenization phase the search was summary with important degree [5]. In adding to decent use of storage space required to store important tokens identified from input documents with less storage spaces. The tokenization was proposed in any preprocessing methodology in which documents vectors were primary depended on token credentials [6].
- Stop words, are a parts of natural language. The purpose of removing stop-words from a document text is to sort the text to its appearance more order by eliminating the less important word used in parsing process, the eliminating process helps to decrease dimension of token space with high degree. In text documents the articles, pro-nouns, and prepositions etc. are the greatest mutual words. That does not give the important information of the documents. These verses can be considered as stop words and cannot deal with it as keywords in text mining process. Example for stop words: the, able, all, a, etc. [5].
- Stemming, normally many of the documents contain one word that have several forms , thus the stemming gives a mapping of these various forms belonging to the same word, the variants of words into their base word called the stem. Stemming process is used in information retrieval as a way to improve retrieval performance based on the assumption that terms with the same stem usually have similar meaning. To do stemming operation on large data, the more

https://doi.org/10.30684/etj.2017.138624
2412-0758/University of Technology-Iraq, Baghdad, Iraq

computation time and power was required, to cope up with the need to search for a particular word in the data [7]. Example on stemming is shown in figure (1).

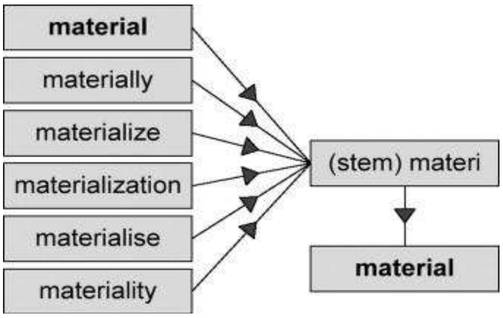

Figure1: Example of stemming process

Types of stemming algorithms are [6];

1) **Look up table Approach:** One of the most important and simple method in stemming. It works mental by excluding all terms with its key and their stems term in list (table). Then any terms derive from the system or the terms key are compared with lookup table stemmed approach, using hash tables, such lookups are very fast, but there are problems through using this table, First there is no such data for English, even if there were they may not be represented because they are domain detailed and may be required to use another approach of stemming, in addition to the storage overhead [1].

2) **Affix Removal Stemmers:** this type of stemming used to delete the suffixes or prefixes from the terms and will remain the root of the words. One of the examples of the affix removal stemmer is one which deletes all the multitude of documents terms. Some set of rules for such a stemmer are as follows (Harman) [1];

- Replace the “ies” from the word ends to “y”
- Replace the “es” from the word ends to “e”
- Replace the “ss” from the word ends to “s”.
- Replace the “s” from the word ends to “NULL”

3) **Porter’s stemming:** Is one of the very widely used stemming method in English language to terminate the end of the words syntactical. Word regulation represents the main use of porter stem. Terminating the word ends addition (suffixes) was based primarily by porter stemming approach , such as gerunds (traveling - travel), plurals (boys - boy), and swapping words ending with "i" for example with "y" , etc. all the task was focused into procedures where each of these procedures contracts with a exact suffix and having sure form(s) to fulfill. An assumed word’s suffix is patterned beside every instruction in a serial way till it equals one, and thus the situations in the instruction are verified on the stem that can outcome in a suffix elimination or adjustment [8].

2. Related works

The approach presented in [3] is to find the damage to the use of electronic documents over databases. The solution is by text illustration which is the critical step for text pre-processing .Text (document) is a pool of words, in [3] Research was recognized in numerous steps. Text assembly, Format cancelling, Data pre-processing on numerous levels, with subsection serial identification. And used a stretch sequence identification. With stop words subtraction and paragraph sequence identification with stop words elimination and a sentence order identification. In [5] the paper discussed about the data mining which used for finding the useful information from the large amount of data. It tries to find interesting patterns from large databases. It uses different pre-processing techniques likes stop words elimination and stemming. In [9] their methodology was used an actual preprocessing stages to protect both galaxy and time supplies by using developed stemming algorithm. Stemming algorithms was castoff to alter the words in texts into their correct origin formula. In [10] Mining text document from a preprocessed stage was calmed as relate to natural languages documents. Thus, preprocessing phase it a significant process in text mining application. The paper was talk about shrink the dimensionally of the words space, different procedures such as cleaning (filtering) and stemming are practical. Filtering methods eliminate those words from the regular of wholly words, which do not offer related evidence; stop word filtering is a typical filtering manner.

