International Journal of
Science and Engineering Investigations
vol. 6, issue 61, February 2017
ISSN: 2251-8843

Rule-Based Filtering Algorithm for Textual Document

Nurul Syafidah Jamil[1], Ku Ruhana Ku-Mahamud[2], Aniza Mohamed Din[3]
[1,2,3]School of Computing, College of Arts and Sciences, Universiti Utara Malaysia
([1]jamil.nurulsyafidah@gmail.com, [2]ruhana@uum.edu.my, [3]anizamd@uum.edu.my)

Abstract- Textual document is usually in unstructured form and high dimensional data. The exploration of hidden information from the unstructured text is useful to find interesting patterns and valuable knowledge. However, not all terms in the text are relevant and can lead to misclassification. Improper filtration might cause terms that have similar meaning to be removed. Thus, to reduce the high-dimensionality of text, this study proposed a filtering algorithm that is able to filter the important terms from the pre-processed text and applied term weighting scheme to solve synonym problem which will help the selection of relevant term. The proposed filtering algorithm utilizes a keyword library that contained special terms which is developed to ensure that important terms are not eliminated during filtration process. The performance of the proposed filtering algorithm is compared with rough set attribute reduction (RSAR) and information retrieval (IR) approaches. From the experiment, the proposed filtering algorithm has outperformed both RSAR and IR in terms of extracted relevant terms.

Keywords- *Topic Identification, Filtering Algorithm, Synonym, Textual Document*

I. INTRODUCTION

The trend to store data in electronic format has increased and requires an efficient effort to organize these important yet beneficial documents [1]. Data can be stored in several formats such as image, video and audio. However, text is commonly used to store knowledge and exchange information [1] [2]. Text is usually unstructured and not restricted to any specific format [2] [3]. The exploration of hidden information from this unstructured text is useful because interesting patterns and valuable knowledge can be discovered from the text [1]. Finding relevant terms from the text is a challenging task as human ability to analyze, gather, understand and store massive datasets is slow and rather costly to process all the information. In fact, analyzing texts is a tedious task prior to the curse of dimensionality in every text document, in which the complexity of natural language, misspelling, mispunctuation and misinterpretation that leads to misunderstanding to occur [1] [2].

The main issues in term filtering are to extract relevant terms from high dimensional data and to solve synonym problems of the extracted terms [4]. Terms in a text may be irrelevant, or other words, it might have no effect on the processing which could give impact on the processing performance [5] Terms with high occurrences in the text shows the importance of the terms itself. However, not all terms with high occurrences numbers are relevant [5]. Hence, retrieving the most relevant term is necessary. Retrieving relevant terms from text is a non-trivial task because of the high dimensionality nature in text and lack of formal structure in text document [6]. In other words, getting a handle on what is important and what is not important from textual data is not easy. Due to the complexity of the text, dimensionality reduction technique is needed to reduce the irrelevant terms in the text and make them easier to handle [7].

Feature selection is a process to retrieve the quality of features since not all features are relevant for classifying the text [8]. Feature selection is commonly used in text classification to reduce the dimensionality of features and improve the efficiency and accuracy of classifiers. Several methods have been widely applied for feature selection in text classification such as information gain (IG), mutual information feature selection (MIFS) [9], document frequency-based selection, term strength, and entropy-based ranking, but these statistical-based methods only provide information access instead of analyzing information to locate patterns. This method aims at selecting some of features or words that have the highest score according to the predetermined measure of the importance of the word [9]. The selection process is performed by applying either the filter approach or the wrapper approach. The filtering approach is employed most of the time for feature selection stage [12]. The approach is based on applying a scoring method to evaluate the features. The filtering approach is based from the document frequency in finding and retaining the terms that occur in the highest number of documents. The commonly used filter methods are such as document frequency, mutual information, information gain, chi-square, and Gini index [10].

The advantages of filter approach are that it is easily scaled to high dimensional datasets, computationally simple and fast. In addition, the filter approach is independent as it only has to be performed only once [11] However, according to [12], the size of the feature space is not reduced by implementing methods in feature selection since the size of the full feature set is reduced and time consuming. In view of this, the drawback of filter approach is that it tends to ignore the effect of the selected feature set on the classifier algorithm. To add to this matter, this is also supported by [14] who claimed that most of the filter approaches tend to ignore the interaction with the classifier and since the techniques are univariate, this means that each feature is considered separately.

