Batna 2 University

Text Mining and Analytics for Extracting/ Discovering Knowledge from the Holy Quran

by
Rahima BENTRCIA

A dissertation submitted for the degree of
Doctorat 3rd cycle LMD

Department of Computer Science, College of Math and Computer science, Batna 2 University

October 2017

République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieure et de la Recherche Scientifique
Université de Batna 2

Thèse

Présentée au Département d'informatique
Faculté des Mathématiques et Informatique

Pour l'obtention du diplôme de

Doctorat 3ème cycle LMD
Spécialité: Informatique

par
Rahima BENTRCIA

Thème

Fouille de Texte et Analyse pour Extraction/Découverte de la Connaissance du Saint Coran

Thèse dirigée par

Dr. Samir Zidat	Rapporteur
Pr. Farhi Marir	Co-rapporteur

Soutenue le : 02/10/2017
Devant le jury composé de :

Pr. Bilami Azeddine	Professeur	université Batna 2
Dr. Guezouli Larbi	MCA	université Batna 2
Dr. Behloul Ali	MCA	université Batna 2
Pr. Benmohamed Mohamed	Professeur	université Constantine 2

Dedication

To my family ☺

Acknowledgement

First and foremost, my unconditional praise and thankfulness goes to Allah, the Most Compassionate, and the Most Merciful for his countless bounties and blessings.

I would like to present my deep thanks for my supervisors Dr. Samir Zidat and Pr. Farhi Marir for their encouragement and continuous support throughout different phases of this dissertation.

Most importantly, I express my deep gratitude to my family Mom and Dad, brothers and sisters, for their boundless support, priceless sacrifice, and continuous prayers...

My father Pr. Mohammad Bentrcia, my brothers Dr. Abdelouahab and Dr. Toufik... thank you very much for your ideas and suggestions.

I cannot forget my lovely little family, my husband Seifeddine, my daughter Tesnim Nour Essojoud, and my twins Sadjed Lirrahmane Elbaraa and Aabed Errahmane Ouis...thank you my stars for enlightening my life by your love, support, and patience.

Abstract

There is an immense need to information systems that rely on Arabic Quranic text to present a precise and comprehensive knowledge about Quran to the world. This motivates us to conduct our research work which uses Quran as a corpus and exploits text mining techniques to perform three different tasks: extracting semantic relations that exist between words linked by AND conjunction, analyzing the order of the words of that conjunctive phrase, and finally measuring the similarity between Quran chapters based on lexical and statistical measures.

Since semantic relations are a vital component in any ontology and many applications in Natural Language Processing strongly depend on them, this motivates the development of the first part in our thesis to extract semantic relations from the holy Quran, written in Arabic script, and enrich the automatic construction of Quran ontology. We focus on semantic relations resulting from proposed conjunctive patterns which include two words linked by the conjunctive AND. These words can be nouns, proper nouns, or adjectives. The strength of each relation is measured based on the correlation coefficient value between the two linked words. Finally, we measure the significance of this method through hypothesis testing and Student t-test.

Moreover, some aspects of semantic relations that may exist between words are inspired from patterns of word co-occurrences. Hence, statistics performed on these patterns are very useful to provide further information about such relations. This fact induces conducting the second part in our research, which is an analytical study, on one type of these patterns called the AND conjunctive phrases, that exist in the holy Quran. First, we propose a set of AND conjunctive patterns in order to extract the conjunctive phrases from the Quranic Arabic Corpus which we convert to Arabic script. Then, we analyze the order of the two words that form the conjunctive phrase. We report three different cases: words that have occurred in a specific order in the conjunctive phrase and repeated only once in the Quran, words that have occurred in a specific order in the conjunctive phrase and repeated many times in the Quran, and words that have occurred in different positions in the conjunctive phrase and repeated one or many time(s) in the holy Quran. Finally, we show that different word orders in the conjunctive phrase yield different contextual meanings as well as different values of association relationship between the linked words.

Similarity Measure between documents is a very important task in information retrieval. However, a crucial issue is the selection of an efficient similarity measure which improves time and performance of such systems.. In the last part of our thesis, we present a lexical approach to extracting similar words and phrases from Arabic texts, represented by Quran chapters (Surah). Furthermore, we measure the similarity value between these chapters using three different statistical metrics: cosine, Jaccard, and correlation distances.

الملخص

هناك حاجة ماسة إلى أنظمة المعلومات التي تعتمد على النص العربي للقرآن الكريم لتقديم معرفة دقيقة وشاملة حول القرآن للعالم. هذا الهدف دفعنا للقيام بهذا البحث الذي يستخدم القرآن الكريم كذخيرة لغوية و يستغل تقنيات تعدين النصوص للقيام بثلاثة أجزاء من البحث. الجزء الأول و هو استخراج العلاقات الدلالية الموجودة في القرآن الكريم بين أي كلمتين مربوطتين بأداة العطف (الواو) اعتمادا على قواعد اللغة العربية و على علم الاحصاء و مبدأ برهنة الفرضيات. الجزء الثاني يقدم دراسة تحليلية لمبدأ التقديم و التأخير في المعطوفات الموجودة في القرآن الكريم و قياس كمية المعلومات التي يقدمها كل لفظين معطوفين وردا في القرآن الكريم بترتيبين مختلفين. الجزء الثالث و الأخير يعتمد على استخراج الألفاظ و الآيات المتشابهات من السور القرآنية و قياس نسبة التشابه مستخدمين عدة مقاييس رياضية.

تعد العلاقات الدلالية عنصرا رئيسيا في الانتولوجيا التي تستثمر في مختلف التطبيقات و لذلك قمنا باستخراج هذا العنصرمن النص العربي للقرآن الكريم لتعزيز البناء التلقائي لانتولوجيا القرآن. ركزنا في الجزء الأول من البحث على استخراج العلاقات الدلالية الموجودة بين الألفاظ القرآنية المعطوفة بالواو و التي يجب أن تشترك في المعنى كما هو منصوص في قواعد النحو العربي و ذلك باقتراح عدة أنماط لألفاظ معطوفة بالواو. هذه الألفاظ قد تكون أسماء أو أسماء أعلام أوصفات. بعد ذلك قمنا بقياس قوة العلاقات الدلالية المستخرجة باستخدام معامل الارتباط كما قمنا بالتأكد من فعالية هذه الطريقة المقترحة بتطبيق مبادئ في علم الإحصاء مثل برهنة الفرضيات.

من جهة أخرى، بعض خصائص العلاقات الدلالية بين الألفاظ مستوحاة من أنماط التشارك بين هذه الألفاظ و كيفية ظهورها في الذخيرة اللغوية و هذا ما حفزنا للقيام بالجزء الثاني من البحث الذي يستخرج جميع المعطوفات بالواو من القرآن الكريم و يقدم دراسة تحليلية شاملة حول ظهور هذه المعطوفات بأنماط معينة في القرآن الكريم من خلال تغير الترتيب بين اللفظين المعطوفين حيث يتقدم أحدهما على الآخر أو يتأخر و توصلنا الى ثلاثة نتائج: هناك قسم من المعطوفات ذكرت في القرآن الكريم مرة واحدة فقط و بطريقة واحدة و معينة من حيث الترتيب و التأخير. قسم آخر من المعطوفات ورد في القرآن الكريم عدة مرات و لكن بطريقة واحدة معينة. القسم الأخير من المعطوفات و هي التي وردت في القرآن الكريم مرة أو عدة مرات و لكن بطريقتين مختلفتين حيث يتقدم أحد اللفظين على الآخر في موضع و يتأخر عنه في موضع آخر. و أخيرا بينا أن التقديم و التأخير في معطوفات القرآن الكريم يؤدي الى اختلاف في المعنى السياقي للنص و قوة علاقة التشارك بين اللفظين المعطوفين.

في الجزء الثالث من بحثنا هذا، تطرقنا الى مبدأ استخراج النصوص المتشابهة لما لها من أهمية في نظم استخراج المعلومات. ورغم ذلك، يعد اختيار مقياس دقيق للتشابه عملا حاسما لتحكمه في فعالية و سرعة هذه الأنظمة. بالنسبة للنصوص العربية فإنها تفتقر لهذه الأبحاث لعدم توفر الموارد اللغوية الخاصة باللغة العربية و لطبيعة الخط العربي. ركزنا في هذا الجزء على استخراج الألفاظ و الجمل المتشابهة بين أي سورتين قرآنيتين كما قمنا بقياس مقدار التشابه مستعملين ثلاثة معايير في علم الإحصاء و هي جيب التمام و جاكارد و معامل الارتباط.

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Problem Statement

Arabic language is widely used in more than 25 countries as a second language or as a mother tongue. For Muslims, it has a very special position because it is the language of the holy Quran. For this reason, many studies were performed to help Arabic and non-Arabic Muslims for better understanding the Quran. However, there is a lack in the developed approaches that deal with Arabic script due, for example, to the nature of Arabic writing, the semantic ambiguity of words, and the shortage in resources and tools that support Arabic such as corpora, lexicons, and machine-readable dictionaries, which are essential to advance research in different areas.

The main theme of this work is to extract knowledge from the holy Quran using Arabic traditional grammar (قواعد النحو العربي) and text mining techniques. This is motivated by the immense need to provide a precise and a comprehensive study about Arabic Quran to the world. This could be exploited very efficiently in many linguistic, scientific, and religious researches such as building ontology and developing a search engine and a question answering system for Quran. Furthermore, it consolidates many theoretical rules relying on scientific evidences such as those related to medical science and discovering origin of diseases, diet, environment, etc.

In spite of the extensive work done on text mining for Latin and Asian documents, a small number of research papers and reports are published on extracting knowledge from Arabic texts. This is certainly due to difficulties associated with Arabic language, which are enlightened in the next sections.

1.1.1 Arabic Variations

Arabic language can be divided into two main forms based on significant differences in vocabulary, where each form has its own register of lexicon. Classical Arabic (CA) is the ancient form of Arabic that appeared from Umayyad and Abbasid times (7th to 9th centuries) and includes the language of the Quran and early Islamic literature which is the main reason why the language has preserved its purity throughout the centuries. Modern Standard Arabic (MSA), as its name indicates, is the modern version of Classical Arabic that is used in all life styles including writing and formal speech. Although (MSA) is growing in size over time and new technical terms are being introduced, but the richness of Arabic vocabulary is superior in the Classical Arabic register. However, recent computational and linguistic studies are focusing on (MSA) because of its easy vocabulary and increasing usage in different areas while ignoring (CA) for its hard lexicon and style which strongly abide by Arabic grammar. This is considered as one of the main challenging issues which motivate conducting our work on Arabic Quran.

1.1.2 Nature of Arabic Writing

The Arabic alphabet consists of twenty eight letters, twenty five of which are consonants and the remaining three letters are long vowels. Each letter has between two to four different shapes of writing based on its location in the word; at the beginning, in the middle, at the end, or isolated. Some letters have only two forms: the isolated form and the final form, as shown in Table 1.1.

All Arabic letters can be connected at least from one side. This cursive nature of Arabic writing is a challenging problem in developing systems that need to find the boundaries of each letter. Moreover, Arabic is the richest natural language in the world in terms of morphological inflection and derivation, where most words are built up from roots with adding segments of letters called prefixes and suffixes. Determining these segments correctly

is a critical issue in any Part-of-Speech (POS) tagging system. One example is shown in Table 1.2.

Table 1.1 Sample of Arabic letters forms

Name	End	Middle	Start	isolated
ألف	ـا	ـا	ا	ا
باء	ـب	ـبـ	بـ	ب
تاء	ـت	ـتـ	تـ	ت
ثاء	ـث	ـثـ	ثـ	ث
جيم	ـج	ـجـ	جـ	ج
حاء	ـح	ـحـ	حـ	ح
خاء	ـخ	ـخـ	خـ	خ
دال	ـد	ـد	د	د

Table 1.2 The tokenization of the Arabic word (فستذكرون)

Word	**Suffix**	**Root**	**Prefix**
فستذكرون	ون	ذكر	فست

1.1.3 Semantic Ambiguity

Ambiguity can be defined as the property of having two or more distinct meanings or interpretations of the same word or the same sentence. A word or sentence is ambiguous if it can be interpreted in more than one way. Arabic uses short vowels, or diacritics to disambiguate words. There are four diacritics in Arabic: the fatHa (ً) is a character put on the top of a letter to give the "a" sound, the Kasra (ٍ) is put under a letter to give the "i" sound, the Dhamma (ٌ) is a character put on the top of a letter to give the "u" sound, and finally the Sukoun (ْ) which is a character put on the top of a letter to indicate the absence of any of the first three sounds. There exists a set of Arabic words that have more than one meaning based on the context they appear, creating a one-to-many ambiguity. Such examples from the Quran are depicted in Table 1.3.

Table 1.3 Example of one word with one-to-many ambiguity

The word	The different meanings	The context in the Quran
ٱلْإِثْم	**الشِّرْك** Polytheism	لَوْلاَ يَنْهَاهُمُ الرَّبَّانِيُّونَ وَالأَحْبَارُ عَن قَوْلِهِمُ الإِثْمَ
	المَعْصِيَة Sin	فَمَنِ اضْطُرَّ فِي مَخْمَصَةٍ غَيْرَ مُتَجَانِفٍ لِإِثْمٍ
	الذَّنْب Guilt	فَمَنْ تَعَجَّلَ فِي يَوْمَيْنِ فَلا إِثْمَ عَلَيْهِ
	الْخَطَأ The error	فَمَنْ خَافَ مِن مُّوصٍ جَنَفًا أَوْ إِثْمًا

Arabic uses diacritics to disambiguate words since many words in Arabic can have the same body (letters) but different diacritics. However, most modern Arabic texts do not include diacritics, and this, yields another case of ambiguity which depends on the context to find out the suitable meaning. Table 1.4 shows examples from the Quran.

Table 1.4 Example of one word with different meanings based on diacritics

The word	The different meanings		The context in the Quran
سنة	**سِنَة**	نعاس drowsiness	لاَ تَأْخُذُهُ سِنَةٌ وَلاَ نَوْمٌ
	سَنَة	عام Year	يَوَدُّ أَحَدُهُمْ لَوْ يُعَمَّرُ أَلْفَ سَنَةٍ
	سَنَة	جدب و قحط Waterless and drought	وَلَقَدْ أَخَذْنَا ءَالَ فِرْعَوْن بِالسِّنِينَ وَنَقْصٍ مِنَ الثَّمَرَاتِ
	سُنَّة	طريقة و نهج Method and approach	إلاّ أنْ تَأتِيَهُمْ سُنَّةُ الأَوَّلِينَ

1.1.4 POS Tagging

Part-of-Speech tagging (POS) is the process of assigning grammatical part-of-speech tags to words based on their context. For Arabic, basic tag sets include verbs, particles, and nouns, which can be subcategorized into adjectives, proper nouns, and pronouns (Khoja, 2001; Kanaan et al., 2003).

Assigning the correct tag is another major challenging issue in Arabic resulting from the absence of diacritics. This leads to different possible POS tags and hence ambiguity, which can only be solved using contextual information. The following statement suffers from the ambiguity, as described in Table 1.5.

Table 1.5 Part-of-Speech tagging for the Arabic sentence (حام الفراش حول الورد)

Word	**Transliteration: POS Tag**	**Meaning**
حام	Haam : Proper Noun Haama : **Verb** Haamin : Adjective	Haam Hover Hot
الفراش	Al-Farash: Noun Al-Firash: Noun	Butterflies Mattress
حول	Hawla: **Preposition** Hawl:Noun Hawal:Noun Hawwala:Verb Huwwila:Verb	Around Year Squint Diverted Was diverted
الورد	Al-Ward:Noun Al-Wird: Noun	Roses Part

1.2 Novel Contributions of this Thesis

So many studies are conducted by Muslim scholars to investigate the holy Quran like Quran exegesis books. Since Quran was revealed in Arabic, most of these approaches are dealing with the linguistic and religious aspects of Quran such as (Adhima, 1972; Al- Soyouti, 1973; Al-Samiraii, 2006). Our proposed work is totally different from all the reported research in that it combines text mining techniques, statistics, and Arabic grammars in order to understand and extract knowledge from the Quran as follows:

1. Exploiting the Arabic grammatical rule which states that any two words linked by AND conjunction must share a relationship of some sort. Hence, we extract AND conjunctive phrases from the whole Quran using a set of predefined patterns and then we reveal the semantic relations that exist between any two words in those phrases. This contribution aims to improve the construction of Quran ontologies by providing important types of semantic relations.
2. Exploiting statistical analyses to measure the strength of the extracted semantic relations. This contribution would save time and efforts of specific domain experts and

validate their decisions very efficiently compared to traditional procedure which depends only on Muslim scholars books.

