*International Journal of Engineering & Technology*, 7 (2.14) (2018) 13-16

International Journal of Engineering & Technology

*Website: www.sciencepubco.com/index.php/IJET*

***Research Paper***

# Single Term Concepts from English Translated Qur'an Using Statistical Methods

**Rohana Ismail[1,2]*, Nurazzah Abd. Rahman[2], Zainab Abu Bakar[3]**

*[1]Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Kampus Besut, Besut, Terengganu, Malaysia*
*[2]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia*
*[3]Faculty of Computer and Information Technology, Al-Madinah International University, Selangor,*
**Corresponding author E-mail:rohana@unisza.edu.my*

**Abstract**

Ontology is essential for the success of knowledge based systems because it has the opportunity to share vocabulary, integrate knowledge easily and discover new instances or relations. However, the development of ontology via manual is time consuming and tedious task. Thus, ontology learning comes to play it roles. The ontology learning tries to extract ontological elements to support the ontology development. Concept extraction is one of the important tasks in ontology learning. The Hajj domain of Quranic study, concepts have not fully discovered. Hence, this paper tries to discover concepts by extracting the single terms from Qur'an translated version. It provides result on extracting the single terms as concepts by using statistical methods. Apart from that, it has been experimented for English Translated Quran by Hilali Khan. Result shows that the performance of using *tf* method as a statistical method is significant with the f-measure value is 0.509. Based on the *tf*, the comparisons have been made for other statistical methods such as *tfidf*, *Avetf* and *Ridf*.

***Keywords***: *Ontology Learning, Concepts, Qur'an, Statistical.*

## 1. Introduction

In Semantic Web, ontology is important element to support the construction of formal specification of shared conceptualization for a domain. In specific domain such as Qur'an study, ontology is able to model the abstract concepts of Quran into a formal representation. However, the Qur'an ontology still not complete[1,2]. It only focuses on limited domains by using related verses of Qur'an. This happened because of the proposed used of the ontology. Some of them are used for semantic search [3,4], knowledge based and knowledge representation[5,6]. However, the Qur'an Ontology from Leeds is rather having a complete ontology for the Qur'an but still lacking in terms of description of concepts. Therefore, more concepts and other domains need to be explored to support the completeness of the ontology development for Qur'an.
The problem with manual ontology development is the task much more time consuming and tedious[7]. Thus, the ontology learning comes to play it roles by reducing the manual effort of the ontology development. In fact, it has seen much interest in developing automatic or semi- automatic solution for the ontology development. Even though the ontology learning is quite mature in other fields or domains[8], but it still being explored for other domains such as software project and Arabic ontology [9,10]. Various methods can be employed to support the ontology learning activities. The statistical methods have seen as a simple method that can be used for extracting single concepts for the ontology development[11]. The methods are based on term frequency (*tf*), term frequency inverse doc frequency (*tfidf)*, average *tf (Avetf)* and *Ridf.*
This paper has tested and evaluated the statistical methods used in the Qur'an verses related to Hajj domain. It has been tested in order to identify concepts for the Hajj ontology development. The experimental results have been presented by comparing the statistical methods.

## 2. Related Works

An Ontology learning is a technique that used various methods from other field to extract ontological elements to form ontology[12]. The aim of the technique is to reduce effort of manual ontology construction by at least providing semi-automatic manner. It is because the construction of ontology is tedious, expensive and time consuming task[13]. To date various methods have been developed and introduced to construct an ontology by using the ontology learning technique[8]. For a certain domain, its quite established but it still being explored for other fields of domains[8,9,10].
The ontology learning has layers of activities to be full-fill. It includes identification terms and concepts, synonyms, concepts hierarchy, relations and axioms. Mostly, the statistical based and linguistic based methods are extremely used in the ontology learning[8]. Examples of the statistical-based methods are *tfidf, tf* and *glossex*. Meanwhile the linguistic based methods are much more depends on POS tagging, matching syntactic patterns. Moreover, the linguistic based methods can also use the WordNet resources as a related resources or thesaurus. Another category of method is the logic based methods. It has been used to find the axioms or rules in the ontology[8].
Concepts identification is one of the main activities in ontology learning. Concepts should provide the intensional aspect of a domain. Also, it will have a set of instances and a set of multilingual

**Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.**

terms because it will guiding the way the ontology will be constructed[13]. Concepts in CRCTOL have been identified by their multi terms for terrorism text[14]. In CRCTOL, the small number of relevant single terms also can be as a concept by finding their appearing frequently in the multi terms or inferred based on the multi terms. Meanwhile, Concepts in Quran has either in single terms or multi terms such as *Hajj*, *mountain*, *House of Allah, Allah Bounties*, *raging Fire* [15,16,17].