3. Proposal of Preprocessing: The proposed system was described by the following sequential steps to extract useful words. The flowchart of proposed preprocessing steps is shown in (Figure 2).

Step 1. Extraction the Documents

The proposed system was selected a domain from Reuters 21578 datasets. collected whole documents from datasets by using body based feature: All body-based features existing in the body of Reuter’s document that includes: (body-keyword), (<body >), (body-java script), and etc. after these body the content of document begin, each body document in datasets was represented using the bag-of-words approach, also these representation known as Vector space model (VSM): it includes the words as column and the documents as rows in VSM matrix. The file content must tokenized into individual word by the algorithm shows the first function of the proposed system

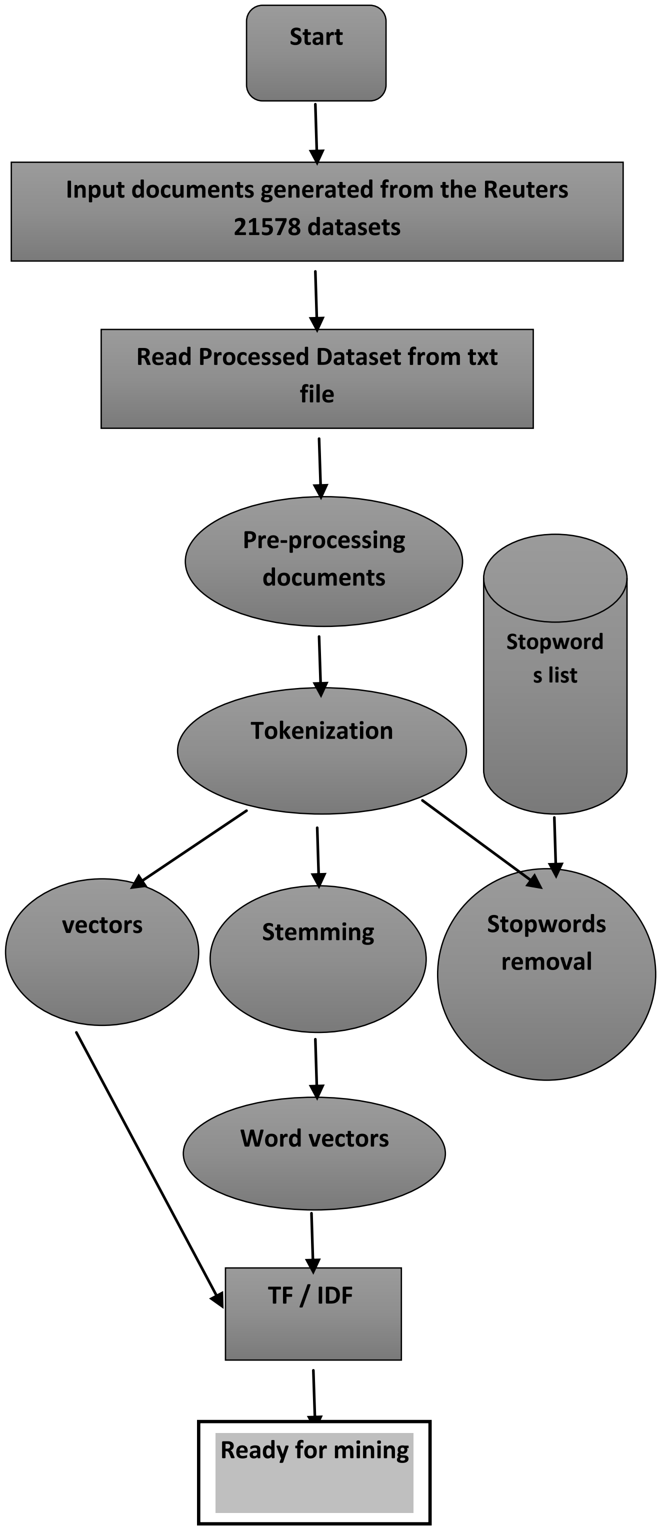

Figure 2: pre-processing methods

Algorithm1: Tokenization Algorithm
Input: Dataset
Output: Tokens (Ti)