3. Performing an analytical study about AND conjunction in Quran and revealing the secrets behind different orders of words that linked by AND conjunction. We utilize statistical approaches to prove that different word orders in the conjunctive phrase yield different values of the association relationship between the combined words. This contribution supports the religious studies which state that different word orders in the conjunctive phrase yield different contextual meanings of Quranic verses.
4. Measuring similarity between Arabic Quran chapters using lexical and statistical approaches. This contribution is very important in extracting similar words/verses and estimating the semantic similarity between the chapters.
5. All the above contributions can also be applied to any other Arabic corpus rather than the holy Quran because all of them follow the same Arabic grammars.

All these contributions are carried out either to support earlier reported research, or are established from scratch in order to serve the holy Quran as follows:

1. Extracting knowledge from the holy Quran and presenting to the researchers in Islamic studies to expand and deepen their understanding of Islam.
2. Extracting knowledge from the holy Quran based on scientific procedures and motiving Muslims and non-Muslims to further investigate and understand the Quran and the values it brings to different domains of human life.
3. Serving the Arabic language by employing its grammatical rules and linguistic features to analyze a miraculous text such as Quran and other holy books.
4. Exploiting this research to develop information retrieval systems for Quran and other text mining applications such as semantic similarity and relatedness between Quran verses.

1.3 Thesis Organization

The proposed work discusses many different topics; each topic is explained in details in a separated chapter as follows:

An overview about the main theme of the proposed work is presented in Chapter 1. We discussed the major issues that make developing information systems for Arabic Quran very challenging. The main contribution of this thesis has been the development of three novel approaches that seek mining Quran to produce a comprehensive and precise knowledge about this sacred text to the whole world.

Chapter 2 introduces several topics related to the holy Quran such as its revelation, the classification of its chapters based on revelation location, and the different themes that Quran talks about. Also, a detailed description of the Quranic Arabic Corpus used in this study was provided.

Chapter 3 reports detailed literature review of previous approaches and techniques developed in the field of Arabic text mining in general and the Quranic text in particular.

Chapter 4 presents theoretical topics related to ontology learning from a text. A description of the main components that form an ontology of any domain was provided besides the common approaches used to develop them. In addition, efficient ontology tools and languages were briefly demonstrated.

A novel approach that aims at enriching the construction of Quran ontology is detailed in Chapter 5. We exploited Arabic grammar to capture all semantic relations that exist in AND

conjunctive phrases. Furthermore, we utilized statistical techniques namely correlation, Student t-test, and testing hypothesis to validate and measure the strength of the extracted relations. This aids domain experts to estimate and validate their final decisions very efficiently. Finally, we categorized manually those relations into Antonymy, Gender, and Class.

Chapter 6 introduces the main Quran mining approaches that exist recently. Then, it discusses our proposed work which relies on the concept of order between words that are linked by AND conjunction. In particular, we conducted an analytical study about words that take different positions/orders in the conjunctive phrase. We demonstrated that different positions of one word in a phrase yield different meanings and values of association relationship between the two words. Finally, we measure these values using Pointwise Mutual Information method (PMI) (Bouma, 2009) and the Sketch Engine tool function (Word Sketch Difference) (the Word Sketch Difference help).

One text mining application is described in Chapter 7, where we tried to find similarities between any two chapters of Arabic Quran. We extracted similar words, phrases, verses, and their frequencies from the two chapters looking for estimating their semantic similarity. Moreover, we explored three different similarity metrics, which are cosine, Jaccard, and correlation distances, to find out the degree of similarity between any two Quranic chapters.

Finally, in Chapter 8, we concluded the discussed approaches in this work and we provided our future directions and perspectives toward discovering other issues related to Quranic knowledge and achieving better performance.

2 The Holy Quran

2.1 Overview

Quran is the holy book of more than 1.6 billion Muslims all around the world. They believe that Quran is the word of God, revealed in Arabic through the angel Gabriel to prophet Muhammad (PBUH) over 23 years beginning in 609 CE when Muhammad was 40, and concluding in 632, the year of his death.

The text of Quran is organized in 114 chapters (Sura) of different length where each chapter consists of few or many verses (Aya). The longest chapter is called Al-Bakara and includes 286 verses whereas Al-Kawthar is the shortest chapter and consists of three verses.

الم ﴿1﴾ ذَلِكَ الْكِتَابُ لاَ رَيْبَ فِيهِ هُدًى لِّلْمُتَّقِينَ ﴿2﴾ الَّذِينَ يُؤْمِنُونَ بِالْغَيْبِ وَيُقِيمُونَ الصَّلاةَ وَمِمَّا رَزَقْنَاهُمْ يُنفِقُونَ ﴿3﴾ والَّذِينَ يُؤْمِنُونَ بِمَا أُنزِلَ إِلَيْكَ وَمَا أُنزِلَ مِن قَبْلِكَ وَبِالآخِرَةِ هُمْ يُوقِنُونَ ﴿4﴾

"AlifLaamMeem. That is the (Holy) Book, where there is no doubt. It is a guidance for the cautious (of evil and Hell). Who believe in the unseen and establish the (daily) prayer; who spend out of what We have provided them. Who believe in that which has been sent down to you (Prophet Muhammad) and what has been sent down before you (to Prophets Jesus and Moses) and firmly believe in the Everlasting Life". ***(Al-Bakara, 1:4)***

إِنَّا أَعْطَيْنَاكَ الْكَوْثَرَ ﴿1﴾ فَصَلِّ لِرَبِّكَ وَانْحَرْ ﴿2﴾ إِنَّ شَانِئَكَ هُوَ الْأَبْتَرُ ﴿3﴾

"Indeed, We have given you (Prophet Muhammad) the abundance (Al Kawthar: river, its pool and springs). So pray to your Lord and sacrifice. Surely, he who hates you, he is the most severed". ***(Al-Kawthar, 1:3)***

Each chapter is classified as Meccan or Medinan based on whether the verses were revealed before or after the migration of Muhammad to the city of Medina. The title (name) of each chapter was inspired from the major topic that the chapter discussed or from its first letters or words. The chapters' titles, their length, and their classification are clarified in Table 2.1.

Table 2.1 Chapters of the holy Quran

Quran Chapters							
Number	Title	Length (no. of verses)	Meccan vs. Medinan	Number	Title	Length (no. of verses)	Meccan vs Medinan
1	Al-Fatiha	7	Meccan	58	Al-Mujadilah	22	Medinan
2	Al-Baqarah	286	Medinan	59	Al-Hashr	24	Medinan
3	Al Imran	200	Medinan	60	Al-Mumtahanah	13	Medinan
4	An-Nisa'	176	Medinan	61	As-Saff	14	Medinan
5	Al-Ma'idah	120	Medinan	62	Al-Jumu'ah	11	Medinan
6	Al-An'am	165	Meccan	63	Al-Munafiqun	11	Medinan
7	Al-A'raf	206	Meccan	64	At-Taghabun	18	Medinan
8	Al-Anfal	75	Medinan	65	At-Talaq	12	Medinan
9	At-Tawbah	129	Medinan	66	At-Tahreem	12	Medinan
10	Yunus	109	Meccan	67	Al-Mulk	30	Meccan
11	Hud	123	Meccan	68	Al-Qalam	52	Meccan
12	Yusuf	111	Meccan	69	Al-Haqqah	52	Meccan
13	Ar-Ra'd	43	Medinan	70	Al-Ma'aarij	44	Meccan
14	Ibraheem	52	Meccan	71	Nuh	28	Meccan
15	Al-Hijr	99	Meccan	72	Al-Jinn	28	Meccan
16	An-Nahl	128	Meccan	73	Al-Muzzammil	20	Meccan
17	Al-Isra	111	Meccan	74	Al-Muddathir	56	Meccan
18	Al-Kahf	110	Meccan	75	Al-Qiyamah	40	Meccan
19	Maryam	98	Meccan	76	Al-Insan	31	Medinan
20	Ta-Ha	135	Meccan	77	Al-Mursalat	50	Meccan
21	Al-Anbiya'	112	Meccan	78	An-Naba'	40	Meccan
22	Al-Hajj	78	Medinan	79	An-Nazi'at	46	Meccan
23	Al-Mu'minoon	118	Meccan	80	`Abasa	42	Meccan
24	An-Nur	64	Medinan	81	At-Takweer	29	Meccan
25	Al-Furqan	77	Meccan	82	Al-Infitar	19	Meccan
26	ash-Shu`ara'	227	Meccan	83	Al-Mutaffifeen	36	Meccan

27	An-Naml	93	Meccan	84	Al-Inshiqaq	25	Meccan
28	Al-Qasas	88	Meccan	85	Al-Burooj	22	Meccan
29	Al-`Ankabut	69	Meccan	86	At-Tariq	17	Meccan
30	Ar-Rum	60	Meccan	87	Al-A'la	19	Meccan
31	Luqman	34	Meccan	88	Al-Ghashiyah	26	Meccan
32	As-Sajdah	30	Meccan	89	Al-Fajr	30	Meccan
33	Al-Ahzab	73	Medinan	90	Al-Balad	20	Meccan
34	Saba'	54	Meccan	91	Ash-Shams	15	Meccan
35	Fatir	45	Meccan	92	Al-Lail	21	Meccan
36	Ya seen	83	Meccan	93	Ad-Dhuha	11	Meccan
37	As-Saffat	182	Meccan	94	Al-Inshirah	8	Meccan
38	Sad	88	Meccan	95	Al-Teen	8	Meccan
39	Az-Zumar	75	Meccan	96	al-`Alaq	19	Meccan
40	Ghafir	85	Meccan	97	Al-Qadr	5	Meccan
41	Fussilat	54	Meccan	98	Al-Bayyinah	8	Medinan
42	Ash-Shura	53	Meccan	99	Az-Zalzala	8	Medinan
43	Az-Zukhruf	89	Meccan	100	Al-Adiyat	11	Meccan
44	Ad-Dukhan	59	Meccan	101	al-Qari`ah	11	Meccan
45	Al-Jathiyah	37	Meccan	102	At-Takathur	8	Meccan
46	Al-Ahqaf	35	Meccan	103	Al-Asr	3	Meccan
47	Muhammad	38	Medinan	104	Al-Humazah	9	Meccan
48	Al-Fath	29	Medinan	105	Al-Feel	5	Meccan
49	Al-Hujurat	18	Medinan	106	Al-Quraish	4	Meccan
50	Qaf	45	Meccan	107	Al-Maa'oun	7	Meccan
51	Ad-Dhariyat	60	Meccan	108	Al-Kawthar	3	Meccan
52	At-Tur	49	Meccan	109	Al-Kafiroun	6	Meccan
53	An-Najm	62	Meccan	110	An-Nasr	3	Medinan
54	Al-Qamar	55	Meccan	111	Al-Masad	5	Meccan
55	Ar-Rahman	78	Medinan	112	Al-Ikhlas	4	Meccan
56	Al-Waqi'ah	96	Meccan	113	Al-Falaq	5	Meccan
57	Al-Hadeed	29	Medinan	114	Al-Nas	6	Meccan

Quran also has other different names, mentioned in the following verses:

- The Book (الكِتَاب)

  (تِلْكَ آيَاتُ **الْكِتَابِ** الْمُبِين)

(These are the verses of the clear **Book**) [26:2]

- The Revelation (التَّنْزِيل)

  (**تَنْزِيلٌ** مِنْ رَبِّ الْعَالَمِين)

  ([It is] a **revelation** from the Lord of the worlds) [56:80]

- The Criterion (الفُرْقَان)

(تَبَارَكَ الَّذِي نَزَّلَ **الْفُرْقَانَ** عَلَىٰ عَبْدِهِ لِيَكُونَ لِلْعَالَمِينَ نَذِيرًا)

(Blessed is He who sent down the **Criterion** upon His Servant that he may be to the worlds a warner) [25:1]

- The Message (الذِّكْر)

(ذَٰلِكَ نَتْلُوهُ عَلَيْكَ مِنَ الْآيَاتِ وَ**الذِّكْرِ** الْحَكِيمِ)

(This is what We recite to you, [O Muhammad], of [Our] verses and the precise [and wise] **message**) [3:58]

- The Eloquence (البَيَان)

(عَلَّمَهُ **الْبَيَانَ**)

([And] taught him **eloquence**) [55:4]

- The Truth (الحَقّ)

(فَلَمَّا جَاءَهُم بِ**الْحَقِّ** مِنْ عِندِنَا قَالُوا اقْتُلُوا أَبْنَاءَ الَّذِينَ آمَنُوا مَعَهُ وَاسْتَحْيُوا نِسَاءَهُمْ ۚ وَمَا كَيْدُ الْكَافِرِينَ إِلَّا فِي ضَلَالٍ)

(And when he brought them the **truth** from Us, they said, "Kill the sons of those who have believed with him and keep their women alive." But the plan of the disbelievers is not except in error) [40:25]

- The Wisdom (الحِكْمَة)

(ذَٰلِكَ مِمَّا أَوْحَىٰ إِلَيْكَ رَبُّكَ مِنَ **الْحِكْمَةِ** ۗ وَلَا تَجْعَلْ مَعَ اللَّهِ إِلَٰهًا آخَرَ فَتُلْقَىٰ فِي جَهَنَّمَ مَلُومًا مَّدْحُورًا)

(That is from what your Lord has revealed to you, [O Muhammad], of **wisdom**. And, [O mankind], do not make [as equal] with Allah another deity, lest you be thrown into Hell, blamed and banished) [17:39]

- The Guide (الهُدَى)

شَهْرُ رَمَضَانَ الَّذِي أُنزِلَ فِيهِ الْقُرْآنُ **هُدًى** لِّلنَّاسِ وَبَيِّنَاتٍ مِّنَ الْهُدَىٰ وَالْفُرْقَانِ ۚ فَمَن شَهِدَ مِنكُمُ الشَّهْرَ فَلْيَصُمْهُ ۖ وَمَن كَانَ مَرِيضًا أَوْ عَلَىٰ سَفَرٍ فَعِدَّةٌ مِّنْ أَيَّامٍ أُخَرَ ۗ يُرِيدُ اللَّهُ بِكُمُ الْيُسْرَ وَلَا يُرِيدُ بِكُمُ الْعُسْرَ وَلِتُكْمِلُوا الْعِدَّةَ وَلِتُكَبِّرُوا اللَّهَ عَلَىٰ مَا هَدَاكُمْ وَلَعَلَّكُمْ تَشْكُرُونَ)

(The month of Ramadhan [is that] in which was revealed the Qur'an, a **guidance** for the people and clear proofs of guidance and criterion. So whoever sights [the new moon of] the month, let him fast it; and whoever is ill or on a journey - then an equal number of other days. Allah intends for you ease and does not intend for you hardship and [wants] for you to complete the period and to glorify Allah for that [to] which He has guided you; and perhaps you will be grateful) [2:185]

- The Light (النُّور)

  (يَا أَيُّهَا النَّاسُ قَدْ جَاءَكُم بُرْهَانٌ مِّن رَّبِّكُمْ وَأَنزَلْنَا إِلَيْكُمْ **نُورًا** مُّبِينًا)

(O mankind, there has come to you a conclusive proof from your Lord, and We have sent down to you a clear **light**) [4:174]

- The Inspiration (الرُّوح)

  وَكَذَٰلِكَ أَوْحَيْنَا إِلَيْكَ **رُوحًا** مِّنْ أَمْرِنَا ۚ مَا كُنتَ تَدْرِي مَا الْكِتَابُ وَلَا الْإِيمَانُ وَلَٰكِن جَعَلْنَاهُ نُورًا نَّهْدِي بِهِ مَن نَّشَاءُ مِنْ عِبَادِنَا ۚ وَإِنَّكَ لَتَهْدِي إِلَىٰ صِرَاطٍ مُّسْتَقِيمٍ)

(And thus We have revealed to you an **inspiration** of Our command. You did not know what is the Book or [what is] faith, but We have made it a light by which We guide whom We will of Our servants. And indeed, [O Muhammad], you guide to a straight path) [42:52]

This holy book is considered as a legislation source which organizes Muslims lives since it contains commandments and laws related to different subjects, divided by scholars into 15, which are: Pillars of Islam, Faith, The call for Allah, The holy Quran, Jihad, Action (Work), Man and the moral relations, Man and the social relations, Organizing financial relationships,

(Trade, Agriculture, Industry and Hunting), Judicial relationships, General and political relationships, Science and art, Religions, and finally, the stories and the history.

