MULTILINGUAL TEXT SIMILARITY ANALYSIS IN ISLAMIC TEXTS

by

PAN HUANG

(Under the Direction of Khaled Rasheed)

ABSTRACT

Text similarity measures have been widely studied and used in machine learning and information retrieval for many years. We present a framework with different text similarity measures to delve into the problem of text similarity in the context of multilingual representations of the Qur'an and the Hadith. For the Qur'an, we compare and contrast the effect of applying five similarity measures across four representations of the Qur'an. We analyze the results along two classes namely: the identical verse pairs and the similar verse pairs. Furthermore, we apply the same methodology to the larger text dataset of the Hadith. We employ multithreading technique for speeding up the similarity computations. We compare and contrast the application of similarity measures across the English and Arabic Representations. Based on the results of our text similarity analysis, we propose interlinking of Hadiths with similar semantic content by investigating different equivalence classes by applying different similarity thresholds.

INDEX WORDS: Similarity, Qur'an, Hadith, Arabic, Hamming, Jaccard

# MULTILINGUAL TEXT SIMILARITY ANALYSIS IN ISLAMIC TEXTS

by

PAN HUANG

B.S., Georgia Southwestern State University, 2013

B.S., Chongqing University of Posts and Telecommunications, China, 2013

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment

of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2016

MULTILINGUAL TEXT SIMILARITY ANALYSIS IN ISLAMIC TEXTS

by

PAN HUANG

| | |
|---|---|
| Major Professor: | Khaled Rasheed |
| Committee: | Tianming Liu |
| | Ismailcem Budak Arpinar |

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2016

## DEDICATION

I dedicate this thesis to my family. A special feeling of gratitude goes to my loving parents and grandparents. For my girlfriend, Chenjin Hou, thank you for your love and your three years of support.

ACKNOWLEDGEMENTS

I would first like to thank my thesis advisor Dr.Khaled Rasheed. Dr. Rhsheed's help was always ready for me whenever I ran into a trouble spot or had a question about my research or writing. I really appreciate your guidance, patience and support during the whole three years.

I would also like to thank Amna Basharat and Usman Nisar for their help. Without their help on the paper review and programming check, I would not have finished my thesis.

Additionally, I would like to thank Dr. Liu and Dr. Arpinar for their patience and support.

Finally, I must express my very profound gratitude to my parents and my girlfriend for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of research and writing this thesis.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

# CHAPTER 1

# INTRODUCTION

With the exponential growth of textual information (Harrag, Cherif & Qawasmed, 2008), text similarity analysis has received increasing attention in many research fields in recent years. Thus, text similarity measures and their applications have made a significant impact towards text related search and applications in tasks in many different fields including information retrieval, machine learning and semantic web. However, few applications of text similarity have dealt with multi-lingual translations of a specific document. The growing number and size of texts with more translations being generated increase the challenge of distinguishing or identifying the similarly between texts across different documents. In this thesis, we present a framework with different text similarity measures to delve into the problem of text similarity in the context of multilingual representations of the Qur'an and the Hadith. These two knowledge sources form the primary foundation of the greater body of Islamic texts.

The Qur'an is one of the most widely read books. It is considered by Muslims as a single-authored text, the direct speech of God (Allah) (Atwell, Brierley, Dukes, Sawalha & Sharaf, 2011). The Qur'an contains 6236 verses and around 80000 words (Basharat, Yasdansepas, & Rasheed, 2015). Qur'an study is becoming increasingly popular. The "Semantic Qur'an dataset" project described and utilized in (Harrag, Cherif & Qawasmed, 2008) and the proposed "Quranic Knowledge Map" (Atwell, Brierley, Dukes, Sawalha & Sharaf, 2011) give our research many inspirations. However, little work has been devoted to the similarity study between verse texts on Qur'an. Chapter 2 presents a study in which we use four semantic translations of the Qur'an as

our dataset for comparative study and analysis. We compare and contrast the effect of applying five similarity measures across these four representations. We analyze the results along two classes namely: the identical verse pairs and the highly similar verse pairs. In addition, we extend our investigation to similar verse pairs in the Qur'an and perform an in-depth comparison across the multiple different translations of the Qur'an.

The Sunnah of the Prophet Muhammad (SAW) is the second of the two revealed fundamental knowledge sources of Islam (Aldhaln, Zeki, & Zeki, 2012). The Hadith text is a larger dataset than the Qur'an text. For the Hadith, we utilize the same methodology which we applied to the Qur'an to apply to the Hadith texts. Chapter 3 presents the work in which we undertook a similarity study between texts on the Hadith. For the Hadith, we introduced the multithreading technique for speeding up the efficiency of similarity computation and Hash map for saving space when storing the text-vector. These two ideas significantly reduce the time complexity of our framework. Based on the experimentation on Hadith, we compare and contrast the application of similarity measures across the English and Arabic representations of the Hadith. Additionally, based on the results of our text similarity analysis, we propose interlinking of Hadiths with similar semantic content by investigating different equivalence classes by applying different similarity thresholds.

# CHAPTER 2

# ANALYSIS OF THE EFFECT OF DISTANCE METRIC ACROSS LANGUAGES ON VERSE SIMILARITY IN THE QUR'AN[1]

[1]Pan Huang, Amna Basharat, Khaled Rasheed. To appear in the 18th International Conference on Artificial Intelligence (ICAI,2016).

## 2.1 Abstract

Text similarity measures have been widely studied and used in machine learning and information retrieval for years. However, few applications of text similarity have been utilized on multi-lingual translations of a specific document. Additionally, the growing number of texts with more translations being generated increases the challenge of distinguishing or identifying the similarity and differences between texts across different documents. In this chapter, we employ different text similarity measures to delve into the problem of text similarity in the context of multi-lingual representations of the Qur'an. Four semantic translations of the Qur'an as our dataset are used for comparative study and analysis. We compare and contrast the effect of applying five similarity measures across these representations. We analyze the results along two classes namely: the identical verse pairs and similar verse pairs. Our analysis provides helpful observations about the impact of applying the five distance metrics for determining the verse similarity in the Qur'an across different languages.

## 2.2 Introduction

Text similarity measures and their applications have made a significant impact towards text related search and applications in tasks in many different fields including information retrieval, machine learning and semantic web (Gomaa & Fahmy, 2013). Text is composed of words. If two words are constructed by similar characters with similar sequence, the words are considered to be lexically similar; if two words represent similar semantics, and are used in the same context, the two words are considered to be semantically similar (Gomaa & Fahmy, 2013).

In this chapter, we employ five suitable text similarity or distance measures to investigate the problem of analyzing similarity between the verses of the Qur'an. The Qur'an is one of the most widely read books. It is considered by Muslims as a single-authored text, the direct speech of God (Allah) (Atwell, Brierley, Dukes, Sawalha & Sharaf, 2011). The Qur'an contains 6236 verses and around 80000 words (Basharat, Yasdansepas, & Rasheed, 2015). The original data format was spoken Classic Arabic (Atwell, Habash, Louw, Shawar, McEnery, Zaghouani, & ElHaj, 2010). Recently, Qur'an study has received increasing attention by many researchers, including the "Semantic Qur'an dataset" project described and utilized in (Sherif & Ngomo, 2009) and the future research proposed by Eric and his colleagues about a structured large-scale online resource for understanding the Qur'an, which is called "Quranic Knowledge Map" (Atwell, Brierley, Dukes, Sawalha & Sharaf, 2011). However, However, little work has been devoted to the similarity study between verse texts on Qur'an. The QurSim introduced in (Sharaf & Atwell, 2012) has some interesting findings, whose content is consisting of semantically similar or related verses linked.

The text of the Qur'an presents an interesting dataset in terms of analyzing lexical and semantic similarity between the different verses. We previously undertook a study to compare the similarity of the verses in the Qur'an particularly across the different representations in the Arabic language and an English translation in (Basharat, Yasdansepas, & Rasheed, 2015). We developed a Verse to Verse similarity computation framework for the Qur'an. We applied four similarity measures and three term weighting methods across different representations of the Qur'an. Specifically, we analyzed four datasets: 1) The original Arabic script of the Qur'an with diacritics or case markings, 2) The Arabic script without diacritics, 3) A dataset for the Arabic word roots in the Qur'an and 4) An English translation of the Qur'an. This study resulted in

several interesting findings. Our analysis was largely limited to the identical verses in the Qur'an. However, it was evident that there are several verse pairs that are considerably similar and need further analysis. Also based on the analysis of the identical verses alone, it was concluded that such insights could indeed reflect upon the precision and accuracy of a given translation of the Qur'an. Thus, the study provided a basis for further analysis and investigation.

In this chapter, we extend our previous work to expand our investigation of the similarity amongst the verses of the Qur'an done in (Basharat, Yasdansepas, & Rasheed, 2015). We keep three of the four best similarity computation measures from (Basharat, Yasdansepas, & Rasheed, 2015) and add two more measures, the Hamming distance and the Manhattan distance. These two measures are classified as term-based similarity measures according to the survey done by Wael and Aly in (Gomaa & Fahmy, 2013). We do this in order to investigate the effect of distance measures on the similarity computation of the verses in the Qur'an in order to derive a conclusion about the most effective distance measure for computing the similarity. More importantly, we apply our similarity measures across four different representations of the Qur'an in three different languages. From the results of the study in (Basharat, Yasdansepas, & Rasheed, 2015), we found the Arabic representation without diacritics to give the most accurate results. We therefore use that as the baseline result for this study. We introduce one more English translation, in addition to the one we experimented with earlier, to give us better grounds for analyzing similarities across different translations in the same language. We also introduce another translation of the Qur'an, in the Urdu language, as one of the datasets, in order to analyze the similarities across different translations of the Qur'an in different languages. Research shows that similarity analysis within and across languages always obtains more interesting and valuable information (Forsyth & Sharoff, 2014). We therefore aim to expand our insights about the

similarity patterns in the Qur'an through this investigation. Moreover, the experiments done in (Basharat, Yasdansepas, & Rasheed, 2015) mainly focused on the analysis of identical verse pairs. In this chapter, we extend our investigation to similar verse pairs in the Qur'an and perform an in-depth comparison across the three different translations of the Qur'an.