The wrapper approach wraps the features around the classifiers to be used to anticipate the benefit of adding or removing a certain feature from the training set [12]. Unlike filter approach, wrapper selects the features that lead to an improvement in the performance of the classifier algorithm. The quality of an attribute subset is directly measured by the performance of the data mining algorithm that is applied to that attribute subset [11]. This means, the wrapper approach might be much slower than the filter approach because data mining algorithm is applied to each attribute subset considered by the search. In some cases, several data mining algorithms need to be applied to the data which makes the wrapper approach become computationally expensive. The wrapper approaches is able to include the interaction between feature subset search and model selection and considers the feature dependencies. However, is risky due to the overfitting issue [11].

Another issue in term filtering is the synonymy of the terms. Synonym is natural linguistic phenomena which Computational Linguistic and Information Retrieval researchers commonly find difficult to cope with [4]. Several attempts which are based on semantic to solve synonymy problem by using WordNet [13], computed semantic relatedness from Wikipedia [14] and extracting word similarity from website [15]. Unfortunately, WordNet is not effective as it did not include any specific terms that are needed [13]. Contents in Wikipedia are not consistent because some articles are missing and anyone can change the content [16]. Different websites might use different word similarity and this can cause inconsistency in solving synonymy problem. Another different solution has been implemented which is to use rule-based term extraction as a controlled environment for synonym distribution such as in [17]. Therefore, a rule-based filtering algorithm is proposed and a keywords library is also developed to solve synonymy of the extracted terms.

Some of the information extraction systems rely on nouns as their features for entity identification such as in [18]. Noun carries a role to interpret the structure of sentence, mostly in particular, fields such as machine translation, text retrieval, information extraction and text classification [19] Nouns in text may portray the topic that has been discussed [20]. In addition, classifying texts that are based on a single linguistic expression can be effective since nouns can represent specific incidents and general events in the sentence and likely produce good topics in the sentence [21]. By this means, mastering noun leads to understanding the main meaning denoted in the text.

To rank the extracted terms, term frequency-inverse document frequency technique or also known as TF-IDF is used to evaluate the relevancy of a term in the document [22] [23]. But, a term that has the highest score has the potential to give vague information to identify one specific topic [22] [23]. In fact, TF-IDF cannot solve synonymy problems because it ignores the relationship between words [24].This matter can be resolved by solving synonymy problem during text pre-processing phase.

To overcome the aforementioned issues, statistical based approach filtering algorithm combined with the computational linguistics technique is proposed. The nature form of unstructured textual data is the challenge to ensure that the selected relevant terms from the dataset. This study focuses on noun as a term candidate; thus Part-of-Speech tagging technique is implemented for identifying the syntactical parts in the text document.

This paper is organized as follows; Section II explains on the text pre-processing processes that will be implemented in this study. Section III presents the proposed filtering algorithm which is based on rule-based approach. Section IV discusses the results of pre-processed text and the performance of the proposed filtering algorithm.

II. TEXT PRE-PROCESSING

Text pre-processing is the initial step in any text mining task such as text classification, text clustering, text summarization and topic identification. Text pre-processing is mainly performed to remove features that are unimportant for topic identification purposes and present text documents into a clear word format. Fig. 1 shows the processes in text pre-processing.

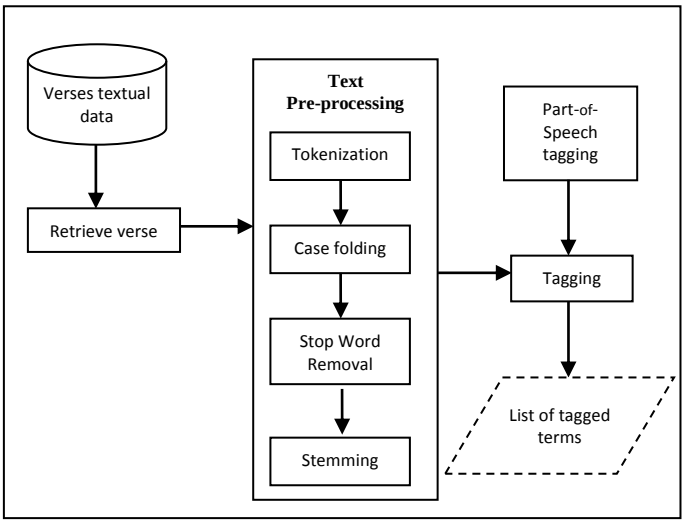

Figure 1. Text pre-processing process [1]