Step1: Input documents from the proposed system and collected as (IDi), i=1, 2, 3….n;

Step2: Abstract useful word (AWi) = IDi, from each input document IDi; // cutting word by separated words using white space (space, tab, newline) as the delimiter process for all documents this process it apply, i=1, 2, 3…n //

Step3: **Tokens** (Ti) will be passed to the second step in preprocessing methods in an IR system.

End

Step 2. Extraction of Meaningful Words by removing stop word

Less meaningful word was removed from a text (stop words) which represent the second important process done by the proposed system. In order reduces the dimensionality of TF-IDF matrix (VSM representation). The greatest mutual words in edition documents are articles, prepositions, and pro-nouns, etc. that was not ensure spring the denotation of the documents. These words are preserved as stop words. In the stop words removal function the proposal system uses the following methods:

The Classic Method: it is traditional and simple method based on removing stop words by the words were compared with the words were stored in list so if there are any match the word was removed from text, the stop words list of the propose system are shown in table(1).

Table (1): Stop words list

1	a
2	A’s
3	able
4	about
5	above
6	according
7	accordingly
8	across
9	actually
10	after
11	Again
12	…. Etc.

Law (Zips-Methods): In addition to the classic stop list, the proposed system was used the stop word creation methods moved by Zipf‟s law, including: the word was delete that shows in the input text once (occur once), i.e. singleton words which means term frequency value of word equal to one (TF1). The proposed system also consider removing words with low term frequency and inverse document frequency (TF-IDF) value by first removed stop words from word vector using stop words list then apply stemming function, and calculate term frequency TF as following:

$$TF = \frac{\text{Number of time terms T appears in a document}}{\text{total number of terms in the document}} \cdots (1).$$

TF value is calculated for all words in document, and calculates inverse document frequency as following:

$$IDF = \log_e \frac{\text{total No.of document}}{\text{No.of document with term T appear in it}} \cdots (2)$$

For all terms, the proposal system was search (low TF-IDF value) and removed it by using standard deviation as threshold to remove the words that have TF-IDF value less than threshold by find the max value of TF-IDF values which was appeared (0.6154) from datasets, and find the minimum value of TF-IDF that appeared (0.0039) then the proposed system calculated (min TF-IDF value / TF-IDF value) to find threshold (0.0063) then used standard deviation for this threshold , ST for TF-IDF value less than 0.0063 is 0.0382, the proposed system removed all the word that have TF-IDF value less than ST value 0.0382 from datasets, the proposed system got fewer words 1516 which represented important word after the number of word are 13195 words previously found .thus in this method contracted from storage space in memory and the processing time become faster to build the final TF-IDF matrix. In addition to use the following algorithm of stop words enhancement algorithm in the proposal system

```
Algorithm2: Stop words removal algorithm
Input: Set of documents (tokenization) and Stop-word list.
Output: Word vector for each document without stop words.
Step1: while ID- document ≤ 925 do
Get new document from datasets; parse and tokenization the document in to set of tokens by using tokenization algorithm.
Step2: Convert all tokens from uppercase letters to lowercase letters.
Step3: check each token by:-
1.      Remove any non-alpha letters and call (Remain word) to the remaining letters.
2.      Let P-L is length of (Remain word) , if P-L ≥ 2  and matching the word in stop words list then (Remain word)is a stop words remove the word , move to the next word in the document and go to step3.
3.      Some word that starts with {“, +, _} thus if the first letters of word is in the set [“, +, _ ] and P-L >= 3 then remove the first letters from the beginning of the word and go to step3.
4.      Some word end with { “,  +, _, :, !,!!, „, •, ?, ??,} thus if the last characters is in the set [“,  +, _, :, !,!!, „, •, ?, ??] Then remove the end letters and go to step3.
5.      If the document is null go to step1.
End.
```

Step 4. Stemming Algorithms

The porter stemming algorithm with enhancement on its rules was used in the proposed system, at each step, a certain suffix is deleted by using the set of rules. These rules are substitution rule which is applied when a set of conditions match to this rule thus to reduce number of words, to have exactly matching stems, and to save memory space and time. The proposed system was used the porters algorithm and look up table approach by having two dictionaries, one for various irregular English words, and another for various suffixes. To applied the following:

Root = past simple or past participle.

