Topic Identification Method For Textual Document

Nurul Syafidah Jamil
School of Computing
College of Arts and Sciences
Universiti Utara Malaysia,
Malaysia
jamil.nurulsyafidah@yahoo.com

Ku Ruhana Ku-Mahamud
School of Computing
College of Arts and Sciences
Universiti Utara Malaysia,
Malaysia
ruhana@uum.edu.my

Aniza Mohamed Din
School of Computing
College of Arts and Sciences
Universiti Utara Malaysia,
Malaysia
anizamd@uum.edu.my

***Abstract*— Topic identification is a crucial task for discovering knowledge from textual document. Existing methods for topic identification suffer from word counting problem as they depend on the most frequent terms in the text to produce the topic keyword. Not all frequent terms are relevant. This paper proposes a topic identification method that filters the important terms from the pre-processed text and applied term weighting scheme to solve synonym problem. A rule generation algorithm is used to determine the appropriate topics based on the weighted terms. The text document used in the experiment is the English translated Quran. The topics identified from the proposed method were compared with topics identified using Rough Set and domain experts. From the findings, the proposed topic identification method was consistently able to identify topics that are mostly close to the topics that have been given by Rough Set and the experts. The result from the comparison proved that the proposed method was able to be used to capture topics for textual documents.**

Keywords—topic identification; filtering algorithm; rule generation; text document

I. INTRODUCTION

The current growth of computing facilities has led the large amount of information to be stored in digital form. Information can be stored in various formats but text is often used to store information and conveying knowledge as well. Text is the natural form to store information and knowledge [1] [2] in documents such as in Electronic Publications, Digital Libraries, E-Mails and World Wide Web [3]. Textual data can be grouped into several forms such as unstructured textual data, structured textual data and semi structured textual data [4]. However, the nature of text is described as unstructured information because there is no specific structure and format for it such as spelling, chapters, sections and paragraphs. The unorganized yet useful information in text needs to be categorized to its class.

To classify text document with appropriate topics, the assurance to extract relevant features is needed [5]. Hence, topic identification method must be able to extract useful data to represent topic of text document without losing important information. By recognizing the correct keywords from text will guide user to capture the whole meaning of text. Topic identification method is a classification problem where the task is the assignment of the correct topic [6] [7]. In other words, topic identification is also known as topic spotting or topic detection and tracking, because it can automatically sort a set of documents into categories or classes or topics from a predefined set [6].

Topic identification method also has been used for interesting works such as identifying topic from social status streams [8] and identifying topics of learning material [9]. The term 'topic' refers to a concept or a subject that can be used to categorize a document. It indicates a certain subject which is discussed in the whole text [9] and usually represented by a single term. A topic usually assists people to search and understand the whole sentence in a text [7] [9] and can be determined by looking at the accuracy of certain words from the sentence [10].

Rule-based topic identification method performs rule-based classification technique such as Rough Set decision rule and other rule-based algorithm. Devasena and Hemalatha (2012) developed a rule reduction algorithm to create token, identify feature to summarize the text and identify the topic [11]. This work focuses on the semantic value in the sentences in order to summarize the whole text. In another work, Massey and Wong (2011), proposed a rule-based algorithm for topic identification which uses single terms from text and single terms extracted from Yahoo web page to determine the topics [12]. However, the proposed method is purely based on statistical approach which did not employ any linguistic techniques such as name entity recognition and tagging.

In a different case, Fuddoly, Jaafar and Zamin (2013) used clustering method such as Bracewell's algorithm to find the similarity of keywords in order to identify topic for Indonesian news documents [13]. Baghdadi and Ranaivo-Malancon (2011) also proposed a method to discover topic employed based on clustering algorithm [7]. The study exploited Chen's algorithm to calculate the IDF which is a weight for each noun and verb identify topic and modified the algorithm by selecting the topic with the highest weight. Anaya-Sanchez, Pons-Porrata and Berlanga-Llavori (2008) proposed an algorithm to obtain label document cluster to identify topic of text collection [14]. Next, an algorithm was proposed by Stoyanov and Cardie (2008) for topic identification of fine-grained opinion analysis. TF-IDF weighting scheme is implemented to count the frequency of topics in texts [15]. These studies exploited POS tagging technique to find the topic candidates syntactic parts.