The first stage in text pre-processing is tokenization where text is broken up into smaller and meaningful units known as tokens before any language processing can be performed. Second, all terms in the sentences are converted to lowercase by using case folding step. Case folding is used to avoid the same term to be counted as a different meaning. For instance, term ‘Mother’ and ‘mother’, which carries the same meaning but the first term contains a capital letter and the other term is in lowercase. Hence, case folding is needed in order to solve this problem. Third stage is to eliminate the noise words in the selected text by using stop word removal. The stop words such as ‘the’, ‘a’, ‘and’ frequently occur and they are assumed as the insignificant words needed to be removed because it is not useful for classification. Next stage is stemming which is to convert different word forms into similar canonical forms, or in other words, it is the process of conflating tokens to their root form.

Once the root word is achieved, the next stage is to tag each of the term into specific values using part-of-speech tagging. POS tagging is the process of assigning grammatical value for each word in the sentence. The purpose of using POS tagging is to identify the potential terms from the text, especially nouns. The expected outcome from this stage is the cleaned texts that are free from any noise. Examples of POS tags are shown in Table 1.

TABLE I. SAMPLE OF PART-OF-SPEECH TAG SET

Tag	Set
AV0 – General adverb	carefully, lovely, intentionally
AVQ – *wh*-adverb	when, why, how, wherever
NN – Common noun	man, woman, girl, mother, town
NP – Proper noun	Yusuf, Mariam, Miss, Mother

III. THE PROPOSED FILTERING ALGORITHM

Fig. 2 shows the proposed filtering algorithm (PFA) where the tagged terms from pre-processing stages are filtered. From text pre-processing stage, a list of tagged terms is produced and the next stage is to collect only terms that have been tagged as 'NN'. The PFA aims to overcome high dimensionality problem of text by removing unimportant terms in the text. In addition, PFA allows checking the availability of the matching keywords in order to find the most precise terms. Though only noun terms are taken, there are several exclusive noun terms that have been determined as keywords, such as 'wed' (verb), 'marry' (verb) and 'will' (future tense). Some of single terms carry similar meaning and eliminating these terms should be avoided. Therefore, the filtering algorithm is designed in term extraction phase to solve the synonymy of the extracted terms and these important terms are not eliminated during the extraction process. Hence, filtration process is executed to avoid the occurrence of missing important terms.