The single terms can be identified using various statistical methods such as term frequency(*tf*), *tfidf*, glossex, Average *tf* and C-Value [11]. The term frequency (*tf*) is a simple method but show significant to be used as a method to find concepts [11]. It will determine concepts by finding the single term occurrences in a text. It also can be normalize by using *tfidf* method. Meanwhile, the multi terms can be identified using N-grams, collocations, Named Entity Recognition (NER) or using linguistic method [16,17,18,19]. The multi terms basically are the combination of terms to form concepts. After getting all terms to be concepts, relations of those concepts can be identified. Relations identification can employ various methods. For example, the use of syntactic pattern from Qur'an text structure for verses related to Solat[5]. The patterns from Qur'an structure has been experimented for five other chapters in Qur'an[20].

## 3. Methodology

This paper has used the English translated version of Qur'an by Hillali Khan. The version is selected because it contains more elaboration on each verse. Since the purpose is to construct the Hajj ontology, verses related to Hajj have been selected. The verses have gone through for a pre-processing step for better retrieve relevant terms. It has removed certain symbols, hyphen marks and also replaced the upper initial pronouns that exist in the middle of verses same as previous experiments[16]. After that, it has been run into a dedicated system to find their tf, tfidf, Avetf and Ridf.

Particularly, given an extracted term in the Hajj verses set, the tf, tfidf, Avetf and Ridf are calculated as follows:

$$\mathrm{tf}(t,d) = f_{t,d}. \quad (1)$$

where $t$ is the occurrences of terms in document $d$.

$$\text{tf-idf}(t,d,D) = \mathrm{tf}(t,d) \bullet \mathrm{idf}(t,D) \quad (2)$$

Where

$$\mathrm{idf}(t,D) = \log\left(\frac{|D|}{df(t)}\right) \quad \text{with}$$

- df(t) = total number of documents in the corpus.
- $D$ = number of documents where the $t$ appears

$$\mathrm{Avetf}(t,d) = \frac{\mathrm{tf}(t,d)}{df(t)} \quad (3)$$

Residual idf =

$$-\log\left(\frac{|D|}{df(t)}\right) + \log\left(1 - \exp\left(-\frac{\mathrm{tf}(t,d)}{df(t)}\right)\right) \quad (4)$$

Lastly, results are evaluated using standard measurement performance i.e. precision (pre), recall (rec) and f-measure. Formula to calculate the performance of extraction is as follows:

$$\text{pre} = \frac{\text{numbers of terms retrieved and relevant}}{\text{numbers of extracted terms by the system}} \quad (5)$$

$$\text{rec} = \frac{\text{numbers of terms retrieved and relevant}}{\text{numbers of extracted terms by manual}} \quad (6)$$

$$\text{f-measure} = \frac{2\,(\text{pre} \bullet \text{rec})}{(\text{pre}+\text{rec})} \quad (7)$$

## 4. Experimental Result

The 53 verses related to Hajj have different topics of Hajj. It produced 614 single terms from 3018 terms. The performance of using *tf* as a statistical method to extract concepts is shown as in *Table 1*.paper

**Table 1:** Performance of ***tf***

| Precision | Recall | F-Measure |
|---|---|---|
| 0.451 | 0.584 | 0.509 |

### 3.4. Figures and tables

Graphs and other numbered figures should appear throughout the text as close to their mention as possible. Figures shouldn't infringe upon the page borders.

Figures and tables must be centered in the column. Large figures and table can be in one column in order to see them more clearly and avoid placing them in the middle of columns. Any table or figure that takes up more than 1 column width must be positioned either at the top or at the bottom of the page

Photos must be crystal clear with such resolution to allow fine details visibility. The elements from any photo must be explained using numbers, letters, etc. The text within a figure or photo must have the same style, shape and height as the caption has.

Any table, figure or picture must have a caption (Fig.1, Table1, etc.) followed by a proper description. All similar graphics must be generated using the same software product (Excel, Origin, Mathematica, etc.). Importing graphics into the article as images (JPG, BMP, PNG, etc.) should be avoided. All similar electronic schematics, charts, program flow, simulated characteristics, etc. from the article should be generated using the same software product. Importing images from other articles or books it's totally forbidden unless they are cite Based on the *Table 1*, it shows that the performance of precision is lower than recall performance. Meanwhile the f-measure is 0.509. In terms of precision result, more relevant terms cannot be retrieved using *tf*. It is because, some of the terms are terms in multi word terms such as the *Bounty of Allah*, *raging Fire*, *remember Allah* and *ways of Prophet Muhammad*. However, the recall shows that, from the retrieved terms, 58% are relevant to be considered as concept. At average, the f-measure 51% can be retrieved.