## 2.3 Methodology

This section presents the methodology that introduced in (Basharat, Yasdansepas, & Rasheed, 2015) for implementing the process of verse similarity evaluation for the different representations of the Qur'an. We also describe the five different similarity measures that we implemented for this study.

### 2.3.1 Dataset with Different Representation of the Qur'an

A Qur'an database was created for this study which contains four different representations of the Qur'an. Table 1 shows the information of the four representations and a sample text in each representation of the Qur'an.

Table 1: Dataset sample and description

| Abbreviation | Description | Example (first verse) |
|---|---|---|
| Q-Arabic | Arabic text without diacritics | "بسم الله الرحمن الرحيم" |
| Q-E-SAHIH | English text translated by the Saheeh team | "In the name of Allah the Entirely Merciful the Especially Merciful" |
| Q-E-YUSUF | English text translated by Yusuf Ali | "In the name of Allah\Most Gracious\ Most Merciful" |
| Q-Urdu | Urdu text | "شروع کا اللہ نام کے جو بڑا مہربان نہایت رحم والا ہے" |

### 2.3.2. Qur'an Text Preprocessing

Each dataset is complete raw data without any preprocessing; each text is simple and plain. Figure 1 shows the framework that has been designed in (Basharat, Yasdansepas, &

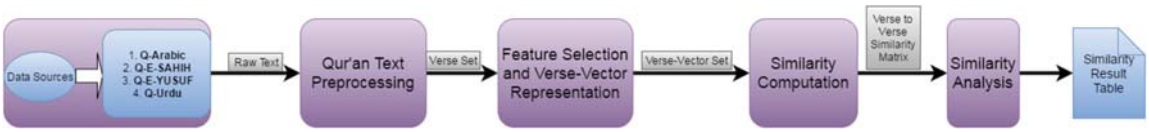


Figure 1: Methodology Flow Chart for Qur'an

Rasheed, 2015). The preprocessing step removes all the stop words and punctuations for each verse in the corresponding representation of the Qur'an.

### 2.3.3. Feature Selection and Verse-Vector Representation

According to our previous work in (Basharat, Yasdansepas, & Rasheed, 2015), assume that $V = \{v_1, v_2, v_3 ...v_n\}$ represents the set of verses in the Qur'an, and $T = \{t_1, t_2, ...t_m\}$ represents the set of the unique terms that construct each v in V. Each verse v is considered as an m-dimensional vector $\overline{v} = \{a_1, a_2, a_3 ...a_m\}$, where $a_i$ represents the weight of the $i_{th}$ term in the vector $\overline{v}$. The next step is to generate the verse-vector matrix. In the matrix, each row represents each verse of the Qur'an and each column represents a unique term of the vocabulary which constitutes the verses in Qur'an. The verse-vector matrix is structured based on the verse order from the Qur'an, and each element is calculated in accordance with one of two term weighting techniques: term frequency–inverse document frequency (TF-IDF) or Frequency (F). We keep these two out of the three term weighing techniques from (Basharat, Yasdansepas, & Rasheed, 2015). At first, the verse set is processed to select the features. After that, the verse-vector matrix where each verse is represented as weighted vector is generated based on the features.

### 2.3.4. Similarity Computation and Similarity Analysis

In the framework, the similarity computation is the most important step of our experiment. We use the similarity computation module to analyze the correlation between verses by calculating the distance or similarity of the corresponding vectors. More importantly, selecting

Table 2: Equations and brief description for 5 different similarity measures

| Similarity Measures | Equations and brief description |
|---|---|
| **Cosine Similarity:** it measures the similarity between two texts by obtaining the normalized their dot product. The range of Cosine similarity value is bounded in [0, 1]. | $S_C(\bar{v}_a, \bar{v}_b) = \frac{\bar{v}_a * \bar{v}_b}{\lvert\bar{v}_a\rvert * \lvert\bar{v}_b\rvert}$ |
| **Manhattan Distance:** it computes the sum of difference in each dimension of two vectors in n dimensional vector space. It is the sum of the absolute differences of their corresponding components (Hasnat, Halder, Hoque, Bhattacharjee & Nasipuri, 2013). We inverse the distance as $1/D_m$ to form the value bounded in [0, 1] | $D_M(\bar{v}_a, \bar{v}_b) = \sum_{i=1}^{n} \lvert t_{ai} - t_{bi}\rvert$ |
| **Pearson Correlation:** computes the linear correlation between two objects. It obtains the correlation coefficient by computing the ratio of the covariance of the two objects and the product of their standard deviations. The value is also bounded in [0, 1]. | $S_P(\bar{v}_a, \bar{v}_b) = \frac{m\sum_1^m t_{a1} \times t_{b1} - \sum_1^m t_{a1} \times \sum_1^m t_{b}}{\sqrt{(m\sum_1^m (t_{a1})^2 - (\sum_1^m t_{a1})^2) \times (m\sum_1^m (t_{b1})^2 -}}$ |
| **Hamming Distance:** The Hamming distance computes the minimum number of substitutions from one string changes to another string. For example: The Hamming distance between 1234567 and 1233497 is 3. We inverse the distance as 1/D_H to form the value bounded in [0, 1]. | $D_H(\bar{v}_a, \bar{v}_b) = \sum_{i=1}^{n} \lvert t_{ai} - t_{bi}\rvert$ (if $t_{ai}$ is different from $t_{bi}$, $t_{ai} - t_{bi} = 1$) |
| **Jaccard Similarity:** computes the similarity between sets. This method is defined as the quotient between the intersection and the union of the entities. The above equation is modified based on the general definition for different cases. The value range is in [0, 1]. | $s_J(\bar{v}_a, \bar{v}_b) = \frac{\bar{v}_a * \bar{v}_b}{\lvert\bar{v}_a\rvert^2 + \lvert\bar{v}_b\rvert^2 - \bar{v}_a * \bar{v}_b}$ |

an appropriate similarity computation measure is the most crucial part. The different vector representation also affects the similarity accuracy. Therefore, we believe that in order to obtain the best result, each combination of similarity measure and vector representation (Frequency and TF-IDF) should be applied to the verses in the Qur'an. Three similarity measures are described in (Huang, 2008) and (Strehl, Ghosh & Mooney, 2000), and two of them are described in (Hasnat,

Halder, Hoque, Bhattacharjee & Nasipuri, 2013) and (Bookstein, Kuliukin & Raita, 2012) respectively. Table 2 shows the relevant equation and brief description for each similarity measure. In the equations, $\overline{v}_a$ and $\overline{v}_b$ are the term vectors corresponding to the two verses $v_a$ and $v_b$ respectively. T = $\{t_1, t_2, ...t_m\}$ represents the weight of each term occurring in V.

The similarity computation step generates the verse to verse similarity matrix which contains all the similarity values of each verse pair. The similarity analysis module extracts the relevant similarity values from the matrix for analyzing the identical and similar verse pairs.

## 2.4. Experimentation and Results

### 2.4.1. Evaluation Measures for Identical Verses

The Q-Arabic dataset contains 775 identical verse pairs. Those verse pairs serve as the ground truth for our identical verse evaluation across the different translations. Each similarity ranges from 0 to 1. A value of 1 means the two verses are lexically identical. Our analysis first focuses on the verse pairs with similarity value 1.

For our analysis we still use the same evaluation metrics used in our previous work (Basharat, Yasdansepas, & Rasheed, 2015), including Precision, Recall and F1score. Equations (1-3) show the formulae of those three measures.

$$P(Precision) = \frac{TP}{TP + FP} \quad (1)$$

$$R(Recall) = \frac{TP}{TP + FN} \quad (2)$$

$$F1(F - Measure) = \frac{2 \times P \times R}{P + R} \quad (3)$$

TP represents true positive, TN represents true negative, FP represents false positive, FN represents false negative. In order to obtain the above three measures, we treat verse pairs with similarity value 1 which are identical verses in the Qur'an as TPs which are correctly classified by the model. We treat non-identical verse pairs with similarity value that is not equal 1 as TNs.

FPs are non-identical verse pairs classified as identical pairs (value 1). FNs are identical verse pairs classified as non-identical (value less than 1).

### 2.4.2. Experimental Results

In our experiments, four datasets including one original Arabic script and three translations of the Qur'an are analyzed. We apply five similarity measures to these datasets. In addition, two term weighting techniques have been used. Therefore, 4 * 5 * 2 = 40 combinations in total are implemented. Every experiment scheme is defined by an abbreviation as follows: dataset representation used - term weighting technique applied - similarity measure employed. For instance, Q-Urdu-M-F indicates Manhattan distance measure is implemented on the Q-Urdu dataset, and the term weighting technique is Frequency. All the experimental results are shown in Table 3.

### 2.4.3. Analysis for Identical Verses

The Arabic text of the Qur'an generates the most perfect experimental results when focusing on the identical verse pairs' similarity. This is clearly indicated by the precision and the recall values given in Table 3. In our previous study (Basharat, Yasdansepas, & Rasheed, 2015) we analyzed three different representations of the Arabic Qur'an namely: Arabic script with diacritics (case markings), an Arabic script without diacritics and a dataset that includes the Arabic word roots. We also included the Q-E-YUSUF in our comparative analysis. We presented an in-depth analysis for the identical verse pairs in this study. From this study, the Arabic script without diacritics provided the best results.

Therefore, the results obtained from this dataset were chosen to be the baseline results for the experimentation and results presented in this chapter. We focus our analysis on the relative comparison between the two English translations: Q-E-SAHIH and Q-E-YUSUF and the Urdu

translation of the Qur'an (Q-E-URDU). We also provide a comparative analysis across these different representations.

#### 2.4.3.1 Analysis on Q-E-YUSUF and Q-E-SAHIH and Comparison between Them

The Q-E-SAHIH and Q-E-YUSUF representations are two different translations of the Qur'an, translated by different English translators. Compared to the original Arabic representation, both produce a relatively lower recall of around 93% for Q-E-SAHIH representation and 85% for Q-E-YUSUF representation.

