Research Article

International Journal of Advanced Computer Research, Vol 7(30)
ISSN (Print): 2249-7277 ISSN (Online): 2277-7970
http://dx.doi.org/10.19101/IJACR.2017.730020

A subject identification method based on term frequency technique

Nurul Syafidah Jamil*, Ku Ruhana Ku-Mahamud, Aniza Mohamed Din, Faudziah Ahmad, Noraziah ChePa, Wan Hussain Wan Ishak, Roshidi Din and Farzana Kabir Ahmad
School of Computing, College of Arts and Sciences, Universiti Utara Malaysia, Sintok, Kedah, Malaysia

Received: 22-February-2017; Revised: 08-April-2017; Accepted: 11-April-2017

Abstract
The analyzing and extracting important information from a text document is crucial and has produced interest in the area of text mining and information retrieval. This process is used in order to notice particularly in the text. Furthermore, on view of the readers that people tend to read almost everything in text documents to find some specific information. However, reading a text document consumes time to complete and additional time to extract information. Thus, classifying text to a subject can guide a person to find relevant information. In this paper, a subject identification method which is based on term frequency to categorize groups of text into a particular subject is proposed. Since term frequency tends to ignore the semantics of a document, the term extraction algorithm is introduced for improving the result of the extracted relevant terms from the text. The evaluation of the extracted terms has shown that the proposed method is exceeded other extraction techniques.

Keywords
Subject identification, Text classification, Term frequency, Term filtering, Text document.

1.Introduction
Organizing information in the text document is a crucial task as it can ease accessing of the information required and reduce collection of unnecessary information [1-2]. Texts can be categorized via predefined or post defined group or subject. Then, it has data category consist of word attributes and higher dimensional frequencies most of the words [3]. Therefore, it is critical grouping the category of text itself. In a predefined category, a group of texts is placed in one of the categories that have been identified earlier. However, if the categories are post defined, the whole text is grouped according to a topic or subject. Currently, for both categories, many techniques have been developed. The predefined category requires classification techniques to place texts into categories such as Naive Bayes and k-nearest neighbour [4-5]. The post defined category requires clustering techniques such as linkage metrics, k-medoids, k-means, or any other clustering techniques to group texts into specific topics or subjects [6]. In this study, the texts classified from related documents that are relevant must be extracted without losing important information.

*Author for correspondence

These extracted texts are then matched to a group of the list of subjects. Then, the closest subject is chosen as the subject or topic [7-8]. The identification of correct subject will allow a user to grasp useful information and has a better understanding of the meaning.

A subject helps human to retrieve meaningful textual documents and assist people searching and understand the whole sentence in text [9]. Subject identification is different from text summarization. Then, the subject identification tends to only give single terms to represent the full theme of a text document. Contrarily, text summarization compresses the content in the text document into a shorter version of information [10]. Then, the text summarization is expected as the way in decreasing the size of text complexity of document without modifying the original text in the document [11]. Subject identification is a classification problem where the task is to assign a correct topic label to a group of text [12].

In other related researches, the term "subject" is not a phrase but is a single word that explains the main content and has been used interchangeably with "theme" or "topic" [13]. An example of subject from a sentence is shown below.

"*Students who enjoy playing outdoor games may be amateur athletes who play recreationally".*

In the above sentence, the subject would be "sports" because the majority of the words (outdoor games, amateur athlete, and recreationally) relates to "sports". In short, subjects can give a basic idea of a group of texts [12, 14]. The works of the subject or topic identification in [15] employs the natural language processing. Others have performed the tasks through a statistical approach [16-18], ontological approach such as in [14,19,20], and combination of statistical approaches and computational linguistic approaches such as in [10] and [12]. For example, identifying subjects from text document is usually performed based on text mining and statistical approach. In text mining, subject identification is usually done through morphological analysis, but this method involves too many linguistic tools such as tokenization, stemming, parsing, and tagging to perform various tasks. This consumes too much time and costs and requires human intervention for confirmation. Due to these problems, subject identification conducted in this research eliminates the unnecessary steps.