There are three approaches for topic identification which are statistical, ontological and rule-based. Based from the review, rule-based topic identification is able to use set of rules in algorithm to make decision. As statistical approach is too robust and ontological approach is computationally expensive, rule-based topic identification is chosen as a suitable approach for the topic identification method proposed in this paper. In Section 2, the explanation on the proposed topic identification method is presented. Results from experiments are described in Section 3. Finally, the conclusion is presented in Section 4.

II. THE PROPOSED TOPIC IDENTIFICATION METHOD

The textual dataset used for this experiment is the English Translated Quran retrieved from Surah.my website (http://www.surah.my). The website is selected due to the website traffic result that it is the most referable in Malaysia There are 224 verses out of 6666 verses used for the experiment. Only verses that contained keywords such as daughter, female, woman, damsel, niece and mother are taken. Generally, there are various topics that can be classified from the Quran such as punishment, History, Rewards and Afterlife. However, three target topic classes have been identified which are related to female from the Holy Quran such as Inheritance, Marriage and Divorce. Hence, this proposed method will identify topic from the textual data which relates to those mentioned target topics. The proposed topic identification method is shown in Fig. 1.

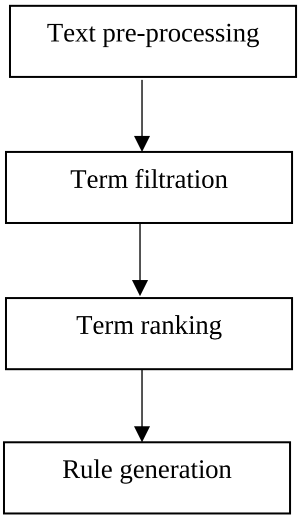

Fig. 1. *The proposed topic identification method*

The method first focuses on text pre-processing, then term filtration is performed. A filtering algorithm is used to filter out the unnecessary features from the text and ensures that only noun is selected. Noun has been used to interpret the structure of sentence and portray the topic that has been discussed by Sagar, Shobha and Kumar (2009) [16]. Classifying texts that are based on nouns can be effective because nouns can represent specific incidents and general events in the sentence and likely produce good topics in the sentence [17]. Once the terms have been filtered, term ranking is performed where the filtered terms are measured by looking at the occurrences number of terms in the text.

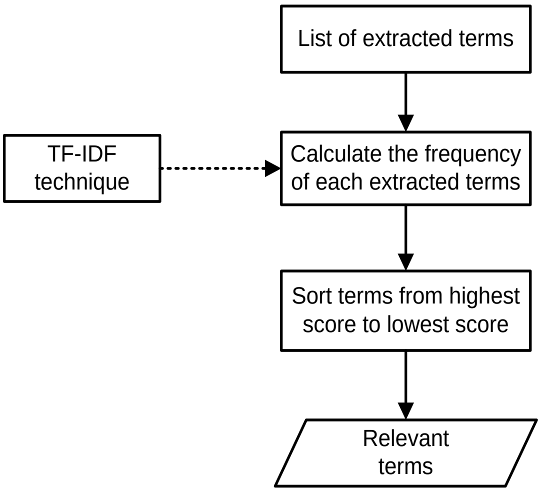
List of extracted terms
TF-IDF technique
Calculate the frequency of each extracted terms
Sort terms from highest score to lowest score
Relevant terms

Fig. 2. *Term ranking*

Fig. 2 illustrates the process of term ranking. In term ranking, the frequency of the extracted terms is calculated and this determines which term has the potential to become relevant terms in constructing decision rules for topic identification. Each term from the list of extracted terms is weighted by calculating the frequency to capture the number of appearances in the text. The weight of a term can express the importance for a particular document. Term frequency-Inverse Diverse Frequency (TF-IDF) technique is selected in term ranking since it can assign weight to a term based on how frequent the term occurs in the document. Each term carries its own score and those terms are sorted in descending order. The term with the highest score is considered as relevant term and will be a candidate for the identification of topic.