From the table, we find that both representations generate a considerable number of FN (false negatives), which is a significant indicator when it comes to the identical verses (Basharat, Yasdansepas, & Rasheed, 2015). As we have stated above, in terms of our experiments, false negatives represent identical verse pairs classified as non-identical. In the Q-E-YUSUF representation, the FNs reach up to 114 or 115, and it is exactly twice the FNs of the Q-E-SAHIH representation. That can be considered as a significant indicator to establish that the Q-E-SAHIH's translation quality is closer in terms of accuracy and precision to the original Arabic script, as compared to the Q-E-YUSUF. Finding the number of FNs and comparing it with the most original Arabic Qur'an script is one of the measures to evaluate the translation quality.

We found some verse pairs that are translated differently even though they are identical in the Arabic Qur'an script. Verse 3:182 and verse 8:51 are identical verses in the Q-Arabic representation. However, these two verses are appearing slightly different in the two English translations. In the Q-E-SAHIH translation, compared to verse 3:182, verse 8:51 has two extra words "of" and "devil" in the middle of the verse, which causes these two verses to be detected as non-identical by our framework. Likewise, in the Q-E-YUSUF translation, semantically, the verses describe the same content, however, the structure and the length of the verse is slightly

Table 3: Experiment Results

| | | Cosine | | Jaccard | | Pearson | | Manhattan | | Hamming | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | TF-IDF | F | TF-IDF | F | TF-IDF | F | TF-IDF | F | TF-IDF | F |
| **Q-ARABIC** | **TP** | 775 | 775 | 775 | 775 | 775 | 775 | 775 | 775 | 775 | 775 |
| | **FP** | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 |
| | **FN** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | **P** | 0.997 | 0.999 | 0.997 | 0.999 | 0.997 | 0.999 | 0.997 | 0.999 | 0.997 | 0.999 |
| | **R** | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| | **F** | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 |
| **Q-E-SAHIH** | **TP** | 718 | 718 | 718 | 718 | 718 | 718 | 718 | 718 | 718 | 718 |
| | **FP** | 12 | 11 | 12 | 11 | 12 | 11 | 12 | 11 | 12 | 11 |
| | **FN** | 57 | 57 | 57 | 57 | 57 | 57 | 57 | 57 | 57 | 57 |
| | **P** | 0.985 | 0.985 | 0.985 | 0.985 | 0.985 | 0.984 | 0.985 | 0.985 | 0.985 | 0.985 |
| | **R** | 0.926 | 0.926 | 0.926 | 0.926 | 0.926 | 0.926 | 0.926 | 0.926 | 0.926 | 0.926 |
| | **F** | 0.955 | 0.955 | 0.955 | 0.955 | 0.955 | 0.954 | 0.955 | 0.955 | 0.955 | 0.955 |
| **Q-E-YUSUF** | **TP** | 661 | 660 | 661 | 660 | 661 | 660 | 661 | 660 | 661 | 660 |
| | **FP** | 12 | 10 | 12 | 10 | 12 | 10 | 12 | 10 | 12 | 10 |
| | **FN** | 114 | 115 | 114 | 115 | 114 | 115 | 114 | 115 | 114 | 115 |
| | **P** | 0.982 | 0.985 | 0.982 | 0.985 | 0.982 | 0.985 | 0.982 | 0.985 | 0.982 | 0.985 |
| | **R** | 0.853 | 0.852 | 0.853 | 0.852 | 0.853 | 0.852 | 0.853 | 0.852 | 0.853 | 0.852 |
| | **F** | 0.913 | 0.913 | 0.913 | 0.913 | 0.913 | 0.913 | 0.913 | 0.913 | 0.913 | 0.913 |
| **Q-URDU** | **TP** | 535 | 535 | 535 | 535 | 535 | 535 | 535 | 535 | 535 | 535 |
| | **FP** | 7 | 6 | 7 | 6 | 7 | 6 | 7 | 6 | 7 | 6 |
| | **FN** | 240 | 240 | 240 | 240 | 240 | 240 | 240 | 240 | 240 | 240 |
| | **P** | 0.987 | 0.989 | 0.987 | 0.989 | 0.987 | 0.989 | 0.987 | 0.989 | 0.987 | 0.989 |
| | **R** | 0.690 | 0.690 | 0.690 | 0.690 | 0.690 | 0.690 | 0.690 | 0.690 | 0.690 | 0.690 |
| | **F** | 0.812 | 0.813 | 0.812 | 0.813 | 0.812 | 0.813 | 0.812 | 0.813 | 0.812 | 0.813 |

different. Besides this sample case, we also found another two verses, which are translated identically in one of the English translations, but differently in the other one. Verse 54:17 and verse 54:22 are identical verses in the Q-Arabic representation. The Q-E-SAHIH translated these two verses to identical English verses. However, in the Q-E-YUSUF, the corresponding translated verses are not the same because the first initial is "and" in the verse 54:17, however, it is the word "but" in the verse "54:22". Similarly, verse 7:108 and 26: 33 are translated to the same English verses in Q-E-YUSUF, but to different English verses in Q-E-SAHIH with two extra words occurring in verse 26:33.

FP (false positives) is another significant indicator to analyze the identical verses. From the perspective of our experiments, false positive cases represent non-identical verse pairs which are classified as identical. In the result table, with different term weighting techniques, 12 or 10 FPs are obtained in the Q-E-YUSUF and 12 or 11 FPs in the Q-E--SAHIH. Both versions contain much fewer FPs as compared to the number of FNs. Arabic is a rich morphological language; in particular, the language of the Qur'an is considered to be precise and the slightest variation or alteration in the arrangement or morphological manifestation of the word implies something significant, which the translation often fails to capture (Basharat, Yasdansepas, & Rasheed, 2015). Table 4 shows three cases, which reflect the difference for specific verse pairs among the three representations of the Qur'an. Case 1 demonstrates two verses which are not identical in the original Arabic script but are translated identically in the two English Qur'an translations. The reason for the difference is that the word "إنا" is widely used in Arabic as an intensifier to emphasize the speech. Verse 37:110 includes this word, but it is missing in verse 37:121. However, this difference is not captured by the two English translations. Similarly, Case 3 shows another two non-identical verses in Arabic. The word "و" is included in verse 20:9, but is not appearing in verse 79:15. This character is a conjunction word in Arabic referring to "and" in English. As a result, the Q-E-SAHIH translation correctly captures this difference; however, it has been ignored in the Q-E-YUSUF. Case 2 shows two verses constructed with same words in the Arabic Qur'an, but three of these words are in different orders within the two verses, which results in the two verses not being lexically identical. This is captured by Q-E-YUSUF translation, but not by the Q-E-SAHIH translation.

For our experiments, we intentionally used two versions of English translation in order to illustrate that the various versions of the Qur'an translated in one language will invariably have

Table 4: 3 cases for analyzing the false positives on three Qur'an representations

| **Case 1** | **Verse Text(Q-Arabic)** | **Verse Text(Q-E-SAHIH)** | **Verse Text(Q-E-YUSUF)** |
|---|---|---|---|
| 37:110 | "إنا كذلك نجزي المحسنين" | "Indeed We thus reward the doers of good" | "Thus indeed do We reward those who do right." |
| 37:121 | "كذلك نجزي المحسنين" | "Indeed We thus reward the doers of good" | "Thus indeed do We reward those who do right." |
| **Case 2** | **Verse Text(Q-Arabic)** | **Verse Text(Q-E-SAHIH)** | **Verse Text(Q-E-YUSUF)** |
| 23:83 | " لقد وعدنا نحن وآباؤنا هذا من قبل إن هذا إلا أساطير الأولين" | "We have been promised this we and our forefathers before this is not but legends of the former peoples" | "Such things have been promised to us and to our fathers before! they are nothing but tales of the ancients! " |
| 27:68 | " لقد وعدنا هذا نحن وآباؤنا من قبل إن هذا إلا أساطير الأولين" | "We have been promised this we and our forefathers before this is not but legends of the former peoples" | "It is true we were promised this we and our fathers before (us): these are nothing but tales of the ancients." |
| **Case 3** | **Verse Text(Q-Arabic)** | **Verse Text(Q-E-SAHIH)** | **Verse Text(Q-E-YUSUF)** |
| 20:9 | "وهل أتاك حديث موسى" | "And has the story of Moses reached you" | "Has the story of Moses reached thee?" |
| 79:15 | "هل أتاك حديث موسى" | "Has there reached you the story of Moses" | "Has the story of Moses reached thee?" |

subtle differences. In addition, this analysis also proves that our framework is helpful in evaluating the translation's quality and investigating the comparison and contrast among different versions of the Quran that are translated in one language.

**2.4.3.2 Analysis of Identical Verses in Q-URDU**

The result table shows that the Q-URDU representation obtains a recall of around 69%, which is much lower than the other three representations. Besides that, the F1 score is also the lowest compared to the others. Similarly, this representation generates both FNs and FPs. The

number of FNs is almost one third of the number of identical verse pairs. On the other hand, the number of FPs is smaller than the numbers generated by the two English representations.