One of the most valuable topics that Islam discussed in the holy Quran is sciences and arts, where God asked people to learn and conduct research in various fields. This is proved decisively when we find that the first chapter (Al-'Alaq) revealed to the prophet Muhammad began by the word read إِقْرَأْ:

اقْرَأْ بِاسْمِ رَبِّكَ الَّذِي خَلَقَ ﴿1﴾ خَلَقَ الْإِنسَانَ مِنْ عَلَقٍ ﴿2﴾ اقْرَأْ وَرَبُّكَ الْأَكْرَمُ ﴿3﴾ الَّذِي عَلَّمَ بِالْقَلَمِ ﴿4﴾ عَلَّمَ الْإِنسَانَ مَا لَمْ يَعْلَمْ ﴿5﴾

"Read (Prophet Muhammad) in the Name of your Lord who created, created the human from a (blood) clot. Read! Your Lord is the Most Generous, who taught by the pen, taught the human what he did not know." ***(Al-'Alaq, 1:5)***

Moreover, Quran discussed many scientific facts and natural phenomena that have been corroborated recently by modern researchers but were a kind of miracles in the Qur'an at that time.

2.2 Quranic Arabic Corpus (QAC)

Quranic Arabic Corpus is an integrated and reliable linguistic resource developed by Kais Dukes in Leeds University (Quranic Arabic Corpus). The corpus provides three levels of analysis: morphological annotation, a syntactic treebank, and a semantic ontology (Dukes et al., 2013). It consists of 77,430 words of Quranic Arabic, divided into 114 documents. Each word is tagged with its part-of-speech as well as multiple morphological features that are based on the traditional Arabic grammar. Also, it is stored as a text file and is available for free.

The data in the corpus is written in Buckwalter Arabic transliteration scheme (Buckwalter scheme) and organized into four columns as follows:

1. LOCATION: consists of four parts: (chapter no: verse no: word no: part no).
2. FORM: consists of the main parts of the word.
3. TAG: includes the part-of-speech tag for each part of the word such as Noun, Verb, and Adjective, etc.
4. FEATURES: describes morphological features of the word such as Root, Stem, and Gender, etc.

Figure 2.1 shows verses of Al-Fatiha chapter, in Buckwalter transliteration scheme.

LOCATION	FORM	TAG	FEATURES
(1:1:1:1)	bi	P	PREFIX|bi+
(1:1:1:2)	somi	N	STEM|POS:N|LEM:{som|ROOT:smw|M|GEN
(1:1:2:1)	{ll~ahi	PN	STEM|POS:PN|LEM:{ll~ah|ROOT:Alh|GEN
(1:1:3:1)	{l	DET	PREFIX|Al+
(1:1:3:2)	r~aHoma`ni	ADJ	STEM|POS:ADJ|LEM:r~aHoma`n|ROOT:rHm|MS|GEN
(1:1:4:1)	{l	DET	PREFIX|Al+
(1:1:4:2)	r~aHiymi	ADJ	STEM|POS:ADJ|LEM:r~aHiym|ROOT:rHm|MS|GEN
(1:2:1:1)	{lo	DET	PREFIX|Al+
(1:2:1:2)	Hamodu	N	STEM|POS:N|LEM:Hamod|ROOT:Hmd|M|NOM
(1:2:2:1)	li	P	PREFIX|l:P+
(1:2:2:2)	l~ahi	PN	STEM|POS:PN|LEM:{ll~ah|ROOT:Alh|GEN
(1:2:3:1)	rab~i	N	STEM|POS:N|LEM:rab~|ROOT:rbb|M|GEN
(1:2:4:1)	{lo	DET	PREFIX|Al+
(1:2:4:2)	Ea`lamiyna	N	STEM|POS:N|LEM:Ea`lamiyn|ROOT:Elm|MP|GEN
(1:3:1:1)	{l	DET	PREFIX|Al+
(1:3:1:2)	r~aHoma`ni	ADJ	STEM|POS:ADJ|LEM:r~aHoma`n|ROOT:rHm|MS|GEN
(1:3:2:1)	{l	DET	PREFIX|Al+
(1:3:2:2)	r~aHiymi	ADJ	STEM|POS:ADJ|LEM:r~aHiym|ROOT:rHm|MS|GEN

Figure 2.1 Verses of Al-Fatiha chapter in Buckwalter transliteration scheme

Because Buckwalter transliteration scheme is not an easy way for users to read and understand the corpus, we need a pre-processing step to represent it in a clearer and more readable format such as the Arabic script. Therefore, in our work, we have developed a conversion method to transfer back each character from Buckwalter scheme to its equivalent Arabic character. These include Arabic alphabet and diacritics. Figure 2.2 demonstrates a sample of the Quranic Arabic corpus converted to Arabic script.

LOCATION	FORM	TAG	FEATURES
(1:1:1:1)	بِ	P	PREFIX|bi+
(1:1:1:2)	سْمِ	N	STEM|POS:N|LEM:اِسْم|ROOT:smw|M|GEN
(1:1:2:1)	اللَّهِ	PN	STEM|POS:PN|LEM:اللَّه|ROOT:Alh|GEN
(1:1:3:1)	ا	DET	PREFIX|Al+
(1:1:3:2)	رَّحْمَٰ	ADJ	STEM|POS:ADJ|LEM:رَحْمَٰن|ROOT:rHm|MS|GEN
(1:1:4:1)	ا	DET	PREFIX|Al+
(1:1:4:2)	رَّحِيمِ	ADJ	STEM|POS:ADJ|LEM:رَحِيم|ROOT:rHm|MS|GEN
(1:2:1:1)	الْ	DET	PREFIX|Al+
(1:2:1:2)	حَمْدُ	N	STEM|POS:N|LEM:حَمْد|ROOT:Hmd|M|NOM
(1:2:2:1)	لِ	P	PREFIX|l:P+
(1:2:2:2)	لَّهِ	PN	STEM|POS:PN|LEM:اللَّه|ROOT:Alh|GEN
(1:2:3:1)	رَبِّ	N	STEM|POS:N|LEM:رَبّ|ROOT:rbb|M|GEN
(1:2:4:1)	الْ	DET	PREFIX|Al+
(1:2:4:2)	عَٰلَمِينَ	N	STEM|POS:N|LEM:عَٰلَمِين|ROOT:Elm|MP|GEN
(1:3:1:1)	ا	DET	PREFIX|Al+
(1:3:1:2)	رَّحْمَٰ	ADJ	STEM|POS:ADJ|LEM:رَحْمَٰن|ROOT:rHm|MS|GEN
(1:3:2:1)	ا	DET	PREFIX|Al+
(1:3:2:2)	رَّحِيمِ	ADJ	STEM|POS:ADJ|LEM:رَحِيم|ROOT:rHm|MS|GEN

Figure 2.2 Verses of Al-Fatiha chapter converted to Arabic script

3 Literature Review

3.1 Quran Ontologies

Ontology learning from text in general occupies a large area in computer science domain whereas ontology learning from the holy Quran suffers from lack due to the nature of Arabic script and the depth of knowledge needed in this field (Habash, 2010). However, few recent Quranic studies were interested in developing approaches which accomplish ontology learning tasks and represent the Quranic knowledge in a semantic way as sets of concepts and relations. Dukes initiated the Quranic Arabic Corpus (QAC) which is the first online collaboratively constructed linguistic resource with multiple layers of annotation including part-of-speech tagging, morphological segmentation, and syntactic analysis using dependency grammar (Dukes and Habash, 2010; Dukes and Atwell, 2012). Also, the author built ontology from (QAC) which finds relations between proper nouns or any nouns if they represent well-defined concepts such as the names of animals, locations, and religious entities. The ontology was validated based on scholarly sources, namely Tafsir of Ibn Kathir.

A large corpus named QurAna (Sharaf and Atwell, 2012) was created from the original Quranic text, where specific types of words are considered as ontological concepts. Personal pronouns are extracted and tagged with their antecedence. These antecedents are maintained as an ontological list of concepts which improves information systems performance. Another study exploited an existing index of Quranic topics from a scholarly source: Tafsir of Ibn Kathir to develop Qurani (Abbas, 2009), which is a tool that looks for concepts in the holy Quran and provides English translations for the verses containing these concepts. A system was proposed in (Yauri et al., 2012) reused Leeds ontology (Quranic Arabic Corpus ontology) (http://corpus.quran.com/ontology.jsp) to model Quran domain knowledge, using Web

Ontology Language OWL. However, the system added the acts concepts related to specific topics in Quran such as praying, Zakat, sin, and rewards, and showed the relations between them using Description Logic concepts. The user of this model can semantically retrieve important concepts from the holy Quran. Verses referring to particular concepts could also be retrieved.

DataQuest (Ul Ain and Basharat, 2011) is an efficient framework for modelling and retrieving knowledge from distributed knowledge sources primarily related to the holy Quran, with the use of semantic web, information extraction, and natural language processing techniques. The documents are annotated using the domain ontology. Thus, users can query that filtered and concise knowledge using a semantic based intelligent search engine. Another work (Al-Yahya et al., 2010) covered a specific topic in the holy Quran and built a computational model for representing Arabic lexicons using ontologies. The model has been implemented on the Arabic language vocabulary associated with "Time nouns" vocabulary in the holy Quran. The ontology consists of 59 words; only 28 of them are used as a basis for the model design and organized semantically into a hierarchical classification with general concepts at the top, and specific at the bottom.

In addition, (Baqai et al., 2009) developed knowledge based platforms that used semantic web technologies to model, store, publish, reason, and retrieve knowledge from distributed sources related to the holy Quran and associated scholarly texts.

A recent work is conducted by (Aman et al., 2017) to review the ontology development approaches for Islamic knowledge domain and identify the main problems and shortcomings of the existing approaches. By a comprehensive literature review, authors found that extraction of concepts and their relationships is the main challenging task in Domain Ontology Learning. Also, (Tashtoush et al., 2017) proposed a new ontological modeling that models the human social relations in the noble Quran by employing Web Ontology Language

(OWL) as well as Resource Description Framework (RDF). The work involves a descriptive identification of the human relations related concepts that are described in the Noble Quran with identifying the relations among them. As a result, SPARQL queries and DL queries are used in the ontology model to retrieve Quran domains, concepts and Verses in Arabic language.

Despite all these studies aim to extract knowledge from Quran, making a serious comparison between them is unfair because there are no complete Quran ontologies. We find many of them have covered specific topics in Quran or special types of words rather than the whole Quran words (Saad et al., 2010). Also, many researchers have built ontologies for parts of Quran and very few have used the entire Quran. Moreover, each ontology has focused only on one or two types of relations between terms such as synonymy and Part-Of (Shoaib et al., 2009). In the validation process, all the reported approaches (Alrehaili and Atwell, 2014) depend on either domain experts or exegesis books such as Tafsir of Ibn Kathir.

In chapter 5, we introduce a novel approach to extract semantic relations from the whole Quran, written in Arabic. There are no similar approaches to our study since we rely on Arabic grammatical tool (AND conjunction) and statistical techniques to efficiently extract and validate the obtained results rather than using the methods mentioned earlier.

3.2 Quran Mining

The concept of text mining is becoming increasingly popular. Therefore, many studies are carried out to show the different methods used to analyze and extract knowledge from textual data. A study (Momtazi et al., 2010) proposed a term clustering algorithm to retrieve sentences from a corpus. This algorithm is based on assigning similar terms to the same clusters based on their tendency to co-occur in similar contexts. Also, they compared four

different methods for estimating word co-occurrence frequencies from two different corpora and discussed their effects on the system performance. In addition, (Islam and Inkpen, 2006) introduced a corpus-based method for calculating the semantic similarity of pair of words. They used Point-wise Mutual Information (PMI) to measure the common words in the context of the two target words and exploit these PMI values to calculate the relative semantic similarity. The results were evaluated using four different corpora. Furthermore, (Gomaa and Fahmy, 2013) discussed in a survey three different methods of text similarity: String-based, Corpus-based, and Knowledge-based similarities. A hybrid of these approaches was presented and useful similarity packages were mentioned.

For Arabic text, little has been written about text mining due to the nature of Arabic script (Habash, 2010). In (Alrabiah et al., 2014), two empirical studies were performed and applied a number of probabilistic distributional semantic models to automatically identify lexical collocations. They tested the performance of eight different association measures on the holy Quran in the first study, and they constructed a Classical Arabic corpus to be used in the second study. Experiments showed that MI.log_freq association measure achieved the best results in extracting the collocations whereas mutual information association measure achieved the worst results. Another approach (Attia et al., 2008) was presented to design and implement an Arabic lexical semantics Language Resource (LR) that enables the retrieval of the possible senses of any given Arabic word at a high coverage. Instead of tying full Arabic words to their possible senses, they related morphologically and POS-tags constrained Arabic lexical compounds to a predefined limited set of semantic fields across which the standard semantic relations are defined and hence the possible senses of the desired Arabic word are retrieved.

A different method (Thabtah et al., 2012) was introduced to classify Arabic documents, specifically the published Corpus of Contemporary Arabic (CCA), using four classification learning algorithms: Decision trees (C4.5), Hybrid (PART), Rule Induction (RIPPER), and Simple Rule (OneRule). They used WEKA (https://sourceforge.net/projects/weka/), the open source Machine learning tool, to evaluate the performance of these algorithms and they found that C4.5 is the most appropriate algorithm to Arabic text classification in terms of error-rate, precision, and recall. Moreover, Badea system (Al-Yahya et al., 2016) was developed in order to enrich the ontological lexicon of Arabic language. Badea was built semi-automatically to extract lexical relations specifically antonyms using a pattern-based approach. The method used ontology of "seed" pairs of antonyms to facilitate the extraction of lexico-syntactic patterns in which the pairs occur. These patterns are then used to find new antonym pairs in a set of Arabic language corpora. The results showed important findings on the reliability of patterns in extracting antonyms for Arabic.

On the other side, Quran mining occupies a large area in text mining although very few approaches have been developed for Quranic Arabic due to the depth of knowledge needed in this field and the challenges related to Arabic script. A research study (Safeena and Kammani, 2013) reviewed Qur'anic computation methods in term of research and application. The work surveyed the development of Quranic computation using a literature review and classification of journal articles, conference proceedings, and dissertations from 1997 to 2011. This study also covered general Arabic besides Quranic Arabic and helped to facilitate the understanding of Quranic text. Another text mining study (Alhawarat et al., 2015) analyzed the Arabic text of the holy Quran and provided statistical information based on term frequencies that are calculated using both Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF) methods. After the preprocessing step, authors performed a set of

experiments, particularly the most important words in Quran, its word cloud, and chapters with high term frequencies. Different partitions of Quran were tested such as using the whole chapters of the holy Quran, using parts of the holy Quran, and using Document-Term Matrix representation.