Equation (1) shows the calculation of term frequency (TF) which is to count the number of times a word appears in a document, divided by the total number of words in that document. The length of certain documents may vary; hence, it is possible that a term would appear more times in long documents than shorter ones. Therefore, term frequency is divided by the document length which is the total number of terms in the document.

$$TF(t) = \frac{total\ occurance\ of\ t}{total\ number\ of\ terms} \qquad (1)$$

Equation (2) shows the calculation for Inverse Document Frequency (IDF) which is to compute the importance of a term.

$$IDF(t) = \frac{\log_e(total\ number\ of\ documents)}{number\ of\ documents\ with\ term\ t} \quad (2)$$

In the calculation of TF all terms are treated as equally important. However, certain terms may appear many times but have insufficient importance or no discriminating power in determining the relevance. Thus, the frequent terms need to be decreased and the rare ones to be scaled up.

Finally, a rule generation algorithm is used to identify suitable topics for the text. Fig. 3 presents the rule generation algorithm for topic identification, known as TopId.

```
Start
    For each verse
        IF $\sum x = 1$ THEN check imd library
        IF x exist in i THEN y = Inheritance
        ELSE
            IF x exist in m THEN y = Marriage
            ELSE
            IF x exist in d THEN y = Divorce
    ELSE
        IF $\sum x > 1$ THEN assign cardinality as |x|
        For each x in \x\
        Check imd library
        IF x exist in i THEN y = Inheritance
        Topic = topic + 1
        ELSE
            IF x exist in m THEN y = Marriage
            Topic = topic + 1
            ELSE
            IF x exist in d THEN y = Divorce
        Topic = topic + 1
    IF topic > 1
    Choose  first topic
            ELSE
            IF $\sum x = 0$ THEN y = out of topic
End
```

Fig. 3. *Rule generation algorithm*

The algorithm determines which terms should be appointed as the topic of the text document. The input needed by TopId is the highest ranked relevant term from term ranking. Assume *x* is the term that has the highest frequency score and *y* is the target topic. There are three conditions based from the number of *x*. First, if *x* equal to 1, then it will directly check the keywords in library IMD that can match with the term. If the term is matched with keywords in library I, then the topic is Inheritance. If the term is matched with keywords in library M, then the topic is Marriage. Lastly, if the term is matched with keywords in library D, then the topic is Divorce.

Otherwise, in the case where the number of terms with the highest frequency score is more than one, then each found term is checked with the library one after another. The number of found terms is classified as cardinality of x (|x|). The process is continued until the condition of |x| is equal to zero. If two terms have the same score, TopId takes the first topic. Meanwhile, if there is no match between terms and keywords, it is considered as out of topic. The output from TopId are the rules along with the topic for each verse. The condition of the decision rules depends on the availability of certain terms from the keywords database. A rule-based classifier is used to determine term patterns which are related to different classes. The produced rules are represented in First Order Predicate Logic (FOPL). For instance, the rule for verse 2_49 is:

$\forall X\ term(son) \land belongs_{to}(son,libraryMarriage) \rightarrow is\_topic(son,Marriage)$

In the set of rule, the left-hand side corresponds to a term pattern, and the right-hand side corresponds to a class label. This rule is used for the purpose of classification. Rule-based classifier is employed in this study because it is highly expressive and it is almost equivalent to a decision tree. Rule-based classifier also allows multiple rules to be triggered for a given record and the interpretation is understandable. The quality of the classification rules can be later evaluated using accuracy measurement based on the produced topics. In order to compare the rules and topics produced by TopId, an experiment involving Rough Set technique and expert opinions are also employed.

III. RESULTS AND DISCUSSION

The experiments are conducted on 224 verses from English translated Quran. Table I shows a sample of pre-processed and filtered terms, taken from Verse 2_237, using the filtering algorithm. The filtered terms consists of nouns and some important terms from library myNounVerse. From the result, there are terms that occur two times in Verse 2_237, which is term ‘dower’.

TABLE I. SAMPLE OF PRE-PROCESSED AND FILTERED TERMS FROM VERSE

Verse No	Verse	Filtered terms
2_237	And if ye divorce them before consummation, but after the fixation of a dower for them, then the half of the dower (Is due to them), unless they remit it or (the man's half) is remitted by him in whose hands is the marriage tie; and the remission (of the man's half) is the nearest to righteousness. And do not forget Liberality between yourselves. For Allah sees well all that ye do.	divorce dower dower man marriage tie man

These filtered terms are then calculated and ranked using TF-IDF technique. No technique comparison has been made in term ranking as it is only been adopted in the method. The calculation is shown in Table II.