For the Q-URDU representation, 240 identical verse pairs are classified mistakenly as non-identical. 240 is a big number in terms of the influence of FNs reflecting the relative precision or accuracy of the translation relative to the original script. Within the 240 false negatives, most verse pairs have a fairly common small difference which causes the misclassification. For instance, verses 5:10 and 5:86 are identical in the Arabic script but not identical in the Urdu version. The difference is strongly visible, where the word "جنہوں" in verse 5:10 is different from the "جن لوگوں" in verse 5:86. However, "جنہوں" is another way of expressing "جن لوگوں", so basically this change doesn't affect the semantics of the verse. It indicates that the translator has just expressed the same meaning with a different expression. Another representative verse pair is indicated by the verse 7:121 and the verse 26:47. There are a few differences between them. First, the word "تمام" is absent in verse 7:121, and present in verse 26:47, but this is a minor difference. Second, the word "پروردگار" used in verse 7:121 means Lord or God which accurately captures the Arabic meaning; the word "مالک " used in verse 26:47 means King or owner which is less accurate. Nevertheless, the semantic content of both verses is still same. Third, the word "لائے " and "آئے لے " are separately used in two verses. Again, these two words are generally still meaning the same thing in the two verses. From the analysis of the 240 FNs, we can deduce that the differences in translation of identical verses occur for two main reasons. Firstly, Urdu has a tendency to be a more verbose language compared to Arabic. The same verse expressed using few words in Arabic requires more words in Urdu in order to convey the meaning. Also, the choice of prepositional and conjunction words can cause lexical differences in the expression, while maintaining the semantics. The second reason is that of

Table 5: 2 Cases for analyzing the false positives from Q-URDU

| **Case 1** | **Verse Text(Q-Arabic)** | **Verse Text(Q-Urdu)** |
|---|---|---|
| 37:34 | "إنا كذلك نفعل بالمجرمين" | "ہم گنہگاروں کے ساتھ ایسا ہی کیا کرتے ہیں" |
| 77:18 | "كذلك نفعل بالمجرمين" | " ہم گنہگاروں کے ساتھ ایسا ہی کیا کرتے ہیں" |
| **Case 2** | **Verse Text(Q-Arabic)** | **Verse Text(Q-Urdu)** |
| 52:11 | "فويل يومئذ للمكذبين" | "اس دن جھٹلانے والوں کے لئے خرابی ہے" |
| 77:15 | "ويل يومئذ للمكذبين" | "اس دن جھٹلانے والوں کے لئے خرابی ہے" |

using a different synonym for a significant word, which may imply subtle semantic differences. This may also be reflective of the precision of the translation. Whether identical verse pairs ought to be translated differently or not is a question in itself.

As we can see from the table, the Q-Urdu representation also generates some FPs (i.e. verse pairs which are different in the original Arabic script but found to be identical in Q-Urdu). Table 5 shows two such FP cases. In case 1, there is an additional word at the beginning of verse 37:34 in Arabic, which is present for emphasis. However, this additional emphasis is not reflected in the translation for both verses. Case 2 shows a rather subtler difference of a single character at the beginning of verse 52:11. However, it should have been captured by the translator, but was not. The analysis of the FPs provides a good basis for determining cases where subtle differences in the Arabic script are not reflected in the translation.

**2.4.3.3 Comparative Analysis across Different Languages**

We considered an additional thread of analysis across the three different translations of the Qur'an by finding common FNs or FPs among them. Table 6 shows the number of common FNs or FPs between two or three different translations of the Qur'an. For the convenience of analysis, all the statistics from Table 6 are based on the FNs and FPs which are generated by the Cosine method with frequency as term weighing measure. Since the generated FPs from each translation are few, as shown in Table 6, the numbers of common FPs between various

Table 6: The FPs and FNs in common between relevant translations

| Comparative Cases | FPs in common | FNs in common |
|---|---|---|
| 1. Q-E-YUSUF vs Q-E-SAHIH | 3 | 46 |
| 2. Q-E-YUSUF vs Q-URDU | 1 | 99 |
| 3. Q-E-SAHIH vs Q-URDU | 1 | 56 |
| 4.Q-E-YUSUF vs Q-E-SAHIH vs Q-URDU | 1 | 45 |

translations are much smaller than FNs. Between the two English translations, there are 3 verse pairs which are FPs for both translations. They are 37:80&37:110, 37:110&37:121 and 37:110&37:131. If we look more closely at these verse pairs, all of them involve the same verse 37:110. The original verse 37:110 is composed of 3 simple Arabic words. However, the other 3 verses which are similar to verse 37:110 include one extra emphasis word "إنا" at the beginning of the verse in the Arabic Qur'an script. Since this difference is not reflected in the English translations, the 3 verse pairs become false positives in both Q-E-SAHIH and Q-E-YUSUF. Interestingly, the only one common FP between all three translations is 37:80&37:110. However, the difference in the other two verse pairs is captured by the Urdu translation. It is not clear why that difference is ignored in one of the verse pairs in the Urdu translation. It is nevertheless obvious that the number of false positives falls within a small range. The major reason that the false negatives still exist is because the translator more or less ignored the emphasis word from some original Arabic verses.

Indeed, each translation of the Qur'an generated many FNs, especially in the Q-URDU. In case 1, the common FNs are 46 which is almost 80% of FNs of Q-E-SAHIH. There are another 11 FNs unique to Q-E-SAHIH and 69 FNs unique to Q-E-YUSUF. In case 2, the common FNs are 99, which is almost 86% of FNs of Q-E-YUSUF. There are another 16 unique FNs to Q-E-SAHIH and 141 unique to Urdu Qur'an. In case 3, the common FNs are 56; the FNs of Urdu include all the FNs generated by Q-E-SAHIH. From these three cases, we may draw at

least one conclusion that there are many FNs in common between the different translations. In other words, a certain number of identical verse pairs in the original Qur'an are consistently translated as different verse pairs by the translators with different reasons. The case 4 is an example, 45 common FNs among three translations. Compare this number with the number in case 1, only one difference, which means there is one identical verse pair correctly captured by the translators of Urdu Qur'an, but not captured by the translators of both English translations. One of the reasons behind this could be the criteria of translation set up by different language. Therefore, based on case 4, if we add one more translation of the Qur'an, the number of common FNs may be fewer than 45 because of the new criteria of the additional translation.

Finally, if we evaluate the quality of a translation of the Qur'an based on recall, F1 score or the number of FNs or FPs without considering the language itself and the criteria of translation set up by different languages, the Q-E-SAHIH definitely is the closest to the Arabic Qur'an based on the comparative analysis on identical verse pairs.

### 2.4.4 Analysis for Similar Verses

Although, much of our analysis is based on the identical verse pairs in the Qur'an, these pairs are a handful compared to the total number of verse pairs. There are several verse pairs, which are similar to a great extent but not identical. It is therefore worthwhile to explore the similarities of those verse pairs. We aim to establish the effectiveness of the various similarity measures employed in our experiments for the analysis of similar verses across the difference representations of the Qur'an.

One of the challenges with analyzing similar verse pairs, unlike the identical verse pairs, is the lack of any baseline standard similarity values or thresholds. It is therefore impossible to investigate and analyze all similar verse pairs; instead, we manually selected a few representative

Table 7: Selected verse for similarity analysis

| Verses | Verse Text (Q-Arabic) |
|---|---|
| 1:3 | "الرحمن الرحيم" |
| 1:1 | "بسم الله الرحمن الرحيم" |
| 27:30 | "إنه من سليمان وإنه بسم الله الرحمن الرحيم" |
| 11:61 | وإلى ثمود أخاهم صالحا ۚ قال يا قوم اعبدوا اللـه ما لكم من إله غيره ۖ هو أنشأكم من الأرض واستعمركم فيها فاستغفروه ثم توبوا إليه ۚ إن ربي قريب مجيب |
| 11:84 | وإلى مدين أخاهم شعيبا ۚ قال يا قوم اعبدوا اللـه ما لكم من إله غيره ۖ ولا تنقصوا المكيال والميزان ۚ إني أراكم بخير وإني أخاف عليكم عذاب يوم محيط |

verses to study from the Arabic Qur'an. As shown in Table 7, it is clear that the first three verses are similar to each other. Specifically, verse 1:1 contains all the content of verse 1:3, and verse 1:27 contains all the content of verse 1:1. The last two verses lexically contain a portion of common words. In looking at the first three verses from Table 7, we clearly see that the content of verse 1:3 is around half that of verse 1:1 and around a quarter of that of verse 1:27; verse 1:1 is also around half of verse 1:27. Table 8 shows relevant similarity values obtained for these verse pairs across the different representations of the Qur'an. For the Q-Arabic version, the Jaccard method gives the most logical similarity values with both F and TFIDF term weighing measures for these three pairs. Since the Jaccard method is one of the term-based similarity measures, it is not surprising for it to return the best results. We also studied the other four methods, and compared the relevant similarity values to that of Jaccard's. The Hamming method returns the same results with the two different term weighting measures, which are close to the Jaccard's evaluations, especially on verse pair 1:1 and 27:30. The Manhattan method obtained the same results as the Hamming's with Frequency as weighting measure and different results with TFIDF as the weighting measure, but overall, the results still approach the Jaccard's values.

On the other hand, the results obtained with the other two weighting measures, Pearson and Cosine, are not reflecting the observed similarity values.

The values obtained for these two methods vastly exceed the ground truth values. Contrary to the first three verses in Table 7, the last two verses do not contain only one contiguous common portion but rather a few common portions. Therefore, it cannot be intuitively evaluated. However, after analyzing the first 3 verses, we conclude that Jaccard method with Frequency as the term weighing measure is the most suitable combination to evaluate our cases. Look into this pair, Jaccard with F returns 0.28, which we consider intuitively close to the true value. Among the other combinations, only Cosine method with TFIDF and Jaccard method with TFIDF are close to this result, so we conclude that the rest of the combinations do not reflect the true similarities. To sum up, the Jaccard method with F as the weighting measure can objectively and appropriately reflect the intuitive values of all cases based on our selected verses. The Manhattan, Hamming and Cosine methods are applicable to some cases. As for the Pearson method, the overall evaluation results are much higher than any intuitive reasonable values.