In statistics, term frequency-inverse diverse frequency (TF-IDF) has been widely used. The technique is based on frequency calculation that counts the occurrences of words in the documents. However, counting words can be insignificant and a waste of time if the word being counted does not carry impact to the whole meaning. For example, in the sentence have given earlier the word count for "*who*" is 2. However, the word "*who*" does not give the useful meaning in the sentence. Therefore, this word cannot be considered for being a subject.

Due to the weakness of TF-IDF, this research has included another step that filters words that are insignificant. This was to avoid calculating unnecessary words. In this research, two methods, computational linguistic and TF-IDF have been applied to produce an algorithm that identifies subjects from text data. The integrated approach is undertaken to ensure that a correct subject is identified and tagged to a group of texts. Combination of computational linguistics and statistics and has been used to identify the appropriate subject. Computational linguistic was used to identify significant words and statistical approach. TF-IDF was used to determine the appropriate subject for a group of texts.

2.The proposed method

The method that proposed in this approach was beginning with preparing the data as a requirement is; The English Translated Quran, retrieved from Surah.my website (http://www.surah.my) was used as the dataset. The Surah.my website was chosen because statistics from the website showed that the website is mostly referred to by Malaysians [21]. It's 224 verses (out of 6666 verses) were used for experimentation and only verses that contained 16 keywords on female were chosen. The keywords used were daughter, female, woman, damsel, niece, mother, aunt, consort, divorcee, girl, lady, maiden, sister, widow, wife and queen. Based on the content of the verses, an appropriate subject will be identified and tagged to it and three subjects, inheritance, marriage and divorce that have been predefined will be used. These subjects and the verses will be matched together and finally a subject is appointed to each verse. Furthermore, the process of the method consists of four essential phases were involved: text preprocessing, term extraction, term calculation and ranking, and subject identification as shown in *Figure 1*.

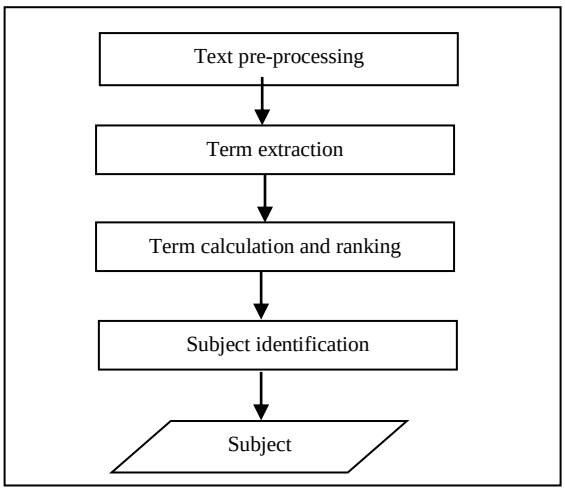

Figure 1 The proposed subject identification method

Based on *Figure 1*, the method begins with text pre-processing phase and then is a term extracting phase. Text pre-processing is necessary to reduce the high dimensionality problem of processing textual data. This phase expected in filtering the large volume of textual document in order to facilitate the searching for the relevant information. In term extracting, term which are nouns are taken as relevant terms and the other terms are categorized as noise word and irrelevant terms. Next, the term calculation and ranking phase is to calculate the frequency of the

extracted words, then ranked the words based on the frequency. In the last phase (subject identification), a word will be chosen and appointed as the subject of the text document. The following paragraphs show details of these phases:

2.1 Text pre-processing and term extraction

There are various available methods that have been introduced and developed in order to extract and filter valuable information from texts.

The computational linguistic method integrated with the rule based process is chosen because it is capable to produce promising results compared to automated shallow methods such as statistical base approach alone. *Figure 2* shows the flowchart of text pre-processing and term extraction and *Figure 3* shows the pseudo code of text pre-processing and term extraction.