TABLE II. THE TF-IDF CALCULATION FOR EACH TERM IN VERSE 2_237

Verse (*d*)	Total terms in *d*	Terms (*t*)	Number of t occurs in *d*	Tf	Idf	Tfidf
2_237	14	Dower	2	0.1429	0.2350	0.0336
		Man	2	0.1429	0.0979	0.0140
		divorce	1	0.0714	0.1567	0.0112
		marriage	1	0.0714	0.2938	0.0210
		Tie	1	0.0714	1.1751	0.0839

Based on Table II, there are 14 terms in Verse 2_237, represented by *d*. However, only five terms are filtered, which are 'dower', 'man', 'divorce', 'marriage' and 'tie'. Each term is labeled as *t* and number of occurrences for each terms is counted. At this stage, each *t* must be calculated using Equations 1 and 2. For example, to calculate the TF for term 'dower', the total number 'dower' occurs in the verse is 2 and divided by the total number of all terms in the verse which is 14. The score is 0.1429. To calculate the IDF, there are 224 verses and the term 'dower' occurs only fifteen times in these verses. Then, the IDF is calculated as *log* (224/15) and equivalent to 0.2350.

Next, the value for TFIDF is computed as 0.01429 multiply with 0.2350 and is equivalent to 0.0336. There is no threshold to determine the value of score for the relevant terms. Therefore, the term with highest score is considered as the most relevant term. From Table 2,, the most relevant terms in the verse are top-ranked based on its score. Terms 'dower' and 'man' appear twice in Verse 2_237, but the score for both terms are different. Therefore, the potential term in Verse 2_237 is 'dower'. The same calculation is conducted to each term in all 224 verses.

The weighted terms produced from this experiment is then used as input for TopId. Rough set rule generation has been used as comparison with TopId. Rough Set rule generation technique is performed using Rosetta application to produce rules for topic identification. In contemplating of avoiding bias in choosing the best model of rules produced by Rosetta application, 10-Fold Cross Validation experiment have been conducted onto four groups of split factor, which are 0.2, 0.3, 0.7 and 0.8. Ten experiments have been carried out for each split factor.

Rough Set technique is chosen to be compared with TopId since Rough Set has been used for rule generation in several text processing studies. Rosetta application is used to generate rule using Rough Set technique. The comparison has been made by comparing the topics produced by TopId and Rough Set with three experts in Quran and Hadith research domain. The experts are selected according to their expertise in Quran and Hadith research domain. The topics given by the experts are used as a benchmark to evaluate the accurateness of the topics identified by TopId. The results from all comparisons with the experts are not really consistent since some experts identified several topics that are not included in the scope of study.

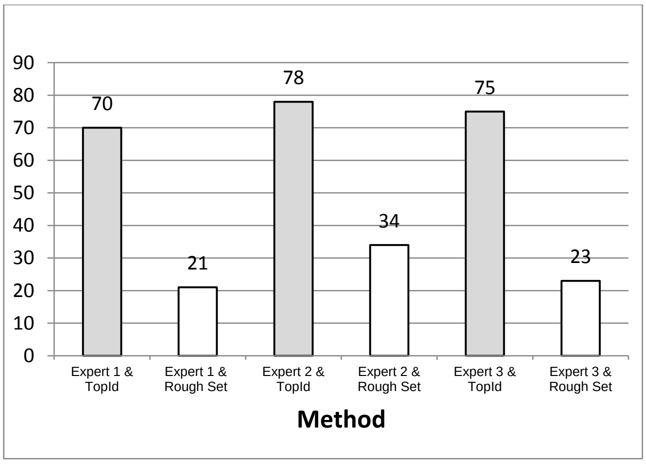

Fig. 4. *Topic accuracy by TopId, Rough Set and experts*

The topic comparisons are measured using the accuracy of the correct topics between TopId with the experts and the topics between Rough Set with the experts. Fig. 4 shows the percentage of accuracy obtained from the comparison with the three experts.

It appears that the accuracies obtained by TopId and the three experts increase moderately by 70% to 78% & and slightly decrease to 75%. These results have proven that TopId is able to identify topics as the total number of matched topics between TopId and the three experts are quite high. As a comparison to TopId, the accuracies obtained by Rough Set technique and the three experts are lower with the percentage of 21% increasing to 34% and decreasing to 23%. The result generated is consistently lower to the result obtained by TopId and the three experts. The reason of this poor result is because there are fewer topics that have been identified by Rough Set technique. Apart from that, Rough Set technique assigned any undefined objects as ‘None’. This result leads to a high number of unmatched topics during the comparison of topics from Rough Set technique and the topics given by the three experts.