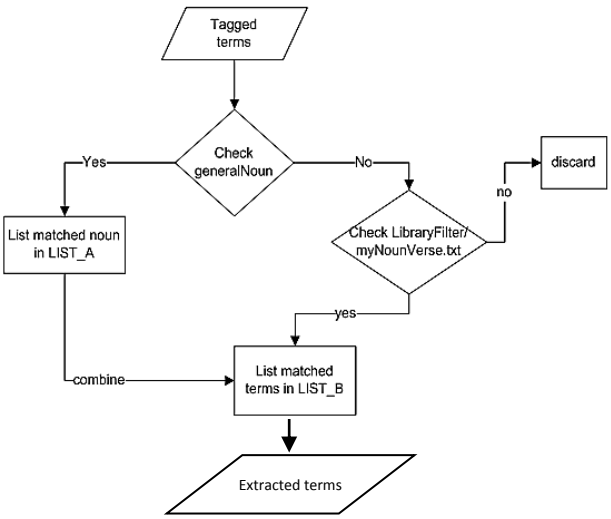

Figure 2. The proposed filtering algorithm

However, in the context of this study there are several important terms that belong in verb class for example 'marry' and 'wed'. Besides, term 'will' is a very strong keyword because in some context of the Quran verses, 'will' is referring to an act to give (property) to another person after one's death. Meanwhile, term 'will' also belongs to the future tense class and considered as noise word by stop word removal. Therefore, two keyword libraries are involved in the filtering process. The first keyword library is known as 'generalNoun' which contains all nouns and retrieved from the open source on the net. The second keyword library is myNounVerse which contained all nouns or any keywords from the Quran such as 'Zihar', 'Iddat' 'Will' and 'Wed'. This library is manually developed by comparing the synonym of certain terms by using Thesaurus.com. Without term filtration phase, it is costly for the next phase to rank relevant terms according to its importance. Too many irrelevant terms can affect the accuracy of ranking results and determination of topic.

This algorithm begins right after the process of assigning POS tagging is finished. In the first filtration, all tagged terms are compared and checked with generalNoun library in order to ensure that all terms are not mistagged. Matched terms which are nouns are listed in List A. The second filtration filters and removes those terms that do not belong to myNounVerse and not tagged as nouns. Matched terms from this process are listed in List B and combined with List A. All umatched terms are discarded. Only matched nouns are listed for further process.

IV. RESULTS AND DISCUSSION

English translated Quran retrieved from Surah.my (http://www.surah.my/) and has been used as a case study. Texts from Surah.my website are chosen since traffic report from bizinformation.com.my shows that the source is frequently accessed and most referable in Malaysia. The experiment to test the performance of PFA has been conducted using accuracy metrics. Table 2 shows the sample of Verse 2_35 that has been pre-processed and filtered. The filtered terms consists of nouns and some important terms from library myNounVerse. For example, each extracted terms from Verse 2_35 occurred only one time, however there are two keyword terms which are 'wife' and 'will'. These two keywords refer to two different topics. Term 'wife' indicates the topic as 'Marriage, meanwhile term 'will' refers topic as 'Inheritance'. In this case, TF-IDF technique has been used to calculate the importance score of the extracted terms.

TABLE II. SAMPLE OF PRE-PROCESSED AND FILTERED TERMS FROM THE VERSES

Verse No	Verse	Filtered terms
2_35	We said: "O Adam! dwell thou and thy wife in the Garden; and eat of the bountiful things therein as (where and when) ye will; but approach not this tree, or ye run into harm and transgression."	wife, garden, thing, will, approach, tree, harm, transgression

The filtered terms produced by PFA are then compared with rough set attribute reduction (RSAR) and information retrieval (IR). RSAR is conducted through Rosetta application for term filtration by selecting the relevant attributes (terms) out of the larger set of candidate attributes. The relevant attributes are defined as attribute subset that has the same classification capability with the overall attributes. RSAR reduces the dimensionality of the data and enables the learning algorithm to operate effectively. The extraction is based on the calculation of reduct values that have been produced in Rosetta. For the filtration based on information retrieval, the interpretation and document ranking are according to the relevancy and the importance of the documents itself. Information retrieval does not consider the linguistic criteria unless it is being assigned to.

Therefore, POS tagging has not been applied during for the experiment. PFA is experimented with RSAR because RSAR is one of the technique in dimensionality reduction and able to reduce the number of features from information system. PFA also has been compared with Information Retrieval because it does not employed POS tagging in most term extraction works. This is because IR focuses only on the terms with the highest score. Meanwhile, PFA chooses noun instead of other regular expressions such as verbs and articles. The comparison that has been made for Verse 2_35 is shown in Table 3.

TABLE III. SAMPLE OF COMPARISON FOR THE FILTERED TERMS

Verse No	Filtered terms					
	PFA		RSAR		IR	
	Terms	#terms	Terms	#terms	Terms	#terms
2_35	wife, garden, thing, will, approach, tree, harm, transgression	8	wife, will	2	say, wife, in, the, garden, eat, bountiful, things, approach, tree, run, harm, transgression	13

As shown in Table 3, IR has the most numbers of filtered terms as compared to the other techniques. However, filtered terms from RSAR is too limited. PFA consistently produced enough numbers of filtered terms. From the experiments, it can be seen that the relevant terms are successfully extracted by the employment of PFA. This is because PFA has its own keyword libraries that store the important terms in it. Besides that, PFA ensures that the terms that belong to the keyword library are not eliminated during the filtering phase. These comparisons are applied to all 224 English translated Quran verses. These filtered terms are then calculated and ranked using TF-IDF technique. No technique comparison has been made in ranking phase as it is only been adopted in the method. The calculation is shown in Table 4.