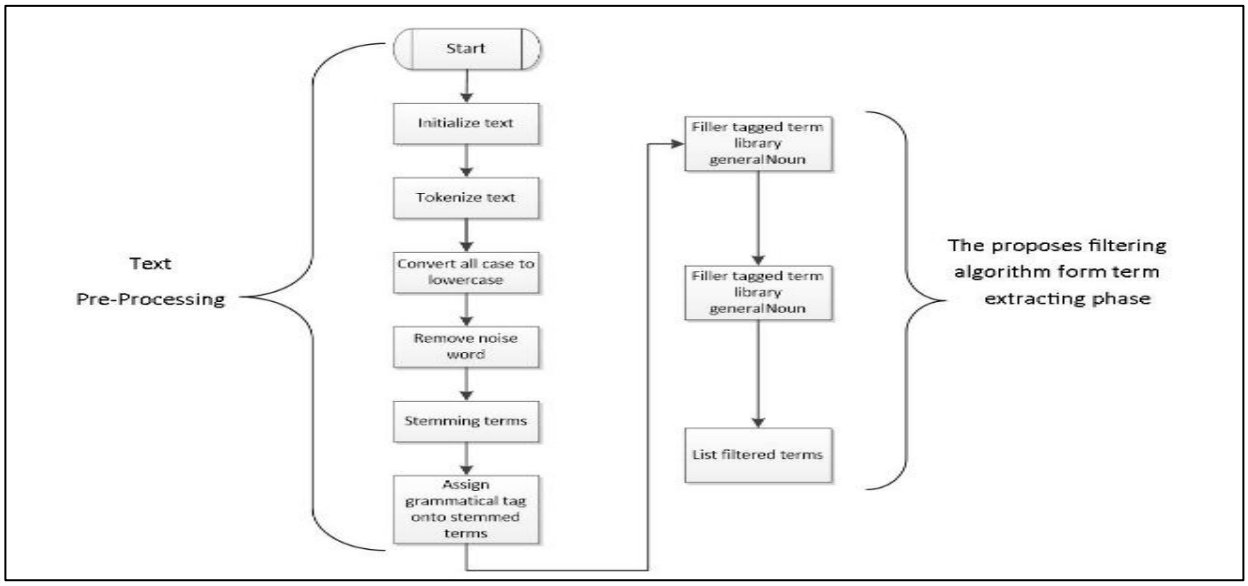

Figure 2 The flowchart of the text pre-processing and term extraction

```
Start
    Execute text nomalization
        Inttialize verse
        Tokenize verse
        Transfrom all case to lowercase
        Remove noise word using stop word removal
        Perform stemming verse
        Assign grammmatical tag using POS tagging
        Each term
    End text pre-processing
    Start term extraction
        For all tagged terms
        Filler tagged terrms with libarry generalNoun
        Filler tagged terrms with libarry generalVerse
        List matched terms from filtrations
    End For
    End term extraction
    End
```

Figure 3 The pseudo code of text pre-processing and term extraction

2.1.1Text Pre-processing and term extraction

Irrelevant words in the text document may create 'noise' that makes the information less distinguishable. Hence, the purpose of this initial phase is to remove noise words, punctuations, numbers, misspelling and others. The process for this phase is described in *Figure 4* below.

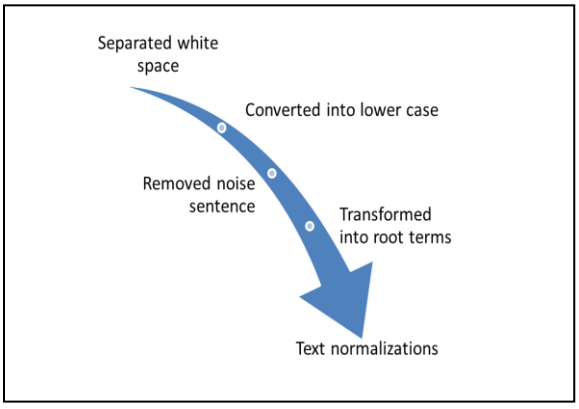

Figure 4 The steps for text pre-processing

First, the texts are tokenized by separating white space between words. The characters are converted into lowercase to avoid same words to be considered as a different meaning. This operation is essential on the account that is necessary to ensure the exact number of the repeated terms in the text. Next, noise in the sentences such as numbers, articles, preposition, and punctuations is removed.