An illustrative graphic-based tool (Hamam et al., 2015) was created to help Quran experts to easily mine Quran. This platform not only links one chapter to another chapter, or one verse to another verse through words, but also connects chapters and verses together through concepts and dependencies. Also, it provides expert users the ability to add new aspects and their dependencies to a shared database.

A different approach (Al-Kabi et al., 2013) classified the verses of Al-Fatiha and Yaseen chapters automatically. The classifier normalizes the verses in the first step then applies the score function to categorize each verse to the class for which it has the highest score value. The accuracy rate reached 91% although it can be improved using a full corpus of the holy Quran and a better stemmer. Furthermore, (Siddiqui et al., 2013) proposed a Probabilistic Topic Model method to discover the thematic structure of the holy Quran. First, they applied a number of preprocessing steps to the Arabic Quranic chapters (Surahs) in order to obtain the final set of features from the raw text in those documents. Then, they used the Latent Dirichlet Allocation (LDA) algorithm (Blei et al., 2003) which was run with different values of the input parameters to identify topics at different levels of granularity. Finally, the topics contained in each surah along with the most important terms that defined those topics were extracted.

A recent work is proposed by (Zakariah et al., 2017) to cover the current state of the art in the majority of areas related to digital Quran applications, their trends and challenges. A comprehensive and detailed survey was provided that encompasses most of the previous work

and emerging issues related to Digital Quran Computing including Quran authentication, e-Learning, mobile and game techniques, memorization techniques, natural language processing (NLP), standardization, and voice recognition. The findings of this work calls on the research community to provide technical solutions to protect the originality of the Quran and monitor the authenticity of online Quran publications.

In chapter 6, we present an analytical study that reveals the rationale behind the order of words in AND conjunctive phrases in the Quranic Arabic. Comparing the previous methods of Quran mining and our approach, there is no research study that analyzed this sacred text the way it is done here. This study presents a totally novel approach since none of the existing methods illustrated the order concept of co-occurred words or even provided statistics about the different positions/ orders that co-occurred words had taken in Quran. The association relationship between the co-occurred words is also measured using different statistical techniques.

3.3 Quran similarity and relatedness

Measuring document similarity is one of the most significant problems of text mining and information retrieval. Quran documents have also received a special attention due to the religious and linguistic value. We find a study (Panju, 2014) conducted to extract and visualize topics in the Quran corpus based on statistical approaches. First, matrix factorization was used to successfully extract meaningful topics from Quran text, written in English. Then, data visualization through t-SNE dimensionality reduction (Maaten and Hinton, 2008) was used to cluster the verses based on their themes and word usage. Another work (Al-Dargazelli, 2004) was presented to identify numerical patterns of number of verses in Quran chapters using five statistical methods: mean mode, median, range, standard deviation (SD), and relative standard deviation (RSD). Also, Quran chapters were classified into Maccan and Madinite based on the number of verses in each chapter.

A hybrid statistical classifier (Nassourou, 2011) consisting of stemming and clustering algorithms was developed for comparing lexical frequency profiles of Quran chapters and deriving dates of revelation for each chapter. The classifier is trained using some chapters with known dates of revelation. Then it classifies chapters with uncertain dates of revelation by computing their proximity to the training ones. (Akour et al., 2014) proposed another approach to retrieve the most similar verses in Quran compared to a user input verse as a query. Moreover, they employed N-gram and LibSVM classifier (Hall et al., 2009) to classify Quran chapters to Makki and Madani chapters. (Dost and Ahmad, 2008) performed a probabilistic study of Makki, Madani, and Mixed chapters of the holy Quran. They relied on word size and length to classify the chapters to Makki and Madani. (Sharaf and Atwell, 2012) presented QurSim which is a large corpus created from the original Quranic text, where semantically similar or related verses are linked together. The authors created online query system where the user inputs a verse number and is returned with both directly and indirectly related verses, in Arabic and English, with the degree of relatedness. The obtained data set includes more than 7600 pairs of related verses collected from scholarly sources with several levels of relatedness degree.

In chapter 7, we address the similarity concept between any two chapters in the Quran. We use a simple matching algorithm to extract similar phrases from the two chapters where the phrase could be a single word, part of the verse, or a complete verse. Further, we apply several statistical methods to measure the similarity degree between them.

4 Ontologies

4.1 Introduction

An early definition of the term ontology appeared in 1993 (Gruber, 1993), where it is defined as a specification of a conceptualization. Maedche and Staab stated the original description of ontology learning from a text as the acquisition of a domain model from data (Maedche and Staab, 2001), where the extracted knowledge from the text is represented by concepts and relationships. Hence, semantic relations are an important element in the construction of ontologies (Alvarez et al., 2007). Besides they hold together the concepts that represent the domain, they solve the ontology structuring problems. Furthermore, providing richer semantics to these relations facilitates selecting the operations that can be performed on them and the task that the ontology can tackle. However, semantic relations have not been given the attention they deserve because of the difficulty to capture the whole information related to the problem domain as well as the various and imprecise interpretations provided for a relation representation.

The process of ontology learning passes through several tasks organized in a layer cake. Each layer is explained deeply in (Cimiano, 2006; Liu et al., 2011). Traditionally, ontology construction depends on domain experts, but it is lengthy, costly and controversial (Navigli et al., 2003). Therefore, automatic ontology construction approach was suggested but it is also still a difficult task due to the lack of a structured knowledge base or domain thesaurus (Lee et al., 2007).

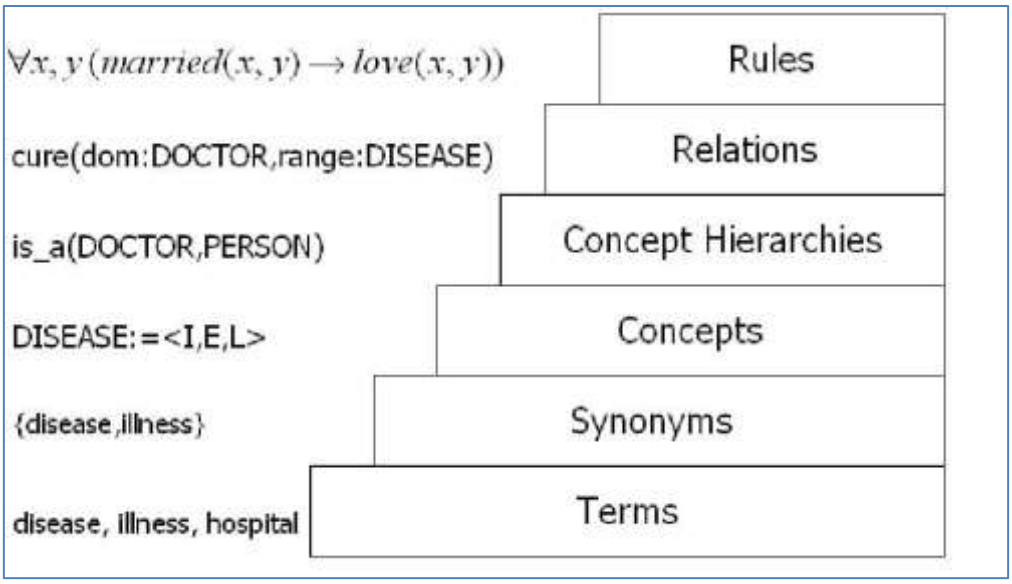


Figure 4.1 Ontology Learning Layer Cake (Cimiano, 2006)

Despite all these challenges, ontologies can provide potential benefits for a lot of applications such as text classification and clustering (Bloehdorn and Hotho, 2004), where additional conceptual features extracted from ontologies are used to enhance the bag-of-words model. In information retrieval and extraction, ontologies can solve the problem of vocabulary mismatch between documents and user queries, and many other problems (Guarino et al., 1999; Elabd et al., 2015). Also, ontologies provide the necessary vocabulary for Natural Language Processing systems and state which semantic relations potentially hold between different concepts (Nirenburg and Raskin, 2004).

In the next sections, we will describe the layers in Figure 4.1 which are basic elements in the ontology learning from text process along the common languages and tools used to accomplish this task.

4.2 Ontology learning from text

There are three different methods used to construct ontologies: Linguistic, Statistical, and Hybrid. In linguistic approaches, linguistic information and natural language processing tools

are used to extract information from text such as part-of-speech tagging, sentence parsing, and syntactic structure analysis (Sabou et al., 2005). Besides that, lexico-syntactic patterns are well known in extracting many types of relations that exist between concepts like hyponym, hypernym, and taxonomic relationships (Brewster et al., 2009). On the other hand, statistical approaches rely on machine learning, information retrieval, and data mining methods. Clustering algorithms (Fortuna et al., 2007) are widely used to extract synonyms of words based on similarity measures or to group words that share similar meaning under one cluster, where the cluster name is a concept, and the words inside this cluster are its instances. Word co-occurrence statistics are also applied to extract terms that tend to occur together and may have probable relationships between them (Punuru and Chen, 2011). Finally, the hybrid approaches exploit combinations of linguistic and statistical methods to construct ontologies (Cimiano and Volker, 2005).

(Al-Arfaj and Al-Salman, 2015) introduced a categorization approach to build ontologies based on five criteria: knowledge sources, level of automation, learning targets, purpose, and learning techniques, as shown in Figure 4. 2.

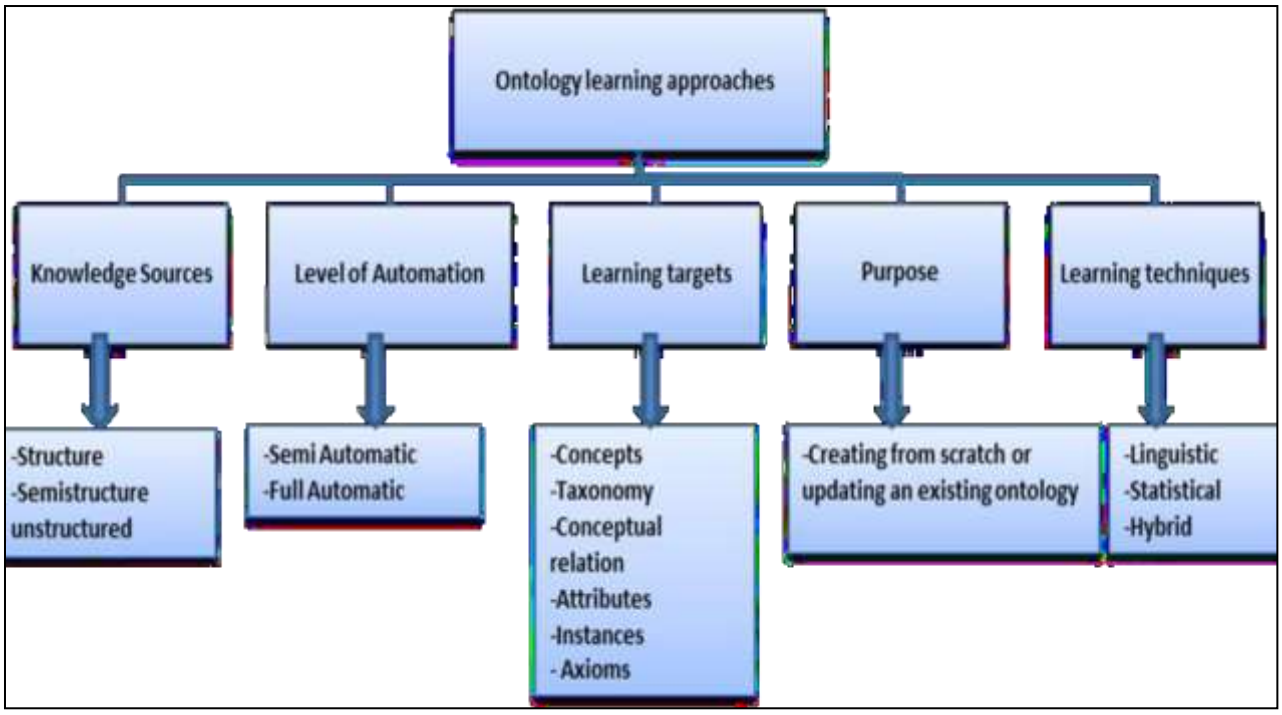


Figure 4.2 Classifications of Ontology Learning Approaches (Al-Arfaj and Al-Salman, 2015)

Although ontologies can belong to different domains, they share the same basic elements which are explained bellow:

4.2.1 Terms

Terms, also known as instances, individuals, and objects, are the smallest component in any ontology and represent the concepts of a specific domain. Ontologies are built based on a lexicon of relevant terms that could be either single words or multi-word compounds extracted from a given corpus. Some methods used linguistic approaches to extract terms such as phrase analysis, dependency structure analysis, and ad-hoc patterns (Frantzi and Ananiadou, 1999). Statistical methods are then applied to select the relevant terms only according to their frequency in the input corpora (Salton et al., 1975).

4.2.2 Synonyms

Synonyms can be defined as terms that have similar meaning and hence represent the same concept/relation in a specific domain. Several methods are developed to identify synonyms among ontological terms in order to eliminate redundant concepts/relations and reduce the

cost and effort of ontology learning. WordNet (Miller, 1990) is a source knowledge that used to find the exact sense of each term to extract synonyms. Clustering techniques also are exploited based on Harris's distributional hypothesis where similar terms in meaning tend to share syntactic contexts (Lin and Pantel, 2001). In addition, Latent Semantic Indexing algorithms are based on dimension reduction approaches to extract inherent relations between terms (Schütze, 1993) and thus group similar ones in one cluster. Recently, statistical techniques over the Web are widely employed for this purpose (Baroni and Bisi, 2004).

4.2.3 Concepts

The concept, also named class, type, and set, is a general abstract of a group of terms that have the same characteristics in a specific domain. According to the ontological view, concept formation should provide (i) an intensional definition of concept, (ii) a set of concept instances, i.e. its extension (iii) the lexical realizations which are used to refer to the concept (Buitelaar et al., 2005). Three main approaches are presented by researchers to extract concepts; clustering techniques are exploited to capture related terms and group them in one cluster (Reinberger and Spyns, 2005) which is the concept in this case. The second approach discovers concepts from an extensional point of view like the systems (Evans, 2003) whereas the third approach treats the intensional definitions of the concept which could be formal or informal definitions. An informal definition might be a textual description, i.e. a gloss of the concept. A formal definition includes the extraction of concept properties such as relations between a particular concept and other concepts (Velardi et al., 2005).

4.2.4 Concept Hierarchies

One type of relations that may exist between concepts is the taxonomic relations, or the hierarchy of concepts, where concepts are organized into sub-super-concept tree structures. One popular approach for taxonomy discovery in textual domains is the hierarchical clustering algorithms (Zavitsanos et al., 2006) which exploited in the concepts discovery task

as well as the process of ordering them hierarchically. Lexico-syntactic patterns such as Hearst patterns (Hearst, 1992) and their variations are also used to extract taxonomic relations like `is a kind of' and `is a part of' relationships. A different approach is called Probabilistic Topic Models that produce a hierarchical modelling of a particular collection (Blei et al., 2004). Each textual document is modelled as a set of concepts across a specific path of the learned hierarchy from the root to a leaf. The identification of concepts in the ontology and their taxonomic arrangement is performed simultaneously.

4.2.5 Relations

Relations or relationships in the ontology describe the way in which concepts are related to each other or specify a concept's properties. Few studies have been conducted to extract non-taxonomic relations (all relations that are not used in the formation of the concept hierarchy) since we found two main approaches. The first one is based on lexico-syntactic patterns that aim at extracting verbs from textual data and the surrounding concepts (Buitelaar et al., 2004). These verbs are supposed to be candidate relations and need further validation. The other approach used association rules and their variant algorithms such as sentence-based term co-occurrence (Maedche and Staab, 2000) to extract anonymous associations, which are named appropriately in a second step.