TABLE IV. THE TF-IDF CALCULATION FOR EACH TERM IN VERSE 2_35

Verse (d)	Total terms in d	Terms (t)	#t occurs in d	TF	IDF	TF-IDF
2_35	8	will	1	0.125	0.6600	0.0825
		wife	1	0.125	0.5790	0.0724
		garden	1	0.125	0.1808	0.0226
		thing	1	0.125	0.0560	0.0070
		approach	1	0.125	0.2938	0.0367
		tree	1	0.125	1.1751	0.1469
		harm	1	0.125	0.4700	0.0588
		transgression	1	0.125	1.1751	0.1469

There are 37 terms in Verse 2_35 and represented by (d). However, only 8 terms are filtered and there are ‘wife, ‘man’, ‘garden, ‘thing, ‘will’, ‘approach’, ‘tree’, ‘harm’ and ‘transgression’. Each term is labeled as (t) and number of occurrences for each term is counted. Amongst these filtered terms, ‘wife’ and ‘will’ are matched from myNounVerse library. At this stage, each (t) must be calculated using TF-IDF formula. For example, to calculate the TF for term ‘wife’, the total number ‘wife’ occurs in the document is 1 and divided by the total number of all terms in the document which is 8. The TF score is 0.125. In this case, both terms ‘wife’ and ‘will’ have same TF score number. Therefore, the calculation to find IDF score has been performed. To calculate the IDF, there are 224 documents and the term ‘wife’ occurs 59 times in these documents. Then, the IDF is calculated as *log* (224/59) and equivalent to 0.5790.

Next, the value for TFIDF is computed as 0.125 multiply with 0.5790 and is equivalent to 0.0724. This calculation process is applied to the whole 224 documents. There is no threshold to determine the value of score for the relevant terms. The term with highest score is considered as the most relevant term. As can be seen, the most relevant terms in the verse are top-ranked based on its score. For example, ‘wife’ and ‘will’ have same occurrences number. Which means, both terms have same TF score number. Through the calculation of IDF and TF-IDF, the final ranking scores are produced. Therefore, the potential term in Verse 2_35 is ‘will’ as it has higher TF-IDF score number as compared to term ‘wife’. The same calculation is conducted to each term in all 224 verses.

V. CONCLUSION

The proposed filtering algorithm for textual document has reduces the high dimensionality of the text and is able to filter the most important terms. The effectiveness of the retrieval results by the proposed filtering algorithm can be judged by the number of relevant terms retrieved from the verses. The proposed filtering algorithm has proven that the extracted terms contribute to the performance of the proposed rule generation algorithm to identify topics which are closer to the topics by the experts. The inclusion of noun selection has enabled the filtering process to produce the most relevant terms. The PFA has also able to solve high dimensionality data and synonym problems in textual document.

ACKNOWLEDGMENT

This research was supported by Universiti Utara Malaysia under PBIT grant [12311 (2012)].

REFERENCES

[1] K. Sumathy and M. Chidambaram, "Text Mining: Concepts, Applications, Tools and Issues–An Overview," *International Journal of Computer Applications*, vol. 80, no. 4, pp. 29–32, 2013.

[2] S. Jusoh and H. M. Alfawareh, "Techniques, Applications and Challenging Issue in Text Mining," *International Journal of Computer Science Issues*, vol. 9, no. 6, pp. 431–436, 2012.

[3] S.S. Kamaruddin, "*Framework for deviation detection in text."* Universiti Kebangsaan Malaysia, Bangi. 2011.

[4] J. I. Sheeba, and K. Vivekanandan, K. "Improved Unsupervised Framework for solving Synonym, Homonym, Hyponymy & Polysemy Problems from Extracted Keywords and Identify topics in Meeting Transcripts." *International Journal of Computer Science, Engineering and Applications*, *2*(5), 85. 2012.