Then all words are transformed into its root term by using stemming technique. Text normalization is a compulsory step in any text processing area. Thus, no changes or enhancements are made in this phase. The expected output of this phase will be terms that are free from noises. The task for this phase was done using Python natural language tool kit. The pseudo code is shown in *Figure 5*.

```
Start
Execute text normalization
        Initialize verse
        Tokenize verse
        Transform all case to lowercase
        Remove noise word using stop
word removal
        Perform stemming verse
End text normalization
End
```

Figure 5 Text normalization pseudo code

2.1.2Term extraction

A library known as *NounList* was created. The library contains only selected nouns, future tense and verbs. These are denoted as keywords. Here, the terms produced from text normalization are checked against the library. Initially, the process starts by assigning part-of-speech (POS) tagging onto each term. For example, term 'marriage' will be tagged as a noun (NN). Another term such as 'marry' will be tagged as a verb (VB). Then, a list of tagged terms is produced and the focus is to collect only terms with noun values and terms that matched with the keywords from the *NounList* library. However, in this research, there are three important terms (marry, wed and will) and these terms are strong keywords which will be based on the subject identification. For example, in Quran context, 'will' is referring to an act to give (property) to another person after one's death. Meanwhile, in linguistic aspect the term 'will' is belongs to the future tense class. Thus, in order to perform filtering process, an algorithm has been developed. The algorithm process is further explained consecutively. *Figure 6* shows the extraction algorithm.

The algorithm starts after the terms are tagged by using POS tagging. During the first filtering process, if there are matched terms with the keywords such as 'marry', 'wed' and 'will' from the tagged terms, then the matched terms are collected and labelled as list A. Next, the remaining terms which have various grammatical tags were also collected and labeled as list B. Later, both lists A and list B were combined and each term in this list is filtered again by comparing the tagged terms with the keywords in the *NounList* keywords database. Only terms that matched with the keywords from the database are collected and this collection of matched terms is labelled as *list C*. At this stage, the number of the relevant features is reduced rather than the number of terms before filtering it. The extracted terms later are ranked in the next following phase.

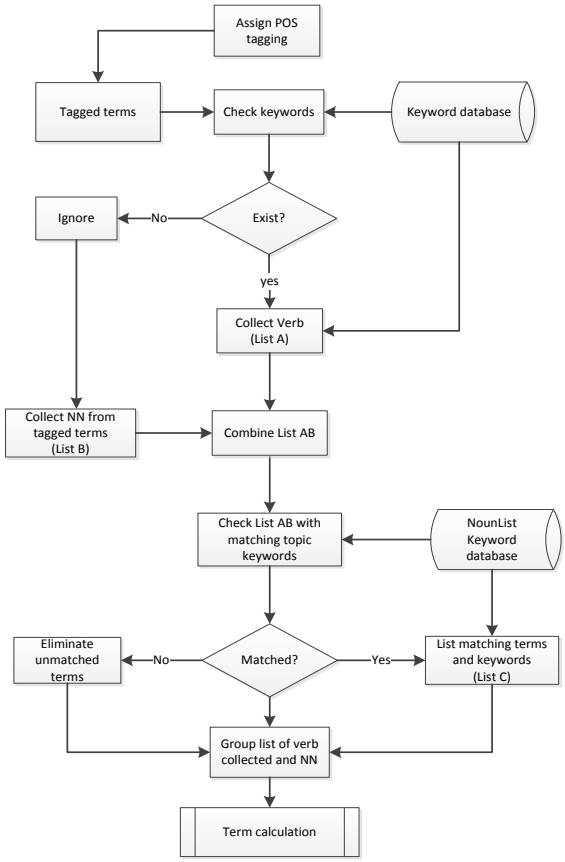

Figure 6 The term extraction algorithm

Table 1 shows the sample of terms produced from the filtration phase.