As for the other language representations, for the first three verses, the relevant similarity values generated by the Jaccard method with F in the Q-E-SAHIH are most similar to that of Arabic. The relevant similarity values from the other Qur'an representations also reflect the same relationships among these three verses. On the other hand, for verses 11:61 and 11:84, the relevant similarity values are 0.37, 0.4, and 0.35 respectively from the three representations, which are all close. Therefore, from what has been discussed above, considering the similarity value comparison of the given cases, we may reasonably conclude that the translation accuracy for similar verses of Q-E-SAHIH is better than that of other representations. Also, its translation

Table 8: Relevant similarity values for selected verse pairs

| Q-ARABIC | **Manhattan** | | **Hamming** | | **Jaccard** | | **Pearson** | | **Cosine** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | F | TFIDF | F | TFIDF | F | TFIDF | F | TFIDF | F | TFIDF |
| 1:3 VS 1:1 | 0.66 | 0.38 | 0.66 | 0.66 | 0.5 | 0.53 | 0.85 | 0.84 | 0.71 | 0.68 |
| 1:3 VS 27:30 | 0.39 | 0.23 | 0.39 | 0.39 | 0.25 | 0.28 | 0.75 | 0.75 | 0.5 | 0.51 |
| 1:1 VS 27:30 | 0.48 | 0.28 | 0.48 | 0.48 | 0.5 | 0.53 | 0.85 | 0.87 | 0.71 | 0.75 |
| 11:61 VS 11:84 | 0.18 | 0.09 | 0.18 | 0.18 | 0.28 | 0.21 | 0.73 | 0.64 | 0.46 | 0.29 |
| Q-E-YUSUF | **Manhattan** | | **Hamming** | | **Jaccard** | | **Pearson** | | **Cosine** | |
| | F | TFIDF | F | TFIDF | F | TFIDF | F | TFIDF | F | TFIDF |
| 1:3 VS 1:1 | 0.43 | 0.40 | 0.43 | 0.43 | 0.18 | 0.58 | 0.89 | 0.89 | 0.74 | 0.79 |
| 1:3 VS 27:30 | 0.27 | 0.22 | 0.28 | 0.28 | 0.24 | 0.30 | 0.77 | 0.77 | 0.53 | 0.54 |
| 1:1 VS 27:30 | 0.34 | 0.28 | 0.36 | 0.36 | 0.53 | 0.51 | 0.84 | 0.84 | 0.72 | 0.68 |
| 11:61 VS 11:84 | 0.14 | 0.09 | 0.14 | 0.15 | 0.37 | 0.26 | 0.64 | 0.64 | 0.62 | 0.29 |
| Q-E-SAHIH | **Manhattan** | | **Hamming** | | **Jaccard** | | **Pearson** | | **Cosine** | |
| | F | TFIDF | F | TFIDF | F | TFIDF | F | TFIDF | F | TFIDF |
| 1:3 VS 1:1 | 0.43 | 0.40 | 0.43 | 0.48 | 0.55 | 0.66 | 0.95 | 0.93 | 0.87 | 0.85 |
| 1:3 VS 27:30 | 0.26 | 0.22 | 0.28 | 0.29 | 0.3 | 0.36 | 0.84 | 0.80 | 0.67 | 0.61 |
| 1:1 VS 27:30 | 0.32 | 0.26 | 0.36 | 0.36 | 0.55 | 0.55 | 0.89 | 0.86 | 0.77 | 0.71 |
| 11:61 VS 11:84 | 0.15 | 0.10 | 0.12 | 0.17 | 0.40 | 0.27 | 0.84 | 0.64 | 0.68 | 0.28 |
| Q-URDU | **Manhattan** | | **Hamming** | | **Jaccard** | | **Pearson** | | **Cosine** | |
| | F | TFIDF | F | TFIDF | F | TFIDF | F | TFIDF | F | TFIDF |
| 1:3 VS 1:1 | 0.34 | 0.24 | 0.34 | 0.34 | 0.38 | 0.46 | 0.81 | 0.84 | 0.62 | 0.68 |
| 1:3 VS 27:30 | 0.22 | 0.18 | 0.24 | 0.24 | 0.21 | 0.33 | 0.70 | 0.78 | 0.41 | 0.57 |
| 1:1 VS 27:30 | 0.27 | 0.22 | 0.28 | 0.29 | 0.48 | 0.59 | 0.85 | 0.86 | 0.71 | 0.73 |
| 11:61 VS 11:84 | 0.12 | 0.08 | 0.11 | 0.14 | 0.35 | 0.24 | 0.82 | 0.64 | 0.63 | 0.29 |

quality is the best among the three translations. Overall, the pattern of similarity is maintained in the other three representations of the Qur'an, which are based on translations. This pattern follows the similarity patterns of the original Arabic script. However, determining the extent to which this conclusion can be generalized needs further investigation and validation of the ground truth measures.

## 2.5 Conclusion and Future Work

In this chapter, we investigated the effect of five similarity measures across four different representations of the Qur'an, in three different languages. We analyzed the results for identical and similar pairs of verses within these representations. We concluded that the Q-E-SAHIH representation demonstrates the most accurate result with highest F1 score among the different translations of the Qur'an. We also concluded that the Jaccard similarity method proves to be effective for each of our tested verse pairs. In addition, the similarity values returned by the Jaccard measure intuitively reflect the observed similarities. We also found that the verse pairs which are similar in the original Arabic script are more or less lexically similar in the three translations.

Regarding future work, this research can be improved and extended in various aspects. First, we can apply the framework developed as part of our research for analyzing similarities within and across other religious texts such as the Bible, to see how the approach scales for larger texts. Second, we aim to adopt more term weighting methods for our future experiments, such as LTC, and relative frequency which are introduced in (Khorsheed & Thubaity, 2013). Finally, we plan to reduce the time complexity of future experiments by applying more efficient data structures or algorithms to reconstruct our term weighing matrix in the data preprocessing stage, or the whole process, using parallel processing techniques.

## 2.6 References

A. Basharat., D. Yasdansepas., & K. Rasheed. (2015). Comparative study of verse vimilarity for multi-lingual representations of the qur'an. Proceedings on the International Conference on Artificial Intelligence (ICAI).

A. Bookstein., V. A. Kuliukin.,& T. Raita. (2012, Octorber). Generalized Hamming distance. Information Retrieval Journal. Vol 5, pp.353-375.

A. Hasnat., S. Halder., A. Hoque., D. Bhattacharjee., & M. Nasipuri. (2013, May -June). A fast fpga based architecture for measuring the distance between two color images using Manhattan.

International Journal of Electronics and Communication Engineering & Technology(IJECET). vol. 4, pp. 01–10.

A. Huang.(2008). Similarity measures for text document clustering. in Pro- ceedings of the sixth new zealand computer science research student conference (NZCSRSC2008). pp.49-56. Christchurch, New Zealand.

A. Strehl., J. Ghosh., & R. Mooney. (2000). Impact of similarity measures on web-page clustering. in Workshop on Artificial Intelligence for Web Search (AAAI 2000). pp. 58–64.

A.-B. M. Sharaf., & E. Atwell. (2012). Qursim: A corpus for evaluation of relatedness in short texts. in LREC, pp. 2295–2302.

E. Atwell., N. Habash., B. Louw,. B. Abu Shawar., T. McEnery., W. Zaghouani., & M. El-Haj. (2010). Understanding the quran: A new grand challenge for computer science and artificial intelligence. ACM-BCS Visions of Computer Science 2010.

E. Atwell., C. Brierley., K. Dukes., M. Sawalha., & A.-B. Sharaf. (2011). An artificial intelligence approach to arabic and islamic content on the internet. Proceedings of NITS 3rd National Information Technology Symposium.

M. A. Sherif., & A.-C. Ngonga Ngomo.(2009). Semantic quran - a multilingual resource for natural-language processing. Semantic Web.

M. S. Khorsheed., & A. O. Al-Thubaity. (2013). Comparative evaluation of text classification techniques using a large diverse arabic dataset. Language resources and evaluation. vol. 47, no. 2, pp. 513-538.

R. S. Forsyth., & S. Sharoff. (2014). Document dissimilarity within and across languages: A benchmarking study. Literary and Linguistic Computing. vol. 29, no. 1, pp. 6–22.

W. H. Gomaa, & A. A. Fahmy. (2013, April). A survey of text similarity approaches. International Journal of Computer Applications(0975-8887) vol. 68 -No.13 pp. 01-06.

# CHAPTER 3

# INTERLINKING HADITH BASED ON MULTILINGUAL TEXT SIMILARITY ANALYSIS[2]

[2] Pan Huang, Amna, Basharat, Usman Nisar, Khaled Rasheed. To be Submitted to The 15th IEEE International Conference on Machine Learning and Applications (IEEE ICMLA'16), 2016.

## 3.1 Abstract

With the development of data technology and science, the size and volume of text data has been increasingly growing in all fields. Similarity analysis of large texts has been receiving much attention. Hadith, is the collection of the sayings of the Prophet Muhammad or the reports about what he did, which contains thousands of texts. In this chapter, based on our previous text similarity research on the Qur'an, we utilize our previously developed similarity computation and analysis methodology for the larger text data that the Hadith comprises. We employ multithreading technique for speeding up the similarity computations in each representation of the Hadith. We compare and contrast the application of similarity measures across the English and Arabic representations of the Hadith. Based on the results of our text similarity analysis, we propose interlinking of Hadiths with similar semantic content by investigating different equivalence classes by applying different similarity thresholds.

## 3.2 Introduction

With the exponential growth of textual information (Harrag, Cherif & Qawasmed, 2008), text similarity analysis has become a focal point in many research areas. Text is composed of words, and two texts can be lexically or semantically similar. A similarity measure or distance is used to measure the similarity between texts, and a reasonable analysis may be established on the similarity result. In this chapter, we utilize the similarity computation methodology previously introduced in chapter 2 to apply on the larger text data that the Hadith comprises. We apply this computation on the Arabic and the English version of the Hadith. We employ specific techniques for speeding up the similarity computations in each representation of the Hadith. Based on our similarity findings, we interlink the Hadiths with similar semantic content in the Arabic representation of the Hadith.

### 3.2.1. Motivation and Related Work.

The Sunnah of the Prophet Muhammad (SAW) is the second of the two revealed fundamental knowledge sources of Islam (Sharaf & Atwell, 2012). A Hadith a narration about the sayings of the Prophet Muhammad or a report about what he did (Atwell, Habash, Louw, Shawar, McEnery, Zaghouani, & ElHaj, 2010). The Hadith collection contains thousands of texts. Hadith forms one of the basis and foundations of the larger body of the Islamic texts and has been the subject of research and intensive study. In (Harrag, Cherif & Qawasmed, 2008), the researchers present an information retrieval architecture for the text mining of Hadith. They develop a novel automatic text miming search tools which is based on the vector space model to classify the Hadiths in accordance with degrees of similarity. In (Sharaf & Atwell, 2012), the researchers use the several data mining techniques such as neural networks, decision trees to carry out knowledge extraction in Hadith. Moreover, in (A.Basharat., B.Abro., I.B.Arpinar., & K.Rasheed, 2016), a semantic Hadith framework has been proposed for interlinking the most important Islamic knowledge sources through the application of the linked data standards (A.Basharat., B.Abro., I.B.Arpinar., & K.Rasheed, 2016).