4.2.6 Rules

One main objective of ontology learning is the ability to derive facts that are presented by the knowledge in the ontology. This is done through a set of statements called rules which describe the logical inferences that can be drawn from an assertion in a particular form. Almost no work has been done to acquire ontological rules except few methods that based on unsupervised learning for discovering inference rules from text (Lin and Pantel, 2001) or analyzing the syntactic structure of a natural language definition and the application of

transformation rules on the resulting dependency tree to produce a list of axioms that can be combined with concepts definitions (Völker et al., 2007).

4.3 Ontology languages

Ontology languages are formal languages used to construct ontologies. They allow the encoding of knowledge about specific domains and often include reasoning rules that support the processing of that knowledge. Researchers addressed the topic of ontology languages and classified the developed languages into three categories explained in details in (Corcho and Gomez-Perez, 2000), as depicted in Figure 4.3.

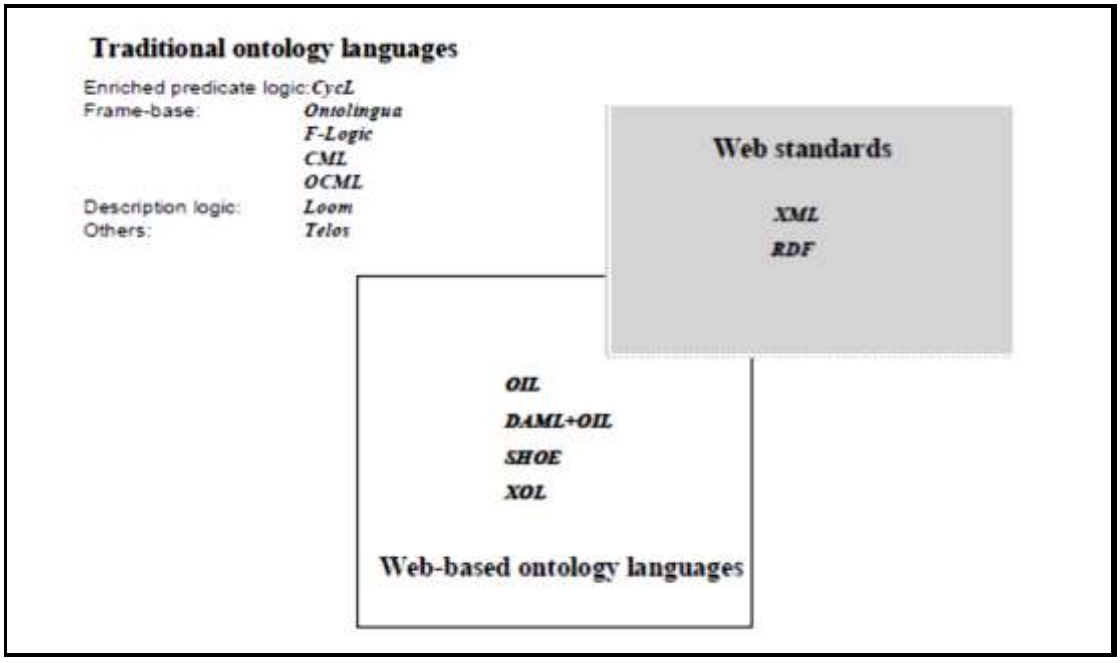


Figure 4.3 The three groups of ontology languages (Su and Ilebrekke, 2002)

In this section, we analyze ontology languages which are widely used by the ontology community.

4.3.1 Knowledge Interchange Format (KIF)

KIF is a computer oriented language based on first order logic created in 1992 to exchange knowledge between distinct computer systems, developed by different programmers at

different times, in different languages (Genesereth and Fikes, 1992). Thus, KIF can also be used as a language for expressing and exchanging ontologies. The language has declarative semantics and is logically comprehensive. It allows the user to make knowledge representation decisions explicit and to introduce new knowledge representation constructs without changing the language. Also, it provides the definitions for the objects, functions, and relations.

4.3.2 Ontolingua

In 1992, the Knowledge Systems Lab (KSL) at Stanford University developed Ontolingua language to support the design and specification of ontologies with a clear logical semantics based on (KIF). Ontolingua (Sireteanu, 2013) extends (KIF) using additional syntax to include the intuitive collection of axioms into definitional forms with ontological significance and a Frame Ontology to define object-oriented and frame-language terms. As a result, Ontolingua ontology is made up of definitions of classes, relations, functions, objects, and axioms that describe these terms.

4.3.3 Resource Description Framework (RDF)

(RDF) is a semantic-network based language developed by the World Wide Web Consortium (W3C) to describe Web resources (Lassila and Swick, 1999). The main purpose of (RDF) is to find a mechanism for describing resources that make no assumptions about a particular application domain nor the structure of a document containing information. Thus, (RDF) provides a model to represent metadata in XML. This data model consists of three object types (a triple): resources (subjects) defined by entities that can be referred to by an address at the WWW, properties which are predicates describing the resources, and statements (objects) that assign a value for a property in a resource.

4.3.4 Web Ontology Language (OWL)

OWL is the most powerful ontology languages that currently exist for the Semantic Web. It is also developed by the World Wide Web Consortium (W3C). OWL is a collection of RDF triples that hold a formally defined meaning to express the semantics of information contents on the web (Maniraj and Sivakumar, 2010). OWL ontology consists of a set of axioms which place constraints on sets of individuals (classes) and the types of relationships permitted between them. These axioms provide semantics by allowing systems to infer additional information based on the data explicitly provided.

4.4 Ontology tools

To develop ontologies in various domains, there is a demand to software tools called ontology editors which allow users to visually manipulate, inspect, browse, and code ontologies, and can be applied to several stages of the ontology life cycle such as creation, population, validation, deployment, maintenance, and evolution. In this section, we provide briefly a description of the most common ontology editing tools:

4.4.1 Protégé-2000

Protégé is an ontology and knowledge base editor produced by Stanford University (Noy et al., 2000). It is an open source, standalone application that used for knowledge acquisition, merging and alignment of existing ontologies, ontology language importation/exportation, and plug-in new functional modules to augment its usability. It allows the definition of classes, class hierarchies, variables, variable-value restrictions, and the relationships between classes and the properties of these relationships by generating graph representations of the editing ontology. A snapshot of this tool is shown in Figure 4.4.

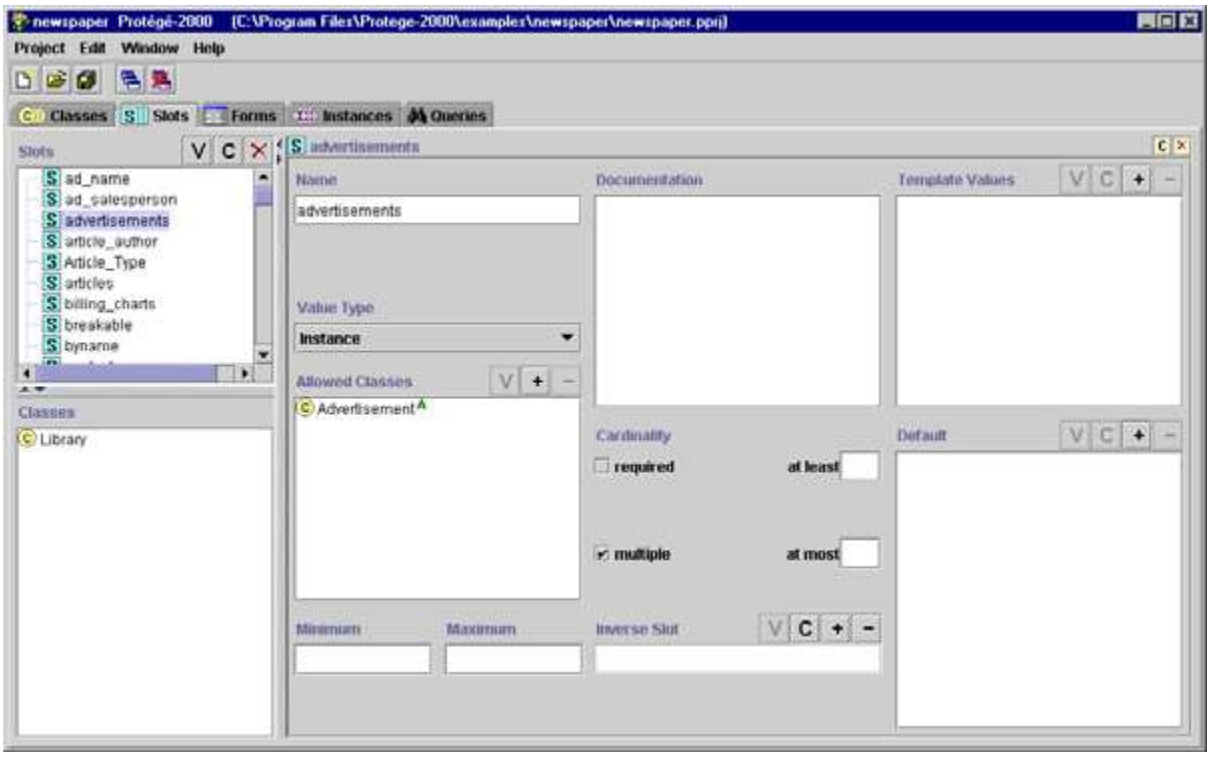


Figure 4.4 Snapshot of Protégé 2000 (Youn and McLeod, 2006)

4.4.2 OntoEdit

OntoEdit is an ontology engineering environment created by the Knowledge Management Group of the University of Karlsruhe to support the development and maintenance of an ontology (Sure et al., 2003). OntoEdit is based on an open plug-in structure where every plug-in provides other features to deal with the domain requirements like OntoKick and Mind2Onto5 to determine concepts and their hierarchical structure. Having a set of plug-ins available such as a domain lexicon, an inferencing plug-in, and several export and import plug-ins, provides user-friendly customization to use this tool in different ontology development phases. In addition, data about classes, properties, and individuals may be imported or exported via different formats such as RDF and OWL.

4.4.3 WebOnto

Another tool developed by the Knowledge Media Institute of the Open University in England called WebOnto which is an ontology editor for visualization, browsing, and development of

Operational Conceptual Modeling Language (OCML) ontologies (Domingue et al., 1999). Its main advantage over other available tools is that it supports editing ontologies collaboratively, allowing synchronous and asynchronous discussions about the ontologies being developed. Also, it provides a graphical interface that facilitates the automatic generation of instance editing forms from class definitions and inspection of elements taking into account the inheritance of properties and consistency checking. A snapshot of this tool is shown in Figure 4.5.

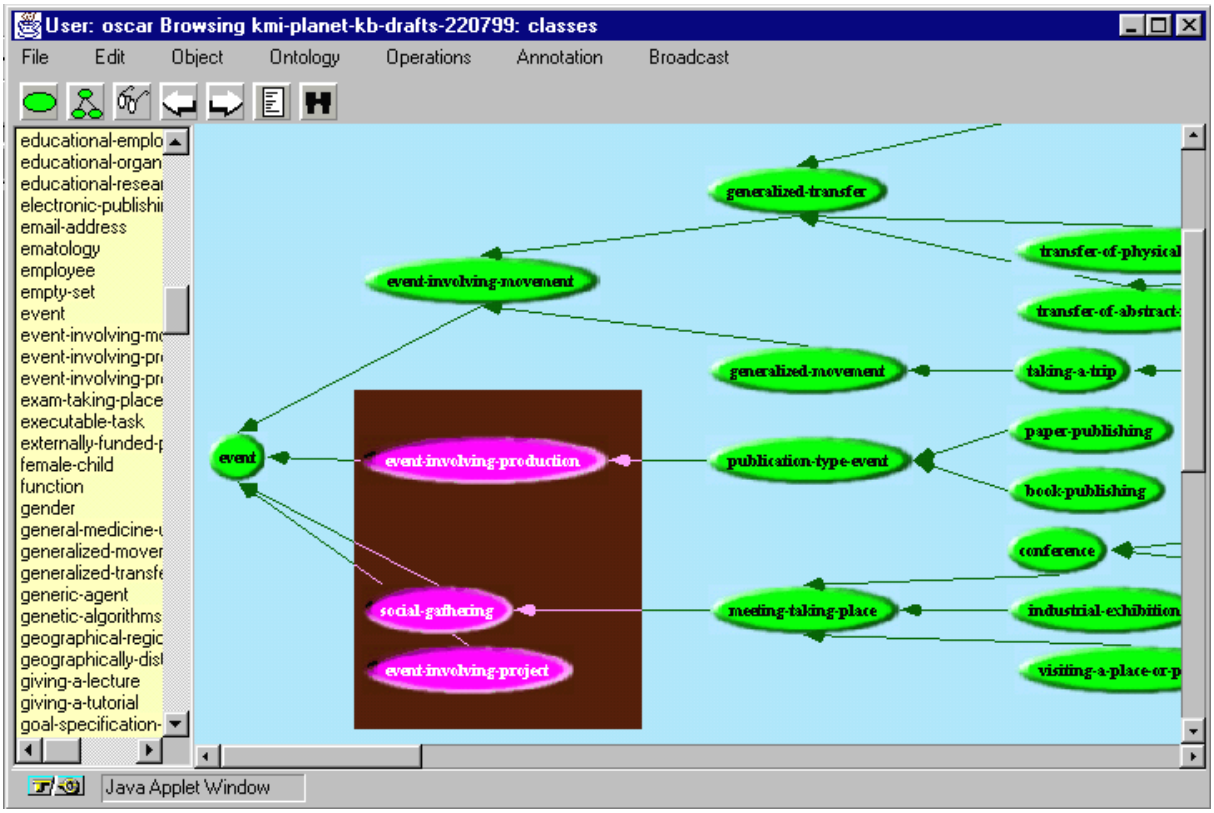


Figure 4.5 Snapshot of WebOnto (Youn and McLeod, 2006)

4.4.4 WebODE

WebODE is a tool built by the Technical School of Computer Science in Madrid as a scalable, extensible, integrated workbench that covers and gave support to most of the activities involved in the ontology development process (Arpirez et al., 2001). Its architecture consists of three tiers, the first tier is the user interface, the second tier is the application server, and the

third tier consists of the database management system. In addition, it provides definitions for the concepts, groups of concepts, relations, constants, and instances. There are several services for ontology language such as importation/exportation, axiom edition, ontology documentation, ontology evaluation, and ontology merging.

4.4.5 Semantic Web Ontology Overview and Perusal (SWOOP)

SWOOP is a simple, scalable, hypermedia inspired OWL ontology browser and editor written in Java and developed by the University of Maryland. It has a reasoning support and provides a multiple ontology environment where different ontologies can be compared against their Description Logic-based definitions, associated properties, and instances (Kalyanpur et al., 2006). SWOOP has main features include browsing and editing multiple ontologies, renderer plugins for OWL presentation syntaxes, semantic search, collaborative annotation, and multimedia markup extension.

4.5 Summary

This chapter highlights the main issues related to ontology learning from a text which may belong to different domains. First, a comprehensive description was introduced about ontology definition and its structural components represented in the layer cake. Second, we have provided a brief state of the art for each layer emphasizing the main approaches used in the development task. Finally, the main features of an important collection of ontology languages and tools were discussed. In the next chapter, we will go deeply in the ontology learning from a text and we will focus on Arabic texts represented by the holy Quran. Furthermore, we will concentrate on the extraction of semantic relationships that hold in particular conjunctive phrases. The chapter will also describe some characteristics of these relations based on Arabic traditional grammars and statistical techniques.

5 Extracting Semantic Relations from the Holy Quran

5.1 Introduction

There is a lack in the developed approaches that deal with ontology learning from texts written in Arabic script due to the nature of Arabic writing, the semantic ambiguity of words, and the shortage in resources and tools that support Arabic (Farghaly and Shaalan, 2009). For Quran ontologies, all studies aim to achieve the purpose of understanding Quran as a source of knowledge and facilitating information retrieval automatically. Therefore, Quran can be presented to the world and employed very efficiently in many linguistic and religious studies. Currently, there are no complete Quran ontologies; many of them have covered specific topics in Quran or special types of words rather than the whole Quran (Saad et al., 2010). Also, many researchers have built ontologies for parts of Quran and very few have used the entire Quran. Moreover, each ontology has focused only on one or two types of relations between terms such as synonymy and Part-Of (Shoaib et al., 2009). As a validation method, the existing ontologies are verified based on a limited procedure which is the domain experts that relied on scholarly sources in their decisions (Ta'a et al., 2013).