In addition, for every cut-off 10% from retrieved terms, the performance of precision, recall and f-measure are shown in the *Figure 1*.

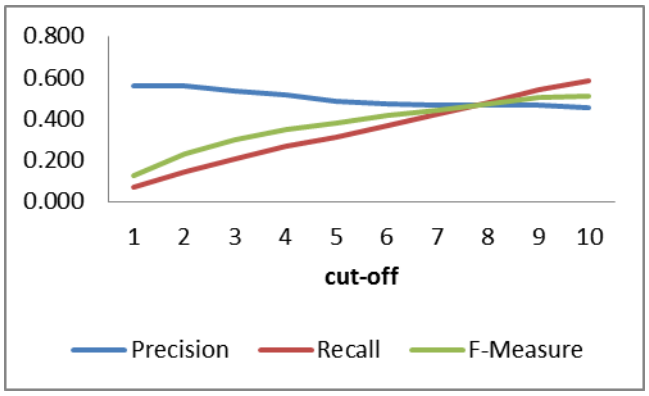



**Fig. 1:** A Cut-Off performance of Precision, Recall and F-Measure

*Figure.1* above shows the cut-off for every 10% retrieved terms. The present result indicates that their performance of precision is higher than recall. It reflects that most terms are being retrieved at first. But, it slowly decreased compared to recall performance. At last, after 100% cut-off or all terms retrieved, the precision, recall and f-measure are 0.451, 0.584 and 0.509 respectively. The recall is higher than the precision. This result shows that, most relevant terms are being retrieved but still lower because less that 60%. More relevant terms needs to be retrieved.

Based on the *tf*, the comparisons have been made for other statistical methods to see their performance of extraction. Such methods are *tfidf*, *Avetf* and *Ridf*. Top twenty results of the methods are shown as in Table 2.

**Table 2:** Results of the terms using *tf*, *tfidf*, *Avetf* and *Ridf*

| ***tf*** | | ***tfidf*** | | ***Avetf*** | | ***Ridf*** | |
|---|---|---|---|---|---|---|---|
| **term** | **Wi** | **term** | **Wi** | **Term** | **Wi** | **term** | **Wi** |
| **Allah** | 131 | Hajj | 0.0040 | having | 0.0013 | having | -2.87766 |
| **Makkah** | 34 | th | 0.0033 | th | 0.0012 | th | -2.93566 |
| **Hajj** | 23 | Makkah | 0.0033 | month | 0.0010 | month | -3.0026 |
| **pilgrimage** | 14 | Kabah | 0.0031 | Allah | 0.0009 | Allah | -3.02538 |
| **House** | 13 | SAW | 0.0031 | Hady | 0.0007 | God | -3.17869 |
| **Kabah** | 13 | days | 0.0031 | afford | 0.0007 | Ilah | -3.17869 |
| **O** | 12 | House | 0.0029 | animals | 0.0007 | beggar | -3.17869 |
| **SAW** | 12 | sacrifice | 0.0029 | God | 0.0007 | come | -3.17869 |
| **mankind** | 12 | mankind | 0.0029 | Ilah | 0.0007 | cut | -3.17869 |
| **Umrah** | 10 | pilgrimage | 0.0028 | beggar | 0.0007 | depart | -3.17869 |
| **sacrifice** | 10 | place | 0.0028 | come | 0.0007 | destroy | -3.17869 |
| **Muhammad** | 9 | whosoever | 0.0028 | cut | 0.0007 | equivalent | -3.17869 |
| **believe** | 9 | Allah | 0.0027 | depart | 0.0007 | fasts | -3.17869 |
| **days** | 9 | people | 0.0026 | destroy | 0.0007 | followed | -3.17869 |
| **good** | 9 | O | 0.0026 | equivalent | 0.0007 | hair | -3.17869 |
| **people** | 9 | Umrah | 0.0026 | fasts | 0.0007 | head | -3.17869 |
| **place** | 9 | cattle | 0.0025 | followed | 0.0007 | houses | -3.17869 |
| **whosoever** | 9 | believe | 0.0024 | hair | 0.0007 | knew | -3.17869 |
| **Verily** | 8 | good | 0.0024 | head | 0.0007 | partners | -3.17869 |
| **cattle** | 8 | men | 0.0024 | houses | 0.0007 | party | -3.17869 |

Based on the above table, *tf* has performed better compared to *tfidf*, *Avetf*, *Ridf* in extracting relevant terms to be concepts. Terms such as *Allah*, *Hajj*, *Makkah* are relevant to be concepts and more meaning full terms compared to terms *having*, *th*.