### 3.2.2. Previous Work.

We previously undertook two studies of text similarity analysis on the Qur'an respectively in chapter 2 and (Basharat, Yasdansepas, & Rasheed, 2015). First of all, we compare the similarity of the verses in the Qur'an across the Arabic representation of the Qur'an and the English representation of the Qur'an in chapter 2. The original methodology of verse similarity analysis was designed and developed in chapter 2. This study was mainly focused on the identical verses in the Qur'an. Based on the similarity analysis in chapter 2, it was concluded that the number of identical verse pairs could be relying on the precision and accuracy of a given

representation of Qur'an. The study presented in (Basharat, Yasdansepas, & Rasheed, 2015) is an extension of our investigation on not only the verse similarity of identical verse pairs but also the verse pairs with high similarity degree. Also, more translations of different language are analyzed through the application of different similarity measures. It was concluded that the similarity patterns through in different representation of the Qur'an are more or less similar.

### 3.2.3. Contributions of this chapter.

In this chapter we make the following contributions:

- We reduce the time complexity of the framework which is designed in previous work (Basharat, Yasdansepas, & Rasheed, 2015) such that it can handle lager data. Specifically, the Hash map for storing the text-vector Hadith and the application of multithreading technique for the framework significantly speeds up the efficiency of similarity computation.
- We compare and contrast the application of similarity measures across the English and Arabic representations of the Hadith.
- We propose interlinking of Hadiths with similar semantic content by investigating different equivalence classes by applying different similarity thresholds.

## 3.3 Methodology

The methodology of this research is derived from our previous work chapter 2 and (Basharat, Yasdansepas, & Rasheed, 2015). Based on the similarity framework designed previously, we improved the functions on several stages of the framework by applying a more efficient data structure to store the processed data, and a multithreading technique is utilized for speeding up the similarity calculations. Figure 2 shows the extension of the framework that has been designed in chapter 2 and (Basharat, Yasdansepas, & Rasheed, 2015) and developed for this research.

Table 9: Date sample and description

| Abbreviation | Description | Example Text |
|---|---|---|
| H-Arabic | Arabic Hadith text without diacritics | حدثني يحيى، عن مالك، عن محمد بن أبي بكر الثقفي، أنه سأل أنس بن مالك - وهما غاديان من منى إلى عرفة - كيف كنتم تصنعون في هذا اليوم مع رسول الله صلى الله عليه وسلم قال كان يهل المهل منا فلا ينكر عليه ويكبر المكبر فلا ينكر عليه . |
| H-English | English Hadith text | Narrated Muhammad bin Abu Bakr Al-Thaqafi: I asked Anas bin Malik while we were proceeding from Mina to `Arafat, "What do you use to do on this day when you were with Allah's Apostle?" Anas said, "Some of us used to recite Talbiya and nobody objected to that, and others used to recite Takbir and nobody objected to that." |

### 3.3.1 Two Representations of Hadith

The Hadith data has been obtained from sunnah.com, which includes Hadith data in English and in Arabic. Other than Hadith text, the dataset also contains the information about each Hadith, such as the URN number, the matching URN number of another language of Hadith and so on. The URN can be considered as the ID for each Hadith. Each Hadith could have only on URN number. Each dataset contains a special feature that the matching URN number indicates URN number in another language of Hadith representation with the same Hadith. For example, the URN number of Hadith A in the English representation is “207420”, so we can easily find out the corresponding Arabic Hadith text in the Arabic representation by looking at the matching Arabic URN number from English representation. However, two dataset contains different number of Hadith. The Arabic representation contains 25932 Hadiths, and the English representation contains 18040 Hadiths. Therefore, in order to keep a unified scheme so as to be able to compare and contrast the similarity, stage 1 extracts common Hadith from both representation of Hadith with same sequence in each file as the basis for analysis. After

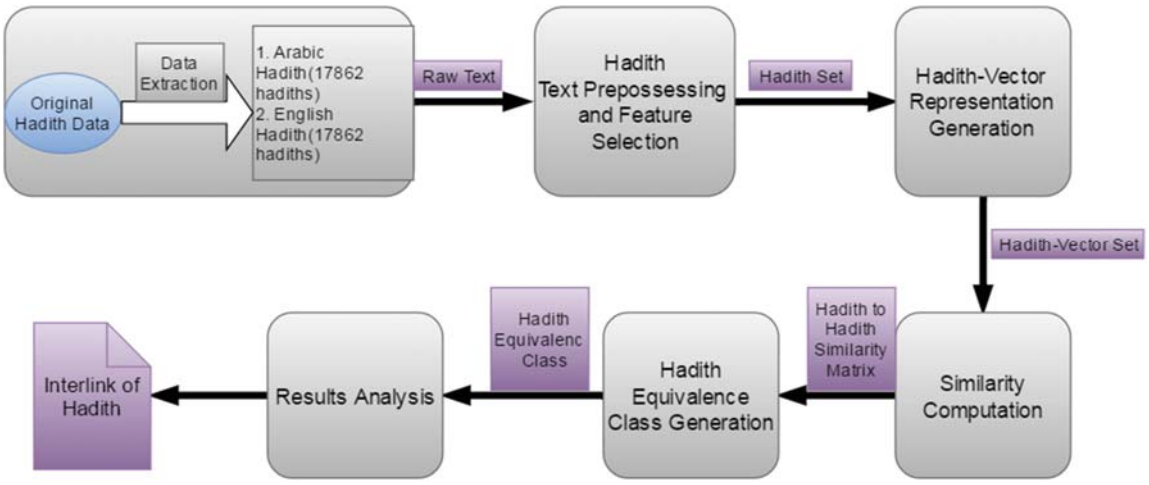


Figure 2: Methodology Flow Chart for Hadith

extraction, both datasets contain 17862 Hadiths. Table 9 shows an example Hadith of different representation in each dataset.

### 3.3.2. Hadith Text Preprocessing and Feature Selection

This stage preprocesses the raw text for the input data representation, which includes three steps. Firstly, it removes all the non-target language characters, such as numbers, special characters. Secondly, the tokenization process is carried out such that it chops each input character sequence into tokens, which are applied for each Hadith (instance) in the dataset. The punctuation is thrown away by this step. The generated tokens which are referred to as terms or words are stored into the term dictionary. Each term or word in the dictionary is unique and constructed by the target language. Thirdly, because the term in dictionary compose the attributes list for the term vector matrix in stage 3, without affecting the accuracy of similarity computation, the lesser the terms in the dictionary, the smaller the size of the matrix. Therefore, this step removes all the terms whose frequency of occurrence are less than five in the entire

Hadith file from the dictionary. After having done that, the number of terms in the dictionary is drastically decreased.

### 3.3.3. Hadith-Vector Representation Generation

Compare this stage with that of our previous work in chapter 2 and (Basharat, Yasdansepas, & Rasheed, 2015), we modified the structure of the term vector matrix. According to our previous work in chapter 2, $H = \{h_1, h_2, h_3 ...h_n\}$ represents the set of Hadiths in the Hadith text, $T = \{t_1, t_2, ...t_m\}$, as the dictionary generated from stage 2, to construct each Hadith. In the Hadith term vector matrix, each Hadith is constructed by an m-dimensional vector $\overline{v} = \{a_1, a_2, a_3 ...a_m\}$. If term i is not appearing in a Hadith, the $a_i$ is 0. Previously, in the matrix, each row represents each Hadith in the Hadith text, and are supposed to be stored sparsely because of a large number of 0s in the matrix. However, during this research, instead, we store each term vector of Hadith in non-sparse way. As a result, this modification saves massive space for our framework. Additionally, instead of using the normal data structure to store the term vector matrix, we import an efficient data structure Hash maps to store the term vector of each Hadith in Key-Value format. Specifically, Hash maps are data structures widely used in modern programming languages like Java for their simplicity and efficiency (Topac, 2015). We believe the specialty of Hash map is helpful for Hadith searching, which speeds up the similarity computation. Based on the previous work in chapter 2, we apply the TF-IDF (term frequency–inverse document frequency) and F(Frequency) as the term weighting techniques to represent each vector in the matrix. In brief, this stage generates the term-Hadith matrix where each Hadith is represented by weighted vectors.

### 3.3.4. Brief Introduction of Similarity Measures and Similarity Computation.

We previously compared and contrasted the effect of applying five similarity measures across four semantic representation of the Qur'an in chapter 2. These five similarity measures are Cosine similarity, Manhattan distance, Pearson correlation, Hamming distance and Jaccard similarity. From the analysis in the previous work, the Jaccard method proves to effective for all the test cases; the Cosine, Hamming, Manhattan method are effective for many test cases; the Pearson method are proved not effective and applicable for most of cases. Thus, based on the summary of similarity measures' comparison from previous research, we discard the Pearson method, and keep one method between Hamming and Manhattan since both of them belong to distance measure. Thus, three similarity measures used for implementing our experiments in this chapter. The original description of Cosine and Jaccard method can be found in (Huang, 2008) and (Strehl, Ghosh & Mooney, 2000), and the basic definition of Hamming method can be found in (Bookstein, Kuliukin & Raita, 2012). Please see Table 2 for the fundamental concept and equations of relevant methods.

Specifically, another important modification that we developed is that we import the multithreading technique into the similarity computation stage. A thread is considered as a call sequence which executes independently. Also, the smallest executed instructions in a program is the thread. In our experiment, we utilize the multi-core processor to execute massive computations simultaneously by employing the multithreading technique, which significantly reduces the time complexity of similarity computation for the large text data.