IV. Conclusion

The proposed rule generation algorithm was able to identify topics similar to topics identified by the experts. The rule generation algorithm was designed to identify a topic based on the highest ranked terms and match it with the topic classes. The generated rules also suggest that there is only single term used to represent a topic to each verse. Apart from the Holy Quran, the proposed method can be used in identifying topic from other textual data such as meeting transcripts, news articles, and social networking posts such as Twitter, Facebook and Tumblr. Further work can be done to improve the proposed method by analyzing the text to understand the syntax and semantic of words using natural language processing.

Acknowledgment

The author would like to acknowledge the Ministry of Higher Education and Center of Research and Innovation Management (RIMC) of Universiti Utara Malaysia for the management and financial support. Appreciation to Assoc. Prof. Dr. Hj. Abd Aziz Bin Hj. Hanafi, INSANIAH for the technical support regarding Quran.

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] P. M. S and K. S. S, “A concise survey on text data mining,” International Journal of Advanced Research in Computer and Communication Engineering, vol. 3, no. 9, pp. 8040–8043, 2014.

[4] R. Jindal and S. Taneja, “U-STRUCT: A framework for conversion of unstructured text documents into structured form,” Communications in Computer and Information Science, vol. 361 CCIS, pp. 59–69, 2013.

[5] Y. Ko, J. Park, and J. Seo, “Automatic text categorization using the importance of sentences,” in The 19th International Conference on Computational Linguistics, vol. 1, pp. 1–7, 2002.

[6] A. T. Sadiq and S. M. Abdullah, “Hybrid intelligent technique for text categorization,” International Conference on Advanced Computer Science Applications and Technologies, ACSAT 2012, pp. 238–245, 2012.

[7] H. S. Baghdadi and B. Ranaivo-Malançon, “An automatic topic identification algorithm,” Journal of Computer Science, vol. 7, no. 9, pp. 1363–1367, 2011.

[8] A. Brun, K. Smaïli, and J. Haton, “Contribution to topic identification by using word similarity,” INTERSPEECH, no. September, 2002.

[9] S. Jain and J. Pareek, “Automatic topic(s) identification from learning material: An ontological approach,” 2nd International Conference on Computer Engineering and Applications, ICCEA, vol. 2, pp. 358–362, 2010.

[10] D. D. D, B. J. Hithaishy, and P. Bhat, “Semantic approach to identify topics from product reviews,” International Journal of Latest Trends in Engineering and Technology, pp. 1–6, 2013.

[11] C. Devasena and M. Hemalatha, “Automatic text categorization and summarization using rule reduction,” in IEEE International Conference on Advances in Engineering, Science and Management, ICAESM, pp. 594–598, 2012.

[12] L. Massey and wilson wong, “A cognitive-based approach to identify topics in text using the web as a knowledge source,” in Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances, pp. 61–78, 2011.

[13] A. Fuddoly, J. Jaafar, and N. Zamin, “Keywords similarity based topic identification for Indonesian news documents,” 2013 European Modelling Symposium, pp. 14–20, 2013.

[14] H. Anaya-Sánchez, A. Pons-Porrata, and R. Berlanga-Llavori, “A new document clustering algorithm for topic discovering and labeling,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 5197 LNCS, pp. 161–168, 2008.

[15] V. Stoyanov and C. Cardie, “Topic identification for fine-grained opinion analysis,” Proceedings of the 22nd International Conference on Computational Linguistics, no. August. pp. 817–824, 2008.

[16] B. M. Sagar, G. Shobha, and R. K. P, “Solving the noun phrase and verb phrase agreement in Kannada sentences,” International Journal of Computer Theory and Engineering, vol. 1, no. 3, pp. 288–292, 2009.

[17] R. Dong, M. Schaal, M. P. O. Mahony, and B. Smyth, “Topic extraction from online reviews for classification and recommendation,” in Twenty-Third International Joint Conference on Artificial Intelligence Topic, pp. 1310–1316, 2013.