In this chapter, we introduce a novel approach that aims at enriching the automatic construction of Quran ontology. We extract the whole relations that exist in the conjunctive phrases. The main contribution is that, we define a hybrid method to extract semantic relations from Quran based on strong and solid rules (Bentrcia et al., 2017). First, we exploit an efficient rule in Arabic grammar, which is AND conjunction, to extract several types of semantic relations. AND conjunction is a well-known grammatical tool that combines terms which have a

degree of association between each other. The proposed set of patterns is used to extract the AND conjunctive phrases from the corpus and not to extract a specific type of semantic relations, as it is known in the pattern-based methods. Second, we use an accurate measurement, which is the correlation coefficient, to find the association value between Quranic Arabic words. This is totally new and worthy in this field and different from other common measurements such as Mutual Information (MI) and t-score. Finally, we combine statistical tests (testing hypotheses and student t-test) and domain experts to validate the results achieved. All the reported approaches in the field of Quran mining (Alrehaili and Atwell, 2014) depend on either domain experts or exegesis books such as Tafsir of Ibn Kathir in the validation process. This step is very essential and cannot be neglected because the holy Quran is a very sensitive and critical text. Basically, we present a scientific validation approach to consolidate the domain experts' decisions and to convince people regardless their religions.

5.2 Ontology learning from the holy Quran

Ontology learning from a specific text means extracting the main terms and relations that represent the domain (Liu et al., 2011). This phase includes two major components: terms extraction and relations extraction. Also, we define another component that seeks extracting conjunctive patterns to accomplish relations extraction task, as detailed in Section 5.2.2.

The whole phases of constructing the Quranic ontology are clarified briefly in Figure 5.1 and in detail in Figure 5.2.

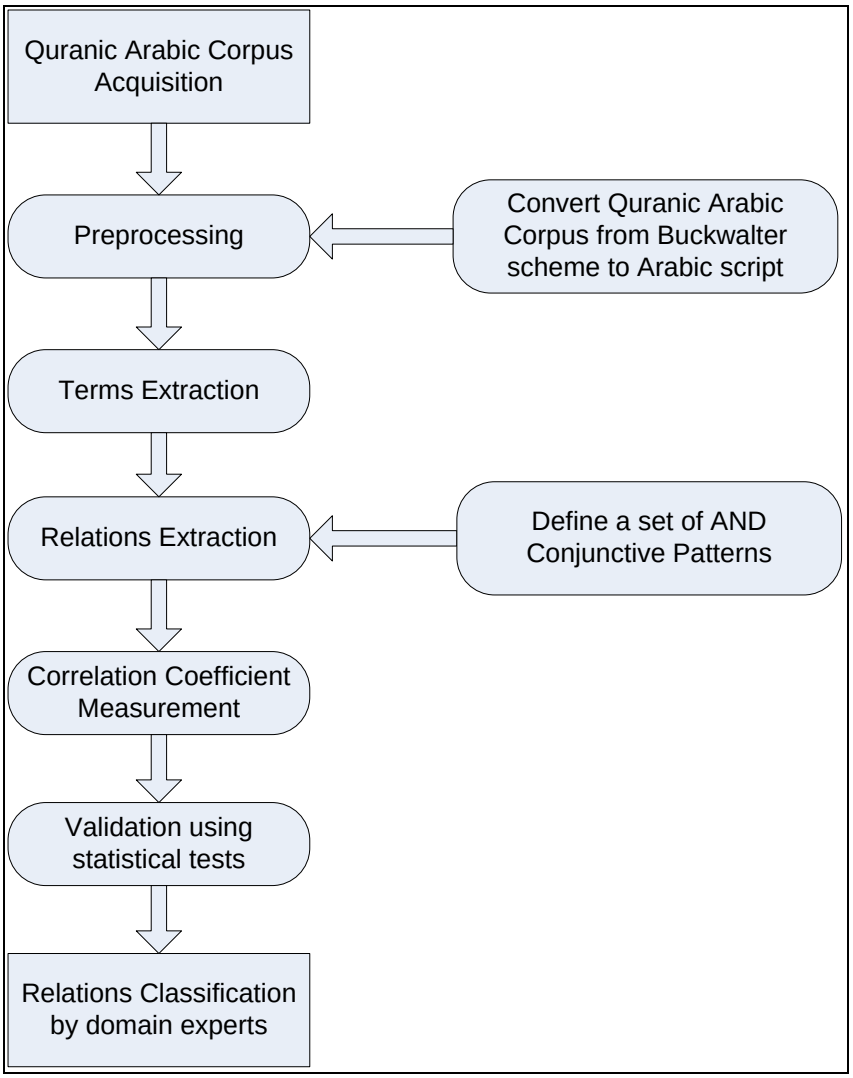


Figure 5.1 Ontology Learning from the Holy Quran Phases

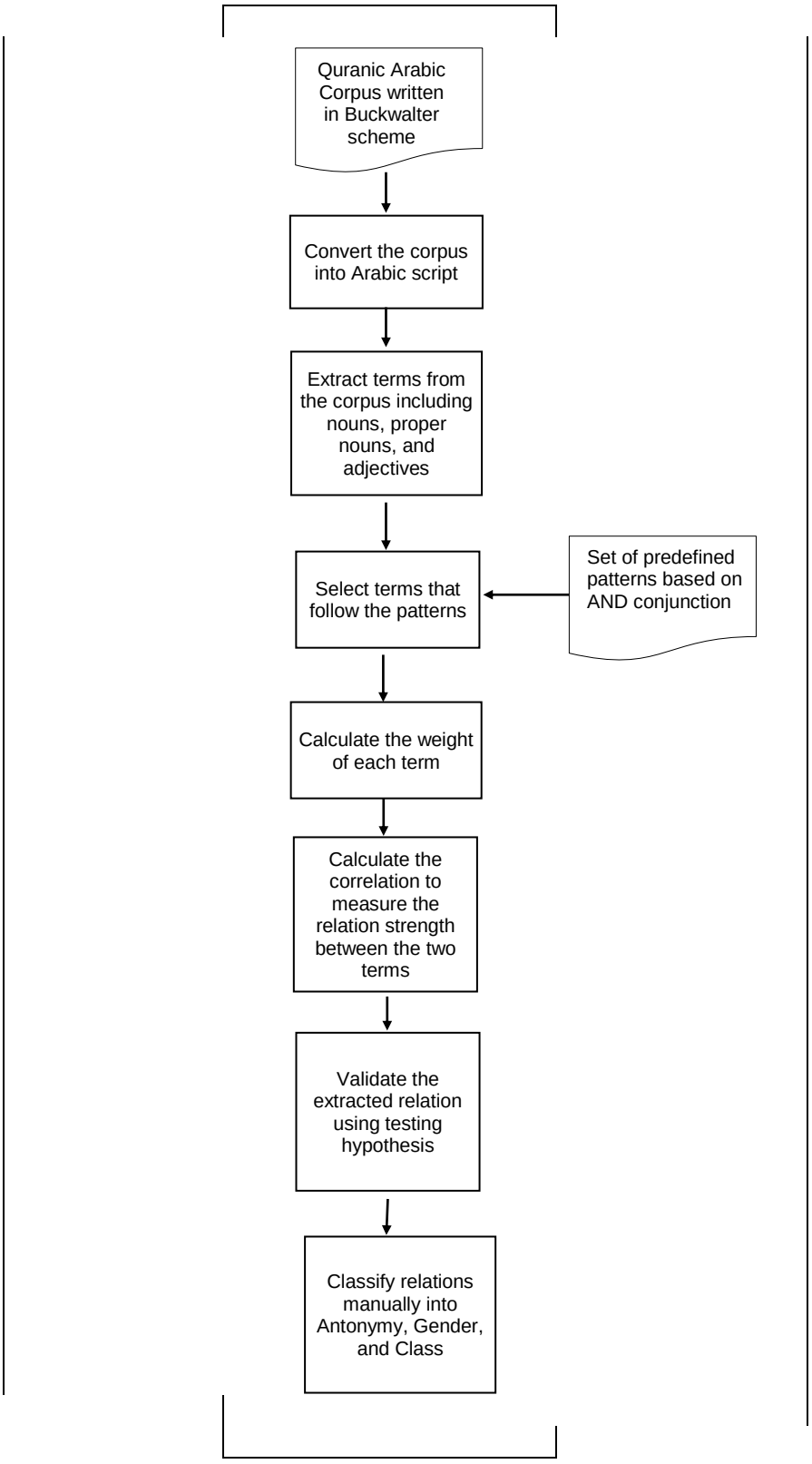


Figure 5.2 Ontology learning from the Holy Quran phases in details

5.2.1 Term Extraction

After the pre-processing phase, mentioned in Chapter 2, we start this phase by extracting terms from the converted-to Arabic corpus (Quranic Arabic Corpus) which include nouns, proper nouns, and adjectives, in their stem form to avoid considering different forms of the same word as different multiple words.

Quranic Arabic corpus is a text file organized into four columns and many rows, as depicted in Chapter 2. To access the file contents, we used the ordinary read/write/search file functions to read the file line by line, search the TAG column looking for words with POS tagging equal to noun, proper noun, or adjective. The stem forms of the resulting words are extracted from the FEATURE column and stored as strings of characters. Finally, we removed the duplicated stems and we stored the unique ones in the term list. Arabic diacritics are characters like letters and we used the same functions to manipulate them. However, they are very important elements that construct Arabic words and distinguish the word's meaning from other words that have similar structure of letters. For these reasons, we did not remove the diacritics but instead, we exploited the POS tagging information available in the corpus and we dealt with the stem form of words. As an example, consider the two Arabic Quranic words (الجَنَّةandالجِنَّة), which have identical letters, different diacritics, and hence different meanings. Based on their stem form (جَنَّةand جِنَّة), we have two different words and not one, which is not the case when we remove the diacritics. For Quranic text, taking the diacritics into consideration with an accurate manipulation is very important and increases the efficiency of the proposed approach.

One basic text mining technique to process textual data is to convert each word in the text into a numerical value that represents word importance in the corpus (Weiss et al., 2005). We achieve this goal by constructing a matrix called term-document matrix where its rows are the extracted Quranic terms and its columns are the Quranic chapters (i.e. surahs). Each term has

a specific weight in each document in the corpus. There is an efficient statistical method to calculate weights called Term Frequency Inverse Document Frequency (tf.idf) (Salton and McGill, 1983):

$$\boldsymbol{w_{ij} = tf_{ij} \cdot log\left(\frac{N}{df_i}\right)} \tag{5.1}$$

where $w_{i,j}$ is the weight of term i in document j, $tf_{i,j}$ is the number of occurrences of term i in document j, N is the total number of documents in the corpus, and df_i is the number of documents containing term i. A high weight in tf.idf is reached by a high term frequency in a given document and a low document frequency of the term in the corpus; the weights hence tend to filter out common terms which are less discriminative.

The result of this process is a matrix of 3267 rows of unique words, namely terms, and 114 columns of Quranic chapters (i.e. surahs). The elements of this matrix are the calculated tf.idf weights of each term in a given chapter.

5.2.2 Conjunctive Patterns Extraction

Arabic grammar is very rich of patterns and clauses that serve different purposes in the sentence. In this work, we call a conjunctive pattern every two terms enclose AND conjunction in between. The considered terms could be nouns, adjectives, or proper nouns. There are nine conjunctives in Arabic, where the two combined terms must have a type of association between each other. However, only six of them have a conjunctive role in the holy Quran, and have been repeated for several times (Adhima, 1972), as shown in Table 5.1.

Table 5.1 The Arabic conjunctions mentioned in the holy Quran

Conjunctive	ثم Then	أو Or	و And	بل But	الفاء Then	أم Or

Before elucidating the conjunctive-based relations, we define a set of conjunctive patterns/ rules based on a deep study of Arabic grammar (Al-Zujaji, 1984; Al-Ghalayini, 2007), POS

tagging, and morphology features found in the Quranic Arabic corpus. We treat only the cases where the two combined terms may be nouns, proper nouns, or adjectives. Other complex cases are beyond the scope of this work because they need specific knowledge resources such as exegesis of the holy Quran. This set is as follows:

1. Noun + Conjunctive "AND" + Noun: this pattern is for extracting any two nouns with AND conjunction in between, such as: "هُدًى و مَوْعِظَةً", which means guidance and instruction. Different cases of this pattern are explained bellow:

   a. Noun + Conjunction "AND" + Noun + Determinant 'ال'+ Noun: the two combined nouns are followed by a third noun which starts with a determinant, like " إِحْسَانًا و ذِي الْقُرْبَى", which means good and to relatives.

   b. Noun + Conjunction "AND" + Noun + Noun: the two combined nouns are followed by a third noun, like "فِتْنَةً و اِبْتِغَاءَ تَأْوِيلِهِ", which means discord and seeking an interpretation.

   c. Noun + Conjunction "AND" + Noun + Determinant 'ال' + Adjective: the two combined nouns are followed by an adjective which starts with a determinant, like " الْأَرْضُ و السَّمَاوَاتُ الْعُلَى", which means earth and highest heavens.

   d. Noun + Conjunction "AND" + Noun + Adjective: the two combined nouns are followed by an adjective like "بُهْتَانًا و إِثْمًا مُبِينًا", which means injustice and manifest sin.

2. Adjective + Conjunction "AND" + Adjective: this rule is for extracting any two adjectives with AND conjunction in between, such as "ثَيِّبَاتٍ و أَبْكَاراً", which means previously married and virgins.

3. Proper Noun + Conjunction "AND" + Determinant 'ال' + Proper Noun: this rule is for extracting any two Proper nouns with AND conjunction in between, and the second one starts with a determinant, such as "يَعْقُوبَ و الْأَسْبَاطَ", which means Jacob and the Descendants.

4. Proper Noun + Conjunction "AND" + Proper Noun: this rule is for extracting any two Proper nouns with AND conjunction in between, for example: "إِبْرَاهِيمَ و إِسْحَاقَ", which means Abraham and Isaac.
5. Proper Noun + Conjunction "AND" + Determinant 'ال' + Noun: this rule extracts any Proper noun followed by a noun which starts with a determinant, for example: " نُوحٍ و النَّبِيِّينَ", which means Noah and the prophets.
6. Noun + Pronoun + Conjunction "AND" + Noun + Pronoun: this rule extracts any two nouns combined with AND, and the first noun ends with a Pronoun, for example: " مَحْيَاهُمْ و مَمَاتُهُمْ", which means their life and their death.

Moreover, we define a set of negative conjunctive patterns, where the negation tool 'لا, NOT' is used with the conjunction "AND", as clarified next.

7. Negation 'NOT لا' + Adjective + Conjunction "AND" + Negation 'NOT لا' + Adjective: this pattern finds out any two negative adjectives combined with AND conjunction, such as "لاَ بَارِدٌ وَ لاَ كَرِيمٌ", which means neither cool nor beneficial.
8. Negation 'NOT لا' + Determinant 'ال' + Noun + Conjunction "AND" + Negation 'NOT لا' + Determinant 'ال' + Noun: this pattern finds out any two negative nouns combined with AND conjunction, such as "لاَ الْهَدْيَ وَ لاَ الْقَلاَئِدَ", which means the sacrificial animals and garlanding.
9. Adjective + Conjunction "AND" + Negation 'NOT لا' + Adjective: this pattern finds out any two adjectives combined with AND, where the second one is directly preceded by a negation, such as "صَغِيرَةٌ و لَا كَبِيرَةٌ", which means small or large.
10. Noun + Conjunction "AND" + Negation 'NOT لا' + Noun: this pattern finds out any two nouns combined with AND, where the second one is directly preceded by a negation, such as "مَالٌ وَ لاَ بَنُونَ", which means wealth or children.