### 3.3.5. Hadith Equivalence Class Generation and Results Analysis.

Hadith to Hadith similarity matrix contains the similarity values of a Hadith pair with every other Hadith from the Hadith representations. By applying different similarity thresholds, it generates different number of pairs. For the experiments, we respectively set a similarity

threshold value of 1, 0.97, 0.95 and 0.9 to obtain the number of pairs whose similarity value is higher than the given threshold. It is possible that one Hadith is lexically equivalent to multiple Hadiths in a representation of Hadith, so some Hadiths pairs with the given threshold could be formed into a bigger group that each entry is equivalent to each other. As a result, the stage 5 is to do this job by forming the generated pairs to the Hadith equivalence class with different similarity threshold. After formalization of the Hadith equivalence class, in last stage, result analysis, we engage experts to validate each equivalence class in order to verify whether all Hadiths in the class are semantically equivalent or not. If for an equivalence class it is verified that each Hadith is semantically equivalent to other class members, the Hadiths in this class are semantically interlinked in the given representation of the Hadith

## 3.4. Experimentation and Results

### 3.4.1. Experimental Results

For our experiments, we use two Hadith datasets: 1) Arabic plain text representation of the Hadith and 2) English text containing the translation of the Hadith. We implement and apply the three similarity measures across both these datasets. Additionally, the term weights were generated using the F and TFIDF term weighting schemes. Thus, the experimentation includes 2 * 3 * 2 = 12 different cases. The definition of each experiment scheme is defined as: dataset representation used – one of the term weighting schemes-similarity measure employed. Therefore, by this scheme English-Cosine-TFIDF indicates that the similarity measure Cosine has been applied using the TFIDF term weighting measure. As stated above, the result of each experiment is firstly to generate Hadith pairs with different similarity thresholds, and secondly form the pairs into a bigger group, which is the Hadith equivalence class. Since each equivalence class contains different number of Hadiths, so it is impossible to show all the corresponding

Table 10: The result of Arabic representation of the Hadith

| Threshold | Cosine | | | Jaccard | | | Hamming | | |
|---|---|---|---|---|---|---|---|---|---|
| | Pairs | Classes | Max Class Size | Pairs | Classes | Max Class Size | Pairs | Classes | Max Class Size |
| 1 | 19 | 19 | 2 | 16 | 16 | 2 | 16 | 16 | 2 |
| 0.97 | 148 | 136 | 3 | 35 | 35 | 2 | 16 | 16 | 2 |
| 0.95 | 324 | 258 | 5 | 55 | 53 | 3 | 16 | 16 | 2 |
| 0.9 | 976 | 631 | 6 | 185 | 168 | 4 | 40 | 38 | 3 |

Table 11: The result of English representation of the Hadith

| Threshold | Cosine | | | Jaccard | | | Hamming | | |
|---|---|---|---|---|---|---|---|---|---|
| | Pairs | Classes | Max Class Size | Pairs | Classes | Max Class Size | Pairs | Classes | Max Class Size |
| 1 | 484 | 11 | 2 | 484 | 82 | 11 | 484 | 80 | 11 |
| 0.97 | 673 | 127 | 13 | 490 | 86 | 11 | 484 | 82 | 11 |
| 0.95 | 849 | 151 | 19 | 516 | 99 | 11 | 484 | 82 | 11 |
| 0.9 | 1896 | 238 | 106 | 704 | 126 | 14 | 782 | 103 | 111 |

Hadith text in each group. Instead, Table 10 and Table 11 respectively show the number of Hadith pairs, the number of formative class, and the biggest class size for the cases with TFIDF term weighting scheme from the representation of English and Arabic Hadith. The reason that we discarded the cases with F term weighting scheme is that when we undertook investigation into the cases with F, we found out most of Hadith equivalence classes, whether in Arabic or

Table 12: The examples shows the Hadiths in one English equivalence class are not equivalent in Arabic

| One English Equivalence Class | Corresponding Hadiths in Arabic |
|---|---|
| This Hadith has been narrated on the authority of 'A'isha by another chain of transmitters. | وحدثنا ابن نمير، حدثنا أبي، حدثنا عبيد الله، حدثنا القاسم، عن عائشة، مثله . |
| This Hadith has been narrated on the authority of 'A'isha by another chain of transmitters. | حدثنا محمد بن المثنى، حدثنا أبو داود، حدثنا شعبة، أخبرني قتادة، قال سمعت مطرف بن عبد الله بن الشخير، قال أبو داود وحدثني هشام، عن قتادة، عن مطرف، عن عائشة، عن النبي صلى الله عليه وسلم بهذا الحديث . |

English, contain some mismatched Hadiths and many Hadiths which are commentaries. Because the purpose of this research is to interlink the actual narrations, not the commentaries, we focus on the cases with TFIDF term weighing scheme, which gives us more meaningful results.

### 3.4.2. Results Analysis

From the results presented in Table 10 and Table 11, it is clear that the H-English dataset produces far more pairs as compared to the H-Arabic dataset for the same similarity threshold.

From these results, one would expect the number of similar pairs from the H-English dataset to be significantly greater. We carried out expert driven validation to verify the correctness of the equivalence classes obtained. We performed this validation for the Arabic equivalence classes and English equivalence classes in H-Cosine-TFIDF experiments. This validation exercise revealed that the Hadiths in most English equivalence classes actually are not correspondingly equivalent in Arabic. One of the examples is shown in Table 12. The table shows that there are instances of Hadith in the dataset, which are commentaries referring to other Hadith, but do not contain any meaningful Hadith content. This indicates the limitation of the current dataset, that it does not distinguish or classify the Hadiths containing meaningful content from those which are only commentaries. The Hadiths in most English equivalence classes are found to be commentaries and their corresponding counterpart Hadith in Arabic may or may not be equivalent, the reason behind this is that English representation uses generic statements to describe the meaning of Hadith or point to some other Hadith, containing the actual text. The commentary does not include the Hadith text itself, it may include the statement about someone who said similar thing to the Hadith previously or in a different narration. However, in Arabic representation of the Hadith, the number of commentaries is much less than that in English representation, and most Hadiths contain meaningful content in them.

Table 13: 7 equivalence classes in Arabic which contain commentaries

| Equivalence Class(Arabic) | Relevant Commentary Hadith |
|---|---|
| Class 1 | وحدثنا أبو بكر بن أبي شيبة، حدثنا حفص بن غياث، عن الأعمش، بهذا الإسناد . |
| | حدثنا أبو بكر بن أبي شيبة، حدثنا حفص بن غياث، عن الأعمش، حدثنا أبو جهمة، بهذا الإسناد . |
| Class 2 | حدثنا محمد بن رافع، حدثنا عبد الرزاق، حدثنا معمر، عن همام بن منبه، عن أبي هريرة، عن النبي صلى الله عليه وسلم بمثله . |
| | حدثنا محمد بن رافع، حدثنا عبد الرزاق، حدثنا معمر، عن همام بن منبه، عن أبي هريرة، عن النبي صلى الله عليه وسلم بمثله . |
| | حدثنا محمد بن رافع، حدثنا عبد الرزاق، أخبرنا معمر، عن همام بن منبه، عن أبي هريرة، عن النبي صلى الله عليه وسلم . بمثله غير أنه قال ينبعث . |
| Class 3 | وحدثناه قتيبة بن سعيد، حدثنا عبد العزيز، - يعني الدراوردي - عن سهيل، بهذا الإسناد . |
| | حدثنا قتيبة بن سعيد، قال حدثنا عبد العزيز، - يعني الدراوردي - عن سهيل، بهذا الإسناد . |
| Class 4 | حدثنا عبد بن حميد، أخبرنا عبد الرزاق، أخبرنا معمر، عن الزهري، بهذا الإسناد . نحوه . |
| | وحدثنا عبد بن حميد، أخبرنا عبد الرزاق، أخبرنا معمر، عن الزهري، بهذا الإسناد نحوه . |
| Class 5 | وحدثناه أبو بكر بن أبي شيبة، وأبو كريب قالا حدثنا ابن نمير، حدثنا هشام، بهذا الإسناد . |
| | وحدثناه أبو بكر بن أبي شيبة، وأبو كريب قالا حدثنا ابن نمير، عن هام، بهذا الإسناد . |
| Class 6 | وحدثناه محمد بن المثنى، وابن، بشار قالا حدثنا محمد بن جعفر، حدثنا شعبة،بهذا الإسناد مثله . |
| | حدثنا محمد بن المثنى، وابن، بشار قالا حدثنا محمد بن جعفر، حدثنا شعبة، بهذا الإسناد مثله . |
| Class 7 | حدثنا يحيى بن يحيى، حدثنا عبد العزيز بن محمد المدني، ح وحدثنا محمد بن، رافع حدثنا ابن أبي فديك، عن هشام، - يعني ابن سعد - كلاهما عن زيد بن أسلم، بهذا الإسناد . |
| | وحدثناه يحيى بن يحيى، أخبرنا عبد العزيز بن محمد المدني، ح وحدثناه محمد، بن رافع حدثنا ابن أبي فديك، أخبرنا هشام، - يعني ابن سعد - كلاهما عن زيد بن، أسلم بهذا الإسناد مثله . |

Because the reasons stated above, we established that meaningful Hadith interlinkages may not be reliably derived from the results obtained from the English representation of Hadith. We therefore focused the analysis more on the Arabic representation. In the Arabic result table, it may be noticed that the number of pairs or equivalence classes with different thresholds obtained from the Cosine method is generally far greater than that from the Hamming and the Jaccard method. We therefore analyze the equivalence classes obtained from the Cosine method. We firstly investigate the equivalence classes with threshold 0.9, the reason that we choose to analyze from the results with 0.9 is that we believe if the bigger equivalence classes contain very few commentary classes, there is no change that small equivalence classes contain the commentary classes more than that. Finally, we conclude that the number of commentary class are not more than five from the results of Hamming and Jaccard method.