Table 1 Sample of relevant terms after filtration phase

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 consummation fixation dower (2 terms) man hand marriage tie remission man righteousness liberality Allah
4_23	Prohibited to you (For marriage) are:- Your mothers, daughters, sisters; father's sisters, Mother's sisters; brother's daughters, sister's daughters; foster-mothers (Who gave you suck), foster-sisters; your wives' mothers; your step-daughters under your guardianship, born of your wives to whom ye have gone in,- no prohibition if ye have not gone in;- (Those who have been) wives of your sons proceeding from your loins; and two sisters in wedlock at one and the same time, except for what is past; for Allah is Oft-forgiving, Most Merciful	marriage mother (3 terms) daughter (4 terms) sister (5 terms) father brother daughter (3 terms) wife (3 terms) guardianship born prohibition proceeding sister wedlock time Allah

2.2Term calculation and ranking

After the list of filtered terms is achieved, each of the terms is weighted by calculating the frequency to capture the number of occurrences in the text. TF-IDF technique is selected for this phase since it can assign weight to a term based on how frequent the term occurs in the document. There are two steps for calculating the TF-IDF. First, the term frequency (TF) is calculated by counting the number of times a word appears in a document. It is then divided by the total number of words in that document. The length of certain document may vary and it is possible that a term would appear much more time in longer documents than shorter ones. Therefore, term frequency is divided by the document length, which is the total number of terms in the document as shown in equation (1);

$$TF(t) = \frac{Na}{td} \quad (1)$$

where,

TF = Total frequency

Na = Number of times term t appears in a document

td = Total number of terms in the document

Based on *Table 2*, it is shown the list of the number appearances in every term. Most of the number of appearances of some terms is 1 and only ‘dower’ and ‘man’ have 2 numbers of appearances.

Table 2 Sample of terms, and its number of applications in the text

Verse no	Terms	Number of appearances
2_237	divorce	1
	consummation	1
	fixation	1
	dower	**2**
	man	**2**
	hand	1
	marriage	1
	tie	1
	remission	1
	righteousness	1
	liberality	1
	allah	1

3. Evaluation

The evaluation of the proposed subject identification is performed to compare the number of the extracted relevant terms as subject candidate. The experiment has been conducted onto 224 verses. The extracted terms and total number of extracted terms obtained from these experiments are compared and analysed. Extraction techniques such as a rough set attribute reduction (RSAR) and information retrieval has been chosen for the experiment. Due to the shortness of the space for this paper, the sample of the extracted terms after the experiments is presented in *Table 3*.

Table 3 The sample for the comparison of the extracted terms

Verse no	Extracted terms produced by:					
	Proposed extraction algorithm		**Rough set attribute reduction technique (RSAR)**		**Information retrieval pre-processing (IR)**	
	Terms	**No of terms**	**Terms**	**No of terms**	**Terms**	**No of terms**
2_237	divorce consummation fixation dower man hand marriage tie remission righteousness liberality	11	divorce dower man marriage treasure	5	divorce consummation fixation dower half dower unless remit man half remit hand marriage tie remission man half righteousness do forget liberality see all do	12
4_23	marriage mother daughter sister father brother guardianship born wife prohibition proceeding wedlock time	13	daughter father marriage mother wife	5	prohibit marriage mother daughter sister father sister mother sister brother daughter sister daughter foster-mother give suck foster-sister wife mother step-daughter guardianship born wife prohibition wife son proceed loin two sister wedlock one same time past merciful	36

Table 3 presented the most relevant terms produced by the proposed extraction algorithm, RSAR and IR. From the result, it can be seen that RSAR extracted a limited number of the relevant terms. The extracted terms cannot be too small since it can cause bias in classification. Contrarily, IR extracted too many terms from the text which the classification can be too cumbersome. Meanwhile, the proposed extraction algorithm produced a sufficient number of extracted terms from the text. In addition, the terms which have been extracted by the proposed extraction algorithm are more relevant since it has keyword checking from the keyword database. Contradictorily, both RSAR and IR did not perform keyword checking process. Hence, the proposed extraction method produces more relevant terms to be ranked.

4.Discussion

Every term has own rank and own score of TF, IDF, and TF-IDF. *Table 4* shows the score of each term and its rank. Later, each term is calculated by using TF-IDF formula. From *Table 4*, the highest score is shown in the second last column. The last column shows the rank. Value 1 shows the highest rank (highest score of TF-IDF) while the lowest score holds the largest rank value. The result of score comparison of the TF-IDF with TF and IDF shows in *Table 4* as follows.