5.2.3 Relation Extraction

Different methods have been proposed in the past to find semantic relations between words in a corpus. All of them belong to one of three categories. The first category includes methods which are based on finding pair of words that may occur together more often than expected by chance (collocations) using statistical tests (Maedche and Staab, 2001). However, the resulting relations depend only on statistical analyses which give less precise decisions in relations extraction and validation.

In the second category, researchers have exploited syntactic dependencies, in particular, the dependencies between a verb and its arguments to detect relations. One problem is how to find a general presentation of the verb arguments regardless the text from where they are extracted (Cimiano, 2006). The third category methods rely on lexico-syntactic patterns to detect very specific types of relations such as part-of and cause (Hearst, 1992).

The main drawback of these methods is the complexity of pattern construction. It is time and effort consuming since for each type of relations, a set of patterns is developed and applied in a specific form and order.

Our proposed approach is a hybrid of pattern-based methods and statistical methods. However, the pattern-based methods depend on using one/many pattern(s) to extract one specific type of semantic relations. The more types of semantic relations we want to extract, the more patterns we should use. On the other hand, our approach uses a limited set of patterns not to extract a specific type of semantic relations but to extract AND conjunctive phrases from the Quranic Arabic Corpus. Each pattern may extract several types of semantic relations which reduces time and effort complexity. For example, the pattern (Noun + Conjunctive "AND" + Noun) extracts three types of relations: Antonymy (مَوْت وَ حَيَواة), Gender (مُؤْمِنِين وَ مُؤْمِنَات), and Class (تِّين وَ زَّيْتُون), as we will discuss in Section 5.2.6.

In order to extract the conjunctive phrases, we search the FORM and the TAG columns in the Quranic Arabic corpus looking for AND conjunction to extract the two terms that occur on the both sides of AND. These phrases that consist of AND conjunction and the two terms are considered as conjunctive phrases only and only if they match one of the patterns defined in the previous section. Each conjunctive phrase indicates the existence of a probable semantic relation which needs further processing, as explained in the next sections.

In our approach, we search the Quranic Arabic corpus, specifically the terms extracted in Section. 5.2.1, looking for those that form 'AND' conjunctive phrases and match one of the defined patterns.

The tables bellow illustrate some examples of conjunctive phrases, where the two combined terms are adjectives as shown in Table 5.2, nouns as shown in Table 5.3, and proper nouns as shown in Table 5.4.

Table 5.2 Sample of conjunctive adjectives

Adjective 1 AND Adjective 2	**English Translation**
صَغِير وَ كَبِير	Small and Big
أَظْلَم وَ أَطْغَى	Unjust and Rebellious
أَعْجَمِيّ وَعَرَبِيّ	Foreign Tongue and Arab
بَشِير وَ نَذِير	Bearer of glad tidings and Warner
أَعْمَى وَ بَصِير	Blind and Seeing
لَا فَارِض وَ لَا بِكْر	Neither old nor young
سَاجِد وَ *قَائِم*	Prostrating and Standing

Table 5.3 Sample of conjunctive nouns

Noun 1 AND Noun 2	**English Translation**
جَنَّة وَحَبّ	Garden and Grain
جَنَّة وَ حَرِير	Garden and Silk
جَنَّة وَ عَيْن	Garden and Spring
جَنَّة وَ مَّغْفِرَة	Garden and Forgiveness
جَنَّة وَ نَعِيم	Garden and Bliss
جَنَّة وَ نَهَر	Garden and Stream

Table 5.4 Sample of conjunctive proper nouns

Proper Noun1 AND Proper Noun 2	**English Translation**
يَأجُوج وَ مَأجُوج	Gog and Magog
جِبْرِيل وَ مِيكَال	Gabriel and Michael
مُوسَى وَ هَارُون	Moses and Aaron

إِنجِيل وَ قُرْءَان	Gospel and Qur'an
فِرْعَوْن وَ هَامَان	Pharaoh and Haman
إِسْحَاق وَ يَعْقُوب	Isaac and Jacob
دَاوُد وَ سُلَيْمَان	David and Solomon
عَاد وَ ثَمُود	'Ad and Thamud

Moreover, we discover a special case of combinations where one term is associated with many different terms, as they occurred in the holy Quran. For example, the term Garden 'جَنَّة' is combined with six different terms as illustrated in Table 5.3.

5.2.4 Correlation Coefficient

One main feature of AND conjunctive is that the two combined terms must hold a kind of correlation between each other (AL-Taweel, 2009), and to find how much two terms are related to each other, we propose a powerful method called Pearson Product-Moment Correlation Coefficient (*r*). It allows researchers to investigate naturally the relation between any two variables very efficiently and interpret the results clearly without any misleadingly incorrect values.

In statistics, (*r*) is defined as a measure of the degree of linear relationship between two variables $\boldsymbol{x}$ and $\boldsymbol{y}$ (Myatt, 2007):

$$r = \frac{\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{x})(\boldsymbol{y}_i-\bar{y})}{(n-1)s_x\boldsymbol{s}_y} \tag{5.2}$$

where $\boldsymbol{x}_i$ are the values of $\boldsymbol{x}$, $\boldsymbol{y}_i$ are the values of $\boldsymbol{y}$, $\bar{x}$ is the mean of the $\boldsymbol{x}$ variable, $\bar{y}$ is the mean of the $\boldsymbol{y}$ variable, s_x and $\boldsymbol{s}_y$ are the standard deviations of the variables $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. (*r*) value ranges between -1 and +1 and its sign defines the direction of the relationship, either positive (+) or negative (-), whereas the absolute value of the correlation coefficient measures the strength of the relationship. Thus, we apply this method to each conjunctive phrase extracted in Section 5.2.2. In this case, the two variables $\boldsymbol{x}$ and $\boldsymbol{y}$ are the two combined terms, and their elements are the tf.idf weights, found in the term-document matrix. As a result, we find either a positive correlation coefficient, which means that as the

weight of one term increases, the weight of the other term increases; as one decreases the other decreases or a negative correlation coefficient, which indicates that as one term's weight increases, the other decreases, and vice-versa. The values of -1 and +1 mean a perfect linear relationship between the two terms, while the zero value indicates the absence of this type of relation. Table 5.5 demonstrates a sample of conjunctive phrases with high positive correlation. Term 1 and Term 2 may share a strong relationship. For example, the two terms Heaven and Earth have a high correlation because they often appear together as a conjunctive phrase in the Quranic verses.

Table 5.5 Sample of conjunctive phrases with high correlation

Term 1 AND Term 2	**English Translation**	**Correlation Coefficient**
رَعْد وَ بَرْق	Thunder and Lightning	0.6115
سَمَاء وَ أَرْض	Heaven and Earth	0.8485
شَاهِد وَ مَشْهُود	Witness and Whom witness has been borne	0.7698

Table 5.6 demonstrates a sample of conjunctive phrases with low positive correlation. Term 1 and Term 2 may share a weak relationship. This is due to the low occurrence percentage of the two terms together compared to their occurrence separately. The term Allah has a weak relation with the term day since we find the term Allah almost in every verse in the Quran and this is not the case for the term day.

Table 5.6 Sample of conjunctive phrases with low correlation

Term 1 AND Term 2	**English Translation**	**Correlation Coefficient**
ظُلُمَات وَ رَعْد	Darkness and Thunder	0.2068
ذِلَّة وَ مَسْكَنَة	Abasement and Destitution	0.1325
الله وَ يَوْم	Allah and Day	0.1678

Table 5.7 demonstrates a sample of conjunctive phrases with close to zero correlation. Term 1 and Term 2 may share no relationship. This set of combined terms may appear together very rarely. In contrary, each term may appear alone or combined with a different term many

times. As an example, the two terms guidance and light may have no relation because the term guidance is also associated with many other terms such as good tidings بُشْرَى, reminder ذِكْرَى, mercy رَحْمَة, healing شِفَاء, criterion فُرْقَان, and instruction مَّوْعِظَة.

Table 5.7 Sample of conjunctive phrases with close to zero correlation

Term 1 AND Term 2	English Translation	Correlation Coefficient
وَجْه وَ يَد	Face and Hand	0.0844
هُدًى وَ نُور	Guidance and Light	0.0054-
هُزُو وَ لَعِب	Jest and Sport	0.0139

In addition, there is a perfect set of conjunctive phrases where both of the two terms occur only together equal number of times. As a result, their correlation coefficient is 1 as shown in Table 5.8.

Table 5.8 Sample of conjunctive phrases with correlation equal to 1

Term 1 AND Term 2	English Translation	Correlation Coefficient
شَفْع وَ وَتْر	Even and Odd	1
شِتَاء وَ صَّيْف	Winter and Summer	1
جَلَال وَ إِكْرَام	Glory and Honor	1

5.2.5 Validation Phase

Texts in general and Quranic Arabic texts in particular can be understood by scholars from different aspects. This reason leads to different linguistic and religious extractions. In order to assess such results, we find that adopting statistical techniques is very useful to give an approximate decision about problems related to text processing.

In our work, because the weight of each term is based on its frequency in the corpus, the correlation coefficient is then highly dependent on this factor. One approach to ensure that two terms are together because of a type of relationship and not due to chance is to use statistical hypotheses testing (Kass et al., 2014) and test the significance of the correlation coefficient.

We suggest two mutually exclusive hypotheses called null hypothesis H_0 and alternative hypothesis H_1.

H_0: There is no correlation between the two terms...it is due to chance.

H_1: There is a significant correlation between the two terms.

Next, we test the two hypotheses to either reject the null hypothesis or accept by applying a statistical test named Student's t test (Siegmund, 1998):

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2} \tag{5.3}$$

It returns a value *t* which shows the validity of null hypothesis. The smaller the t-value, the weaker is the evidence against the null hypothesis. Then, we compare the t-value to an acceptable significance threshold *a* =1.984, taken from statistical t-test tables (Verma, 2013). The tabulated value of *a* is required for significance at .05 level of significance and *n* -2 degree of freedom, *n* is equal to 114, which is the size of each term vector, and *r* is the calculated correlation coefficient obtained from formula (5.2). If $t > a$, the correlation coefficient is statistically significant, the null hypothesis may be rejected and the alternative hypothesis is valid. Otherwise, if $t \leq a$, the null hypothesis is failed to be rejected. Table 5.9 clarifies a sample of probable accepted and rejected conjunctive relations, after applying t-test.

Table 5.9 Sample of accepted and rejected conjunctive relations R after applying t- test

Term 1	**AND**	**Term 2**	**Correlation Coefficient *(r)***	**t-value**	**t-test Decision**
جِنّة Jinn	و	نّاس Men	0.5711	7.3628	R may be accepted
فَاكِهَة Fruits	و	أب Herbage	0.4993	6.0987	R may be accepted
ضُحَى Morn	و	لَيْل Night	0.4006	4.6271	R may be accepted
سَمَاء Heaven	و	طَارِق Morning Star	0.0768	0.8152	R may be rejected
كَذّاب Liar	و	كُل Every	0.0323-	-0.3420	R may be rejected
الله Allah	و	فَتْح Victory	0.0627	0.6649	R may be rejected

5.2.6 Experimental Results

We use the Quranic Arabic corpus which consists of 77,430 words. We start by words that may represent Quranic domain such as nouns, proper nouns and adjectives. Basically, we find a set of 31007 of repeated words, filtered to 3267 of unique terms.

Besides that, conjunctive phrases occurred in the holy Quran almost 2000 times. After eliminating the repeated ones, we get a set of 1047 unique phrases and hence probable relations.

Because the automatically learned ontologies are highly error prone, there is an immense need to domain-specific experts to inspect them, validate, and modify before they can be applied. We may suggest a filtering method to select the most representative relations by defining a threshold and select the phrases with correlation coefficient greater than this threshold. Although we find that the statistical method t-test is very efficient in the filtering process, but threshold- based method can also be used as a next step.

Compared to the previous approaches mentioned in Section 5.2.3, our novel method is a hybrid of the statistical approaches and the lexico-syntactic patterns approaches. However, it is based on a strong Arabic grammar, which is AND conjunctive, to define a small set of patterns that ensures the existence and the extraction of many types of semantic relations from Quran very efficiently. Furthermore, we exploit the statistical methods to measure the strength of the extracted relations and to aid domain experts in estimating the final decision about semantic relations. Classifying each relation as antonymy, gender, or class is performed manually looking for doing it automatically in the future. The proposed approach achieved more accurate and comprehensive results.

The extracted relations are classified manually into three categories:

5.2.6.1 Antonymy

Antonymy is the semantic relation between antonyms which are words with opposite meaning. In Table 5.10, this category of relations includes antonyms that are combined by AND conjunctive. For example, we find the term Sky ¨سَمَاء ¨ is combined with its antonym Land ¨أَرْض ¨ and the term secretly ¨ سِر ¨ is combined with two different antonyms: Openly ¨ جَهْرّ ¨ and openly ¨ عَلَانِيَة ¨.

Table 5.10 Sample of conjunctive phrases with Antonymy relation

Term 1 AND Term 2	English Translation
سَمَاء وَ أَرْض	Sky and Land
سِر وَ جَهْرّ	Secretly and Openly
سِرّ وَ عَلَانِيَة	Secretly and Openly
شَرّ وَ خَيْر	Evil and Good
شَقِيّ وَ سَعِيد	Unhappy and Happy
شِتَاء وَ صَيْف	Winter and Summer
ضَرَّاء وَ سَرَّاء	Adversity and Joy
لَيْل وَ نَهَار	Night and Day
مَشْرق وَ مَغْرب	East and west
مَوْت وَ حَيَواة	Death and Life

5.2.6.2 Gender

This category refers to the relation that exists between masculine and feminine words. Such relation is very common in the holy Quran where God converses both male and female at once. Table 5.11 lists a sample of conjunctive phrases that combine masculine and feminine terms together such as the masculine term charitable men مُصَدِّقِين and its feminine charitable women مُصَدِّقَات.

Table 5.11 Sample of conjunctive phrases with Gender relation

Term 1 AND Term 2	English Translation
مُؤْمِنِين وَ مُؤْمِنَات	The believing men and believing women
مُتَصَدِّقِين وَ مُتَصَدِّقَات	The charitable men and charitable women
مُسْلِمِين وَ مُسْلِمَات	The Muslim men and Muslim women
مُصَدِّقِين وَ مُصَدِّقَات	The charitable men and charitable women
مُنَافِقُون وَ مُنَافِقَات	The hypocrite men and hypocrite women
صَابِرِين وَ صَّابِرَات	The patient men and patient women
صَادِقِين وَ صَّادِقَات	The truthful men and truthful women
صَّائِمِين وَ صَّائِمَات	The fasting men and fasting women
قَانِتِين وَ قَانِتَات	The obedient men and obedient women
بَنِين وَ بَنَات	Sons and daughters

5.2.6.3 Class

A different category of semantic relations consists of terms that belong to the same class because they share the same characteristic features. The holy Quran is full of such examples that are combined by AND conjunctive. Table 5.12 mentioned few of them such as the terms seven and eighth ¨سَبْعَة وَ ثَامِن¨ which belong to the class Numeral. Also, the class Book includes the terms Gospel and Quran ¨إِنجِيل وَ قُرْءَان¨ and the class Animal includes Cow and Sheep ¨بَقَر وَ غَنَم¨.

Table 5.12 Sample of conjunctive phrases with Class relation

Term 1 AND Term 2	**English Translation**
اَللَات وَ اَلْعُزَى	Lat and Uzza
أَبَارِيق وَ كَأس	Ewers and Cups
أصْوَاف وَ أَوْبَار	Wool and Furs
أنف وَ أَذُنٌ	Nose and Ears
أَيُّوب وَ يُوسُف	Job and Joseph
إنجِيل وَ قُرْءَان	Gospel and Qur'an
بَقَر وَ غَنَم	Cow and Sheep
تِين وَ زَيْتُون	Fig and Olive
ذَهَب وَ فِضَّة	Gold and Silver
سَبْعَة وَ ثَامِن	Seven and Eighth

It is clear in our novel approach that we have used only one type of patterns, namely conjunctive, to extract different types of semantic relations. In order to categorize them, we could train classifiers for each type of relations and combine their results and test on different types of extracted ontological relations.