Table 14: 2 out of 624 equivalence classes in Arabic which were identified as narrations.

| Equivalence Class(Arabic) | Relevant Narration Hadith |
|---|---|
| Class 1 | حدثنا قتيبة بن سعيد، حدثنا ليث، ح وحدثنا محمد بن رمح، أخبرنا الليث، عن ابن شهاب، عن ابن المسيب، أنه سمع أبا هريرة، يقول قال رسول الله صلى الله عليه وسلم " والذي نفسي بيده ليوشكن أن ينزل فيكم ابن مريم صلى الله عليه وسلم حكما مقسطا فيكسر الصليب ويقتل الخنزير ويضع الجزية ويفيض المال حتى لا يقبله أحد " . |
| | حدثنا قتيبة بن سعيد، حدثنا الليث، عن ابن شهاب، عن ابن المسيب، أنه سمع أبا هريرة ـ رضى الله عنه ـ يقول قال رسول الله صلى الله عليه وسلم " والذي نفسي بيده ليوشكن أن ينزل فيكم ابن مريم حكما مقسطا فيكسر الصليب، ويقتل الخنزير، ويضع الجزية، ويفيض المال حتى لا يقبله أحد ". |
| Class 2 | حدثنا عبد الله بن يوسف، أخبرنا مالك، عن أبي الزناد، عن الأعرج، عن أبي هريرة ـ رضى الله عنه ـ أن رسول الله صلى الله عليه وسلم قال " رأس الكفر نحو المشرق، والفخر والخيلاء في أهل الخيل والإبل، والفدادين أهل الوبر، والسكينة في أهل الغنم ". |
| | حدثنا يحيى بن يحيى، قال قرأت على مالك عن أبي الزناد، عن الأعرج، عن أبي هريرة، أن رسول الله صلى الله عليه وسلم قال " رأس الكفر نحو المشرق والفخر والخيلاء في أهل الخيل والإبل الفدادين أهل الوبر والسكينة في أهل الغنم " . |
| | حدثني مالك، عن أبي الزناد، عن الأعرج، عن أبي هريرة، أن رسول الله صلى الله عليه وسلم قال " رأس الكفر نحو المشرق والفخر والخيلاء في أهل الخيل والإبل والفدادين أهل الوبر والسكينة في أهل الغنم " . |

The validation of equivalence classes was carried out by an expert to verify if all the members in the equivalence class were actually similar and there were no false positives. In addition, an important verification aspect was to identify if the equivalence class contained only the Hadith with some meaningful content. Any equivalence class that contained only the commentaries were classified differently. The results were promising in the range of similarity threshold 0.9 to 1. We found out that with threshold 1, there are only 2 equivalence classes containing the commentaries; with threshold 0.97, there are 3 equivalence classes contain the commentaries; with threshold 0.95, there are 4 equivalence classes that contain the commentaries and with threshold 0.9, there are 7. Table 13 demonstrates these 7 equivalence classes in Arabic representation of the Hadith (commentary). The remaining equivalence classes were all validated as having similar Hadith with valid and meaningful content. Table 14 shows two of the equivalence classes which were identified as narrations. For the similarity threshold 0.9, 624

equivalence classes were identified as narrations in Arabic, and the largest of this group contains six semantically equivalent Hadiths.

### 3.5 Conclusion and Future Work

In this chapter, we developed and improved the framework which was designed in our previous work in (Basharat, Yasdansepas, & Rasheed, 2015), with the aim to handle larger texts. We compared and contrasted the application of similarity measures across the English and Arabic representations of the Hadith. We investigated the result of Hadith text similarity analysis with different thresholds, and found out 624 classes of Hadith that provide a basis for meaningful interlinking in the Arabic representation of the Hadith. Also, we presented the reasons why the results based on the English Hadith and the results of cases with F term weighting scheme were not considered meaningful. More importantly, our results indicate that our proposed framework and techniques provide a promising basis for identifying relationships based on text similarity for the texts. In future, we aim to extend our analysis by considering alternative similarity thresholds to obtain more equivalence classes. Also, we can apply the same methodology across larger text data from related domains to find cross linkages.

### 3.6 References

A. Basharat., D. Yasdansepas., & K. Rasheed. (2015). Comparative study of verse vimilarity for multi-lingual representations of the qur'an. Proceedings on the International Conference on Artificial Intelligence (ICAI).

A. Bookstein., V. A. Kuliukin.,& T. Raita. (2012, Octorber). Generalized Hamming distance. Information Retrieval Journal. Vol 5, pp.353-375.

A. Huang.(2008). Similarity measures for text document clustering. in Pro- ceedings of the sixth new zealand computer science research student conference (NZCSRSC2008). pp.49-56. Christchurch, New Zealand.

A. Strehl., J. Ghosh., & R. Mooney. (2000). Impact of similarity measures on web-page clustering. in Workshop on Artificial Intelligence for Web Search (AAAI 2000). pp. 58–64.

A.Basharat., B.Abro., I.B.Arpinar., & K.Rasheed. (2016). Semantic Hadith: Leveraging Linked Data Opportunities for Islamic Knowledge, Linked Data on the Web.

A.K.Reinhart. (2010). Juynbolliana, Gradualism, the big bang, and Hadith study in the Twenty-First Century, Journal of the American Oriental Society, Vol. 130, No. 3, pp. 413-444.

F.Harrag., A.H.Cherif., & E.E.Qawasmeh. (2008). Information Retrieval Acrchitecture for 'Hadith' Text Mining, Journal of Dightal Informaiton Management, Vol. 6 Issue 6, p449.

K.A.Aldhaln., Akram.M.Zeki., & Ahmed.M.Zeki. (2012). Knowledge Extraction In Hadith Using Data Mining Technique, International Journal of Information Technology & Computer Science, pp. 13–21.

V. Topac. (2015). Efficient fuzzy search enabled hash map. Soft Computing Applications (SOFA), 2010 4th International Workshop, pp. 39-47.

# CHAPTER 4

## SUMMARY AND CONCLUSION

This thesis presents two studies on the two important Islamic knowledge sources: The Qur'an and the Hadith. In the similarity study on the Qur'an, we have investigated the effect of five similarity measures across four different representations of the Qur'an, in three different languages. We analyzed the results for identical and similar pairs of verses within these representations. We found that the English Qur'an: SAHIH representation demonstrates the most accurate result with highest F1 score among the different translations of the Qur'an. As for the performance of the five similarity measures used, we concluded that the Jaccard similarity measure proves to be effective for each of our tested verse pairs in the Qur'an. Furthermore, the Manhattan, Hamming and Cosine methods are applicable to some test cases, but the similarity results generated by the Pearson method cannot intuitively reflect the observed similarity. More importantly, we also found that the verse pairs that are similar in the original Arabic representation of the Qur'an are more or less lexically similar in the three translations even though the criteria of translation differ across different language.

In the similarity study on the Hadith, we first developed and improved the framework, making it able to handle bigger text datasets. We compared and contrasted the application of similarity measures across the English and Arabic representations of the Hadith. We investigated the results of Hadith text similarity analysis with different thresholds and found 624 classes of Hadith that provide a basis for meaningful interlinking in the Arabic representation of the Hadith.

We gave examples for the commentary Hadiths that are extracted from the Arabic representation of the Hadith and the true narrations contained in the 624 Arabic equivalence Hadiths.

Our results and the similarity analysis from both studies prove that our proposed framework and techniques provide a promising basis for identifying relationships based on text similarity for Islamic texts. While our approach is contextualized to the application of Islamic texts, the framework developed is generic enough to be applied across many other texts.

REFERENCES

A. Basharat., D. Yasdansepas., & K. Rasheed. (2015). Comparative study of verse vimilarity for multi-lingual representations of the qur'an. Proceedings on the International Conference on Artificial Intelligence (ICAI).

A. Bookstein., V. A. Kuliukin.,& T. Raita. (2012, Octorber). Generalized Hamming distance. Information Retrieval Journal. Vol 5, pp.353-375.

A. Hasnat., S. Halder., A. Hoque., D. Bhattacharjee.,& M. Nasipuri. (2013, May -June). A fast fpga based architecture for measuring the distance between tow color images using manhattan. International Journal of Electronics and Communication Engineering & Technology(IJECET). vol. 4, pp. 01–10.

A. Huang.(2008). Similarity measures for text document clustering. in Pro- ceedings of the sixth new zealand computer science research student conference (NZCSRSC2008). pp.49-56. Christchurch, New Zealand.

A. Strehl., J. Ghosh., & R. Mooney. (2000). Impact of similarity measures on web-page clustering. in Workshop on Artificial Intelligence for Web Search (AAAI 2000). pp. 58–64.

A.-B. M. Sharaf., & E. Atwell. (2012). Qursim: A corpus for evaluation of relatedness in short texts. in LREC, pp. 2295–2302.

A.Basharat., B.Abro., I.B.Arpinar., & K.Rasheed. (2016). Semantic Hadith: Leveraging Linked Data Opportunities for Islamic Knowledge, Linked Data on the Web.

A.K.Reinhart. (2010). Juynbolliana, Gradualism, the big bang, and Hadith study in the Twenty-First Century, Journal of the American Oriental Society, Vol. 130, No. 3, pp. 413-444.

E. Atwell., N. Habash., B. Louw,. B. Abu Shawar., T. McEnery., W. Zaghouani., & M. El-Haj. (2010). Understanding the quran: A new grand challenge for computer science and artificial intelligence. ACM-BCS Visions of Computer Science 2010.

E. Atwell., C. Brierley., K. Dukes., M. Sawalha., & A.-B. Sharaf. (2011). An artificial intelligence approach to arabic and islamic content on the internet. Proceedings of NITS 3rd National Information Technology Symposium.

F.Harrag., A.H.Cherif., & E.E.Qawasmeh. (2008). Information Retrieval Acrchitecture for 'Hadith' Text Mining, Journal of Dightal Informaiton Management, Vol. 6 Issue 6, p449.

K.A.Aldhaln., Akram.M.Zeki., & Ahmed.M.Zeki. (2012). Knowledge Extraction In Hadith Using Data Mining Technique, International Journal of Information Technology & Computer Science, pp. 13–21.

M. A. Sherif., & A.-C. Ngonga Ngomo.(2009). Semantic quran - a multilingual resource for natural-language processing. Semantic Web.

M. S. Khorsheed., & A. O. Al-Thubaity. (2013). Comparative evaluation of text classification techniques using a large diverse arabic dataset. Language resources and evaluation. vol. 47, no. 2, pp. 513-538.

R. S. Forsyth., & S. Sharoff. (2014). Document dissimilarity within and across languages: A benchmarking study. Literary and Linguistic Computing. vol. 29, no. 1, pp. 6–22.

V. Topac. (2015). Efficient fuzzy search enabled hash map. Soft Computing Applications (SOFA), 2010 4th International Workshop, pp. 39-47.

W. H. Gomaa, & A. A. Fahmy. (2013, April). A survey of text similarity approaches. International Journal of Computer Applications(0975-8887) vol. 68 -No.13 pp. 01-06.