[5] J. Ventura, and J.F. da Silva. "*Ranking and extraction of relevant single words in text*." INTECH Open Access Publisher. 2008.

[6] H. S. Baghdadi and B. Ranaivo-Malançon, "An Automatic Topic Identification Algorithm," *Journal of Computer Science*, vol. 7, no. 9, pp. 1363–1367, 2011.

[7] A. Khan, B. Baharudin, L.H. Lee, and K. Khan, "A review of machine learning algorithms for text-documents classification." *Journal of Advances in Information Technology*. *1*(1). 2010.

[8] C. H. Bong, and T.K. Wong, "An examination of feature selection frameworks in text categorization. Information Retrieval Technology." 3689. 2005.

[9] J. Bakus, and M.S. Kamel, "Higher order feature selection for text classification." Knowledge Information System. 9, 4. 468-491. 2006.

[10] A. K. Uysal, S. Gunal, S. Ergin, and E. Sora Gunal, "The Impact of Feature Extraction and Selection on SMS Spam Filtering". *Elektronika ir Elektrotechnika*, *19*(5), 67-72. 2012.

[11] S. Beniwal, and J. Arora, "Classification and feature selection techniques in data mining." In *International Journal of Engineering Research and Technology* (Vol. 1, No. 6 (August-2012)). ESRSA Publications. 2012, August.

[12] A. T. Sadiq, and S.M. Abdullah, "Hybrid Intelligent Technique for Text Categorization." In *Advanced Computer Science Applications and Technologies (ACSAT), 2012 International Conference on* (pp. 238-245). IEEE. (2012, November)

[13] J. McCrae, E. Montiel-Ponsoda, and P. Cimiano, "Integrating WordNet and Wiktionary with lemon." In *Linked Data in Linguistics* (pp. 25-34). Springer Berlin Heidelberg. 2012.

[14] E. Gabrilovich, and S. Markovitch, "Computing semantic relatedness using Wikipedia-based explicit semantic analysis." In *IJcAI* (Vol. 7, pp. 1606-1611). 2007, January.

[15] X. H. Phan, C.T. Nguyen, D.T. Le, L.M. Nguyen, S. Horiguchi, and Q.T. Ha, "A hidden topic-based framework toward building applications with short web documents." *IEEE Transactions on Knowledge and Data Engineering*, *23*(7), 961-976. 2011.

[16] M. Hassan, "*Automatic Document Topic Identification Using Hierarchical Ontology Extracted from Human Background Knowledge"* (Doctoral dissertation, University of Waterloo). 2013

[17] T. Wang, and G. Hirst, "Exploring patterns in dictionary definitions for synonym extraction." *Natural Language Engineering*, *18*(03), 313-342. 2012.

[18] S. Berkowitz, "*U.S. Patent No. 7,805,291*." Washington, DC: U.S. Patent and Trademark Office. 2010.

[19] B. M. Sagar, G. Shobha, and R. Kumar, Solving the Noun Phrase and Verb Phrase Agreement in Kannada Sentences. *International Journal of Computer Theory and Engineering*, *1*(3). 2009.

[20] G. Protaziuk, M. Kryszkiewicz, H. Rybinski, and A. Delteil, "Discovering compound and proper nouns." In *Rough Sets and Intelligent Systems Paradigms* (pp. 505-515). Springer Berlin Heidelberg. 2007.

[21] R. Dong, M. Schaal, M.P. O'Mahony, and B. Smyth, "Topic extraction from online reviews for classification and recommendation." *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*. 2013.

[22] T. P. Hong, C.W. Lin, K. T. Yang, and S. L. Wang, "Using TF-IDF to hide sensitive item sets." *Applied Intelligence*, 1-9. 2013.

[23] W. Zhang, T. Yoshida, and X. Tang, "A comparative study of TF*IDF, LSI and multi-words for text classification." *Expert Systems with Applications*, *38*(3), 2758-2765. 2011.

[24] J. Ramos, "Using tf-idf to determine word relevance in document queries." In *Proceedings of the first instructional conference on machine learning*. 2003.