Table 4 Sample of text data scores

Terms	Score (TF)	Score (IDF)	Score (TF-IDF)	Rank
divorce	0.014	0.16	2.24	5
consummation	0.014	2.35	0.03	8
fixation	0.014	1.18	0.02	9
dower	0.028	0.20	5.60	1
man	0.028	0.09	2.52	4
hand	0.014	0.12	1.68	7
marriage	0.014	0.34	4.76	2
tie	0.014	1.18	0.02	6
remission	0.014	1.18	0.02	9
righteousness	0.014	2.35	0.03	8
liberality	0.014	2.35	0.03	8
Allah	0.014	0.02	2.8	3

From the rank column, the term with the highest frequency score is taken as the subject. The term ‘dower’ is top-ranked (Rank 1) because it has the highest score amongst all terms. This is followed by

'marriage', 'man', 'divorce' and 'tie'. The other terms are considered as less important in the document, thus, for the Surah Al-Baqarah verse 237, 'dower' is identified as the subject.

5.Conclusion

This study proposed a subject identification method to identify subjects for groups of text. An algorithm based on term frequency technique was developed for this purpose. This algorithm was tested and results showed that it was able to identify a suitable 'subject' for the selected verse. The algorithm produced showed that combining computational linguistic method and statistical method can be more effective for selecting the best subject.

Acknowledgment

This work was supported in part by the School of Computing, Universiti Utara Malaysia, Sintok, Kedah, Malaysia under Grant PBIT (12311) under UUM RIMC grant.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] Korde V, Mahender CN. Text classification and classifiers: a survey. International Journal of Artificial Intelligence & Applications. 2012; 3(2):85-99.

[2] Weiss SM, Indurkhya N, Zhang T, Damerau F. Text mining: predictive methods for analyzing unstructured information. Springer Science & Business Media; 2010.

[3] Aggarwal CC, Zhai C. A survey of text classification algorithms. In mining text data 2012 (pp. 163-222). Springer US.

[4] Patil TR, Sherekar SS. Performance analysis of Naive Bayes and J48 classification algorithm for data classification. International Journal of Computer Science and Applications. 2013; 6(2):256-61.

[5] Elmehdwi Y, Samanthula BK, Jiang W. Secure k-nearest neighbor query over encrypted data in outsourced environments. In international conference on data engineering 2014 (pp. 664-75). IEEE.

[6] Celebi ME, Kingravi HA, Vela PA. A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Systems with Applications. 2013; 40(1):200-10.

[7] Bouamor D, Semmar N, Zweigenbaum P. Using wordnet and semantic similarity for bilingual terminology mining from comparable corpora. In proceedings of the 6th workshop on building and using comparable corpora 2013 (pp. 16-23).

[8] Gupta R, Pal S, Bandyopadhyay S. Improving MT system using extracted parallel fragments of text from comparable corpora. In proceedings of 6th workshop of building and using comparable corpora 2013 (pp. 69-76).

[9] Ker SJ, Chen JN. A text categorization based on summarization technique. In proceedings of the ACL-2000 workshop on recent advances in natural language processing and information retrieval: held in conjunction with the 38th annual meeting of the association for computational linguistics (pp. 79-83). Association for Computational Linguistics.

[10] Baghdadi HS, Ranaivo-Malançon B. An automatic topic identification algorithm. Journal of Computer Science. 2011; 7(9):1363-7.

[11] Meena YK, Jain A, Gopalani D. Survey on graph and cluster based approaches in multi-document text summarization. In recent advances and innovations in engineering 2014 (pp. 1-5). IEEE.

[12] Sawant Ganesh S, Kanawade Bhavana R. A review on topic modeling in information retrieval. 2014.

[13] Butarbutar M, McRoy S. Indexing text documents based on topic identification. In international symposium on string processing and information retrieval 2004 (pp. 113-24). Springer Berlin Heidelberg.

[14] Jain S, Pareek J. Automatic topic (s) identification from learning material: An ontological approach. In second international conference on computer engineering and applications 2010 (pp. 358-62). IEEE.