Suffixed = root + suffix.

The algorithms below and flowchart in figure (3) showed the stemming function in the proposal system.

Algorithm3: porter stemming enhancement algorithm
Input: Word vector for all documents without stop words
Output: Stemmed word vector

Step 1:-the suffixes (ed , ing, and plurals S) are removes from the end words.
Step 2: the ending of words that contain y replace y to I when there is additional vowel in the word.
Step 3: replace dual suffixes to only ones: -ization, -ational, etc.
Step 4: the suffixes, -full, -ness etc are removed from the end of the words.
Step 5: off Incomes -ant, -ence, etc.
Step 6: Eradicates a last –e
Step7: Eradicate one of dual *b, d, g, m, n, p, r, s, t.*
Step 8: Try deadly *d, r, t, z* into *s.*
End.

4. Results and Experiments:

The proposed system usage the Reuters 21578 datasets for proposed preprocessing steps as tests with number of documents selected are 925 documents. Table 2 shows the setting for the proposed system experiment, while the whole number of tokens made in all effort documents after treating are (13195). Lacking tokenization treating to huge number of tokens, and take a lengthy time in complete tokenization procedure which is right relative to performance measure of an information retrieval system, as it acutely moves the indexing and storing features.

Phase 1: Extraction the Documents from Reuters 21578 datasets as table 2, input the 925 documents to the tokenization procedure

Table 2: Data set Extraction

Do.Id	Document contents
1	Showers continued throughout the week in behin coca zone alleviating drought since early January improving prospects coming tempora
2	Standard oil co and bp north America said they plan form venture manage borrowing investment activities both companies north
3	Texa commerce Bancshares incs texa commerce bank Houston said filed application with comptroller currency
4	Bankamerice corp is not under pressure quickly proposed equity offering would well delay because stocks recent poor
5	The u.s. agriculture department reported farmer-owned reserve national five day average price through February follows dlrs/ bu
6	Argentine grain board figures show crop registrations grains oilseeds their products February thousands tonnes showing
7	Lion inns limited partnership said filed registraction statement with securities exchange commission covering proposed

Phase 2:

Now in phase2, all the input documents are mined to extract the pool of words displayed below:

ID: doc1
[Showers, continued, throughout, the, week, in, ………………… .etc.]

ID: doc2
[standard, north, america, they, said, plan, form, venture, manage…………etc]

ID: doc3
[texas, commerce, Bancshares, incs, texas, commerce, bank, -houston, …… etc]

.
.
.

ID: doc6
[argentin, grain, board, figure, show, crop, register, grain, oilse, their, product, februari, thousand, ton, show, those, …………… , future, shipment, month]

Phase 3 besides Phase 4:

Next mining all the words output, in this phase the proposed system eliminates all stop words, and stemming algorithm was practical, as presented below:

Doc1: shower continue throughout week behia coca zone allevi drought sinc earli januari improve prospect ….

Doc2: standard north America said they plan from ventur manage money market ……

Doc3: texa commerce bancshare incs texa commerc bank Houston said file applic with comptrol ….

Doc4: bankamerica corp under pressur quicki propos equili offer would well delai because…..……………………..

Phase 5: finally the proposal system calculates TF-IDF value for each term in datasets, a small example from huge TF-IDF matrix shown in table (3)