The statistical methods such as *tf*,*tfidf*, *Avetf* and *Ridf* are significantly able to extract single terms as concepts, but it cannot extract multi terms that are also concepts such as *invoking Allah*, *magnify Allah*, *raging fire*. The multi-terms extractions need to be performed to get more relevant terms and the statistical method also can be used on top of that.

## 5. Conclusion

The statistical methods have been presented in extracting concepts. The experiment results show that, the methods can be used to get single terms concepts. However, a lot of improvement can be done to extract other relevant terms such as multi terms because concepts in Qur'an are either in single terms or multi-terms. Future works includes the extraction of multi terms as concepts using pattern-based techniques.

## Acknowledgement

Special gratitude goes to Faculty of Informatics & Computing and Universiti Sultan Zainal Abidin (UniSZA) for the support and conference funding.

## References

[1] Alrehaili S, Atwell E. Computational ontologies for semantic tagging of the Quran: A survey of past approaches. Lr 2014 Proc. 2014;
[2] Petiwala AJ, Sathya SS. A Multi-Agent System to Learn Literature Ontology : An Experiment on English Quran Corpus. 2011;46–51.
[3] Yauri AR, Kadir RA, Azman A, Azrifah M, Murad A. Quranic Verse Extraction base on Concepts using OWL-DL Ontology. 2013;6(23):4492–8.
[4] Khan HU, Saqlain SM, Shoaib M, Sher M. Ontology Based Semantic Search in Holy Quran. 2013;2(6).
[5] Saad S. Ontology Learning and Population Techniques for English Extended Quranic Translation Text. 2013.
[6] Iqbal R, Mustapha A. An experience of developing Quran ontology with contextual information support. 2013;7(4):333–43.
[7] Hazman M, Rafea A. A Survey of Ontology Learning Approaches. 2011;22(9):36–43.
[8] Wong W, Liu W, Bennamoun M. Ontology learning from text. ACM Comput Surv. 2012 Aug 1;44(4):1–36.
[9] Guo J. Ontology Learning and its Application in Software-Intensive Projects. 2016;6–9.

[10] Albukhitan S, Alnazer A. Arabic Ontology Learning Using Deep Learning. 2017;
[11] Saad S, Salim N, Tiun S. Concept Extraction on Quranic Translation Text. Int J Islam Appl Comput Sci Technol. 2014;2(1):1–9.
[12] Liu Z, Ma Z. Establishing Formalized Representation of Standards for Construction Cost Estimation by using Ontology Learning. Procedia Eng. Elsevier B.V.; 2015;123:291–9.
[13] Buitelaar P, Cimiano P, Magnini B. Ontology Learning from Text : An Overview. 2004;1–10.
[14] Jiang X, Tan AH. CRCTOL: A semantic-based domain ontology learning system. J Am Soc Inf Sci Technol. 2010;61:150–68.
[15] Saad S, Salim N, Zainal H. Islamic knowledge ontology creation. 2009 Int Conf Internet Technol Secur Trans. 2009;(NOVEMBER 2009):1–6.
[16] Ismail R, Abd Rahman N, Abu Bakar Z. Identifying Concept from English Translated Quran. IEEE Conf Open Syst Conf Open Syst. 2017;1–5.
[17] Ismail R, Abu Bakar Z, Abd Rahman N. Ontology Learning Framework for Quran. Adv Sci Lett. 2017;23(5):4175–8.
[18] Zaidi S, Laskri MT, Abdelali A. Arabic collocations extraction using gate. 2010 Int Conf Mach Web Intell ICMWI 2010 - Proc. 2010;473–5.
[19] Ghadfi S, Bechet N, Berio G. Building ontologies from textual resources: A pattern based improvement using deep linguistic information. CEUR Workshop Proc. 2014;1302.
[20] Ismail R, Abu Bakar Z, Abd Rahman N. Extracting Knowledge From English Translated Quran Using NLP Pattern. J Teknol UTM. 2015;77(19):1–6.