[15] McDonough J, Ng K, Jeanrenaud P, Gish H, Rohlicek JR. Approaches to topic identification on the switchboard corpus. In international conference on acoustics, speech, and signal processing. 1994 (pp. I-385). IEEE.

[16] Berkowitz S. Method of identifying topic of text using nouns. The United States of America as represented by the Director National Security Agency. United States Patent US 7,805,291. 2010.

[17] Dalal MK, Zaveri MA. Automatic text classification of sports blog data. In computing, communications and applications conference 2012 (pp. 219-22). IEEE.

[18] Van Zaanen M, Kanters P. Automatic mood classification using TF* IDF based on lyrics. In international society for music information retrieval conference 2010 (pp. 75-80).

[19] Coursey K, Mihalcea R, Moen W. Using encyclopedic knowledge for automatic topic identification. In proceedings of the thirteenth conference on computational natural language learning 2009 (pp. 210-8). Association for Computational Linguistics.

[20] Schönhofen P. Identifying document topics using the Wikipedia category network. Web Intelligence and Agent Systems: an International Journal. 2009; 7(2):195-207.

[21] Ku-Mahamud KR, Ahmad F, Mohamed Din A, Ishak W, Hussain W, Ahmad FK, et al. Semantic network representation of female related issues from the Holy Quran. Knowledge management international conference 2012 (pp. 726-30).

Nurul Syafidah Jamil holds both Bachelor in Information Technology (Software Engineering) and Master in Information Technology (by research) from Universiti Utara Malaysia. She is currently pursuing her PhD studies at the same university since August 2016. Her research interest is text mining, topic identification, and sentiment analysis.

Email: jamil.nurulsyafidah@gmail.com

Prof. Dr. Ku Ruhana holds a Bachelor in Mathematical Sciences and a Masters degree in Computing, both from Bradford University, United Kingdom in 1983 and 1986 respectively. Her PhD in Computer Science was obtained from Universiti Pertanian Malaysia in 1994. As an academic, her research interests include ant colony optimization, pattern classification and vehicle routing problem.

Aniza Mohamed Din holds a Bachelor in Information Technology from Universiti Utara Malaysia in 1998 and a Masters degree in Computer Science from Universiti Sains Malaysia in 1999. Her research interests include resource allocation problem, optimization, metaheuristics, genetic algorithm and artificial immune systems.

Noraziah ChePa joined UUM as a lecturer since May 2000. Having a deep interest with neuro modelling in health-related issues. Her research activities include knowledge representation of women's issues in the Quran, Neural Network modelling, simulation agent for stress modelling and neuro–modelling.

Dr. Farzana Kabir Ahmad is a senior lecturer at School of Computing, Universiti Utara Malaysia. She holds a Bachelor degree in Computer Sciences (with Honours) from Universiti Sains Malaysia in 2003 and a Master degree in Computer Science from the same university in 2005. She pursued her PhD in Computer Science (Bioinformatics) from Universiti Teknologi Malaysia in 2012. Her research interests are gene regulatory network (GRN), neuroscience and Neuroinformatics.

Wan Hussain Wan Ishak received the Bachelor in Information Technology and Master of Science in Information Technology from Universiti Utara Malaysia in 2000 and 2003 respectively. Currently, he is a Senior Lecturer at Universiti Utara Malaysia. His research interests include intelligent system, content management, and web application.

Dr. Roshidi Din received his Bachelor of Information Technology and Master of Science in Information Technology degrees from Universiti Utara Malaysia (UUM) in 1996 and 1999 respectively. He later completed his Ph.D from Universiti Sains Malaysia (USM) in 2015. His current research interests more on the application of discrete mathematics in various areas, especially in information security, steganography and steganalysis, and natural language steganology.

Dr. Faudziah Ahmad is an associate professor at Universiti Utara Malaysia (UUM). She has been with UUM since 1990 and have been teaching in several courses at the masters and undergraduate levels. Among the courses taught are Artificial Intelligence, Intelligent Database, Knowledge Discovery in databases Research Methodology. Her research work is grounded in theories and methods found in the field of artificial intelligence, specifically in data mining, text mining, and computational modelling. Currently, she supervises Ph.D and Masters students in research works related to artificial intelligence.