Table 3: Sample of TF-IDF value

Term	TF value	IDF value	TF-IDF value
week	0.0108	4.3027	0.0464
behia	0.0144	8.0163	0.1153
cocoa	0.0216	7.3232	0.1581
come	0.0072	6.9177	0.0498
tempora	0.0072	8.0163	0.0577
have	0.0072	3.7536	0.0270
commissari	0.0180	8.0163	0.1442
said	0.0180	1.7174	0.0309
Period	0.0072	5.9369	0.0427
year	0.0072	2.9226	0.0210
arrive	0.0072	8.0163	0.0577
februari	0.0108	4.8383	0.0522
bag	0.0180	6.9177	0.1244
kilo	0.0072	6.9177	0.0498
total	0.0108	4.7582	0.0513

against	0.0108	5.1831	0.0559
consign	0.0072	8.0163	0.0577
still	0.0108	6.4069	0.0691
crop	0.0180	6.6300	0.1192
export	0.0072	4.3528	0.0313
dlr	0.0504	2.5191	0.1269
port	0.0108	6.4069	0.0691
open	0.0072	6.2246	0.0448
north	0.0476	6.6300	0.3157

5. Experimental Results

The proposed system usage the Reuters 21578 datasets for preprocessing step, which was tested with number of documents selected form datasets are 925 documents. Table 4 shows the setting for the proposed system experiment.

Table 4: Setting for Experiment

Effective preprocessing parameters	Number of documents	Set Randomly
	Tokenization	Set parsing documents
	Stop words removal	Set list of stop words
	Stemming	Set enhancement porter algorithm
	Word vectors	Create TF-IDF matrix

6. Conclusions

The proposed system introduces an enhancement to the pre-processing information retrieval system, this step affects the outcomes of any IR system. The lack of standard porter stemming algorithm and preprocessing steps such as, stop-word removal and stemming also motivates us to bring out these instruments.

The proposed system GUI has many options including reading dataset files, display output in tables, and produce statistics about preprocessing steps. it is careful as a chief step through a standard English language preprocessing systems. The proposed method is tested in two levels, first level uses only vector space model which based on used traditional stop words removal and with traditional porter stemming and the second level uses vector space model with combined features of improved stop words removal algorithm and improved stemming algorithm.

The results show that using second level as preprocessing step for text mining application achieves good performance with an average categorization accuracy of 90%.

In the future research, the proposed method can improve the performance of the text mining application in the field of another datasets from other aspects.

References

[1] Vairaprakash Gurusamy and Subbu kannan, “Preprocessing techniques for text mining”, https://www.researchgate.net/publication/273127322, 2014.

[2] Vikram Singh and Balwinder Saini, “An effective pre-processing algorithm for information retrieval systems”, International Journal of Database Management Systems (IJDMS) Vol.6, No.6, 2014.

[3] Dasa Munkova, Michal Munk, and Martin vozar,” Data Pre-Processing Evaluation for Text Mining: Transaction/Sequence Model”, International Conference on Computational Science 2013.

[4] Anjali R. Deshpande, and Lobo L. M. R. J, “Text Summarization using Clustering Technique”, International Journal of Engineering Trends and Technology (IJETT) - Volume4 Issue8- 2013, 2013.

[5] Vijayarani S., Ilamathi J., and Nithya, “Preprocessing Techniques for Text Mining - An Overview”, International Journal of Computer Science and Communication Networks, vol 5(1), 7-16.

[6] Vikram Singh and Balwinder Saini, “An Effective tokenization algorithm for information retrieval systems”, Department of Computer Engineering, National Institute of Technology, Kurukshetra, Haryana, India David C. Wyld et al. (Eds) : COSIT, DMIN, SIGL, CYBI, NMCT, AIAPP - 2014.

[7] Brajendra Singh Rajput1, and Nilay Khare, “A survey of Stemming Algorithms for Information Retrieval”, IOSR Journal of Computer Engineering (IOSR-JCE) Volume 17, Issue 3, 2015.

[8] Fadi Yamout, Rana Demachkieh, Ghalia Hamdan, and Reem Sabra, “Further Enhancement to the Porter’s Stemming Algorithm”, Faculty of Computer Sciences, C and E American University I., Beirut, Lebanon, Email: fyamout@inco.com.lb.

[9] C.Ramasubramanian, R.Ramya, Virudhunagar, and Tamilnadu, “Effective Pre-Processing Activities in Text Mining using Improved Porter’s Stemming Algorithm”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 12, December 2013.

[10] Nikita P.Katariya1, M. S. Chaudhary, and Nikita P.Katariya, “Text preprocessing for text mining using side information”, International Journal of Computer Science and Mobile Applications, Vol.3 Issue. 2015.