5.3 Ontology evaluation

The process of testing a constructed ontology is very important to avoid applications from using inconsistent or even redundant ontologies. (Brank et al., 2005) proposed four approaches of ontology evaluation:

- Data-driven evaluation where the ontology is compared to a source of data that covers the specific domain to ensure the relatedness of the ontology.

- Application based evaluation where the ontology is exploited in a specific application which its performance and accuracy are used to evaluate the ontology.
- Gold standard which is a predefined ontology, built manually by experts for a specific domain to be compared to the new developed ontologies lexically and conceptually.
- Human evaluation is the most common approach where the developed ontology is verified manually by domain experts to test whether it fits to specific requirements.

For Quranic Arabic, this task is more critical and complicated due to the shortage in the evaluation resources for Arabic language. To assess our approach in extracting semantic relations based on AND conjunction, we did not find a complete or even similar gold standard for Quranic Arabic to verify and compare our results with. Moreover, the existing resources such as the Quranic Arabic Corpus are missing other important features related to Arabic language and Quran in order to be used perfectly. Using human evaluation is also a challenging task due to two main reasons:

- Finding experts in the field of Quran who are experts in Arabic grammar as well, or vice versa, is not easy since the two fields are huge and may intersect in just few elementary basics.
- The existence of different interpretations for the same result is very common in Quran studies especially those based on natural language processing and semantics. Hence, domain experts may return different decisions according to their background and knowledge level. This issue raises the need for another method to support experts' decisions.

The process of extracting conjunctive phrases is totally dependent on the morphological annotation of the Quranic Arabic Corpus, which achieves an accuracy rate of 99%, and on the predefined set of conjunctive patterns which relies on correct Arabic grammars. This provides

a very accurate set of conjunctive phrases which should be tested and validated statistically to select those that may hold strong or weak semantic relations.

To evaluate the accuracy of the relation extraction process, we have used the two performance metrics: precision and recall. In our work, precision is defined as the ratio of the number of relevant retrieved conjunctive relations to the number of retrieved conjunctive relations whether relevant or not, whereas the recall is defined as the ratio of relevant retrieved conjunctive relations to the total number of all relevant conjunctive relations that exist in Quran. The system retrieves 1047 semantic relations based on the predefined conjunctive patterns, 57% are statistically classified as strong relations and 43% are classified as weak relations.

Furthermore, the extracted relations are validated manually by domain experts. The system achieves a precision of 84% and a recall of 92%. More details about the evaluation results are described in Table 5.13.

Table 5.13 The evaluation details of the system

Total number of retrieved relations by the system	1047
Total number of relevant relations retrieved by the system and validated by domain experts	878
Total number of all relevant relations that exist in the Quran	950
Precision	84%
Recall	92%

Using testing hypothesis to validate the correlation results is statistically very sufficient to get precise and reliable results. However, some ambiguous cases appear in two scenarios:

- When a verse is terminated by a noun and the next verse starts by (AND) and a noun. Although this situation follows the first pattern and gives many correct results as shown in the first two examples in the Table 5.14, it may also produce erroneous ones like the other two examples. The extracted conjunctive phrases appear in bold.

Table 5.14 The first scenario of ambiguous conjunctive phrases

The ambiguous case	The achieved result	The correct result
‘و التين و **الزيتون**(1) **و طور** سينين(2)’	Accepted relation	Accepted relation
‘و **الضحى**(1) **و الليل** إذا سجى(2)’	Accepted relation	Accepted relation
‘...و ما أرسلناك عليهم **وكيلا**(54) **و ربك** أعلم بمن في السموات و الأرض...’	Accepted relation	Rejected relation
‘...و أكثرهم **الكافرون**(83) **و يوم** نبعث من كل أمة شهيدا...’	Accepted relation	Rejected relation

This problem also appears when both nouns occur in one verse. These irrelevant cases of erroneous relations happened due to mistakes in the annotation of the Quranic Arabic corpus specifically in the Arabic traditional grammar (الإعراب). Although this corpus has achieved an accuracy rate of 99%, these errors still exist due to the difficulty of the Arabic language and the lack of efficient validation methods.

- When the first term in the conjunctive phrase is separated from the second term by an intermediate term. This shifts the actual conjunctive phrase to a new phrase consists of the intermediate term AND the second term, which may give a result of an erroneous phrase instead of the real one. The examples of Table 5.15 show the actual and the erroneous conjunctive phrases extracted by the proposed method.

Table 5.15 The second scenario of ambiguous conjunctive phrases

The ambiguous case	The actual conjunctive phrase	The erroneous conjunctive phrase
‘ذلكم **أقسط** عند الله **و أقوم** للشهادة **و أدنى** ألا ترتابوا’	‘أقسط و أقوم’	‘الله و أقوم’
‘فمن لم يجد فصيام **ثلاثة** أيام في الحج **و سبعة** إذا رجعتم’	‘ثلاثة و سبعة’	‘الحج و سبعة’
‘**تفريقا** بين المؤمنين **و إرصادا** لمن حارب الله و رسوله من قبل’	‘تفريقا و إرصادا’	‘المؤمنين و إرصادا’
‘إن **الخزي** اليوم **و السوء** على الكافرين’	‘الخزي و السوء’	‘اليوم و السوء’

This problem is also shown due to the lack of grammatical details in the annotated Quranic corpus especially those related to the conjunctive phrase. Although the annotation specifies the type of the tool AND whether it has a conjunctive role or not and the type of the two terms combined by AND, it does not provide much information about their grammatical positions in the conjunctive phrase (المعطوف و المعطوف عليه). This problem leads to extracting wrong conjunctive phrases and hence irrelevant semantic relations.

We conclude that the exploited AND conjunction rule is sufficient to ensure the existence of semantic relations. However, we used statistical techniques to measure the strength of each relation and to help domain experts to decide when conflicts occur.

5.4 The need for Quran experts

Recent Quranic studies need accurate methods to validate the achieved results due to the difficulty of understanding Quranic Arabic text. These methods could be applied indirectly by exploiting Islamic religious books such as Quran interpretations or directly by passing the

results to famous scholars (experts) who perform the validation process based on their knowledge. There are three basic conditions, mentioned in (Al- Soyouti, 1973) that must be available in these experts to be qualified to do this critical job properly. The first condition is the deep knowledge of Arabic language including vocabulary, grammar, and rhetoric. The second condition is related to the knowledge of Quran basics like particulars of Quran revelation, fundamentals of religion and Fiqh, and talks of Prophet Mohammad PBH (Hadith). The last condition is the talent which God gives to every one working sincerely for Islam.

5.5 Summary

Quran ontologies aim at improving the performance of information retrieval systems that deal with Quran. However, current Quran ontologies have a limited construction due to several criteria. In this work, we have exploited the Arabic conjunctive patterns that exist in the traditional Arabic grammar to extract different types of semantic relations from the entire Quran and enrich the automatic construction of the Quran ontology. We have applied correlation coefficient method to measure the strength of the linear relationship which may exist between every pair of nouns, proper nouns, and adjectives that form a conjunctive phrase. Furthermore, we have suggested hypotheses testing and Student t-test to go beyond chance and validate the significance of the extracted relations. We have unveiled manually three categories of semantic relations: antonymy, gender, and class. In future work, we can exploit classifiers to perform this task automatically. Finally, we insist that such a field of research needs statistical techniques besides the domain experts to assess the results achieved.

6 Quran Mining: The Order of Words in AND Conjunctive Phrases

6.1 Introduction

Text mining concept can be defined as "the analysis of observational textual data sets to find un-suspected relationships and to summarize the text in novel ways that are both understandable and useful to the users" (Hand et al., 2001). Word co-occurrences are considered as one of the most powerful text mining approaches that is used to extract statistical and associational relationships from textual documents (Bullinaria and Levy, 2007). Generally, two words co-occur if they are observed together in a given unit of text. However, the unit of text can be a window of a fixed number of words, or a sentence, or a group of sentences that may form a small paragraph or a document. Moreover, different text mining methods have been developed and applied in different fields such as information retrieval, which is widely used to answer queries like the case in search engines (Xu and Croft, 2000). In addition, various ontology-based information extraction systems are also based on such methods either to extract keywords from a specific domain or to find the relationships among them (Abulaish and Dey, 2007).

Text mining and natural language processing methods are highly cooperated to extract information from text where such information is presented in an unstructured format that is not immediately suitable for automatic analysis by a computer. Applying text mining techniques, supported by machine learning methods, can play a significant role to extract useful information which provides potential benefits for a lot of applications such as text categorization, concept/entity extraction, and entity relation modeling.

On the other hand, researchers are also starting to exploit the text mining approaches to extract knowledge from sacred texts such as the holy Quran and the Bible in order to get better understanding of the Islamic and Christian religions (Banchs, 2013). Nevertheless, there is a lack in these approaches that deal with texts written in Arabic script due, for example, to the nature of Arabic writing, the semantic ambiguity of words, and the shortage in resources and tools that support Arabic (Farghaly and Shaalan, 2009). For Quran mining, all studies aim to achieve the purpose of understanding Quran as a source of knowledge and extracting useful information automatically. Therefore, Quran can be presented to the world and exploited very efficiently in many scientific, linguistic, and religious applications. Although few studies have been conducted in the literature on Arabic text mining (Saif and Aziz, 2011; Al-Kabi et al., 2013; Alrabiah et al., 2014), only very few mined Quranic Arabic text. Currently, the existing approaches to mine Quran are divided into computational and statistical methods where statistics are used to extract information from Quran such as word co-occurrence, Quran concordance, and verses similarity (Al-Kabi et al., 2005; Panju, 2014), and morphological and syntactical methods where Quran is analyzed to extract lexical and semantic information, or to construct a knowledge representation model such as ontologies and treebanks (Dukes and Habash, 2010; Dukes and Buckwalter, 2010).

6.2 Quran mining approaches

The main objective of text mining approaches is the extraction of knowledge from the processed textual data in order to be exploited in many useful applications. In this section, we introduce three main approaches that seek mining Quran and we highlight some of the reported works in each field:

6.2.1 Visualization

Data visualization is a machine learning technique that used to represent data in a visual format. For text documents, this method is very helpful in discovering patterns and relations among disparate terms or documents in a corpus. (Alhawarat et al., 2015) introduced a study that applies various text mining approaches to the Quranic text and displays the results in a graphical representation. In the preprocessing stage, the text of the holy Quran is divided into five different parts and converted to CP1256 code. Next, stop words removal and stemming procedure are applied to produce a corpus which is converted to Term- Document and Document-Term matrices. Different experiments are conducted based on Chapters and Parts partitioning methods in order to provide the most frequent terms, word cloud, and clusters that exist in the holy Quran.

In experiments based on Quran chapters, the 114 chapters are used to extract the most frequent terms based on TF measure as shown in Figure 6.1, and based on TF-IDF as shown in Figure 6.2.

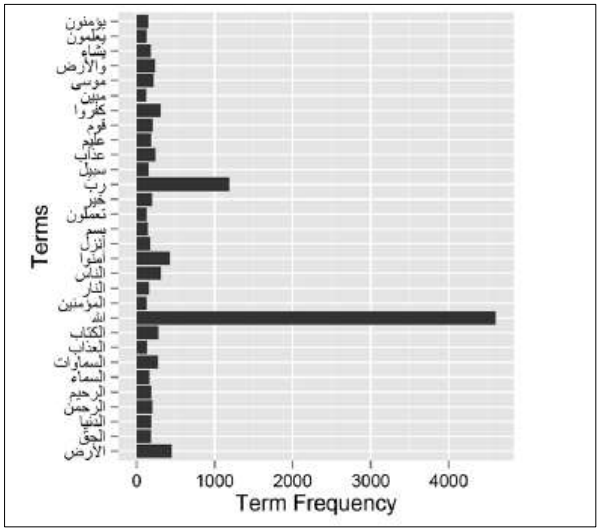


Figure 6.1 Most frequent terms in the holy Quran measured by TF (Alhawarat et al., 2015)

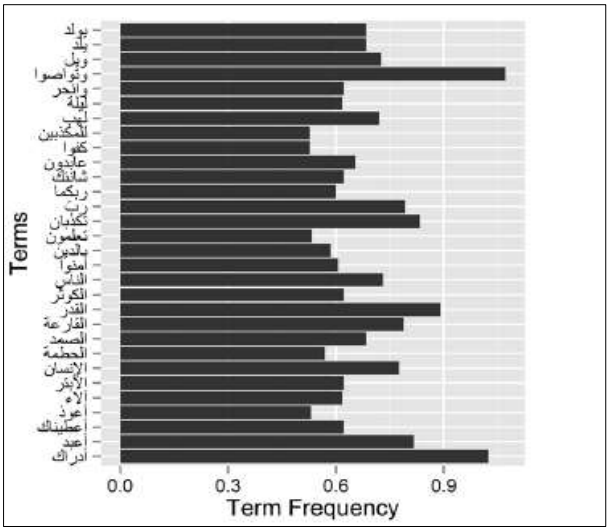


Figure 6.2 Most frequent terms in the holy Quran measured by TF-IDF (Alhawarat et al., 2015)

Moreover, word cloud visualization for the 100 most frequent words shown previously is depicted in Figures 6.3 and 6.4.

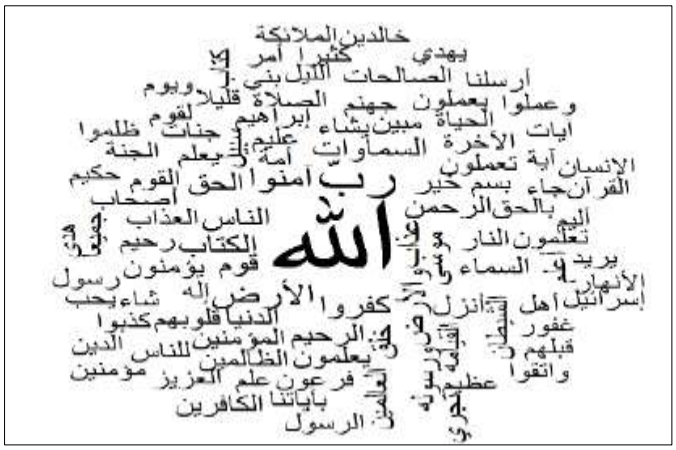


Figure 6.3 Word cloud for the most frequent 100 words in the holy Quran measured by TF (Alhawarat et al., 2015)

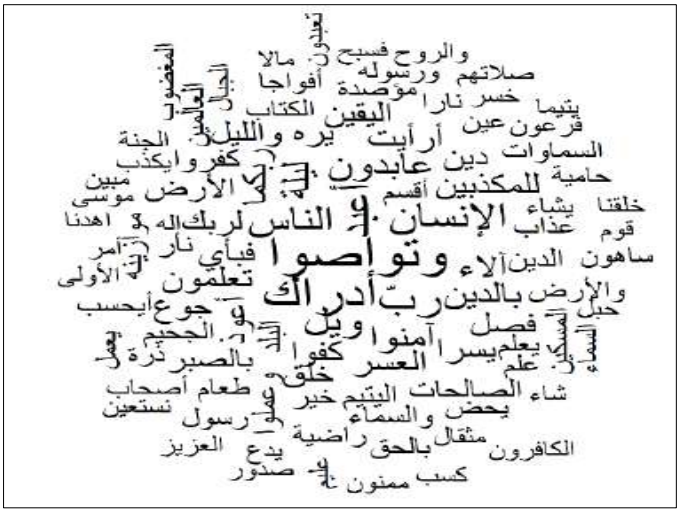


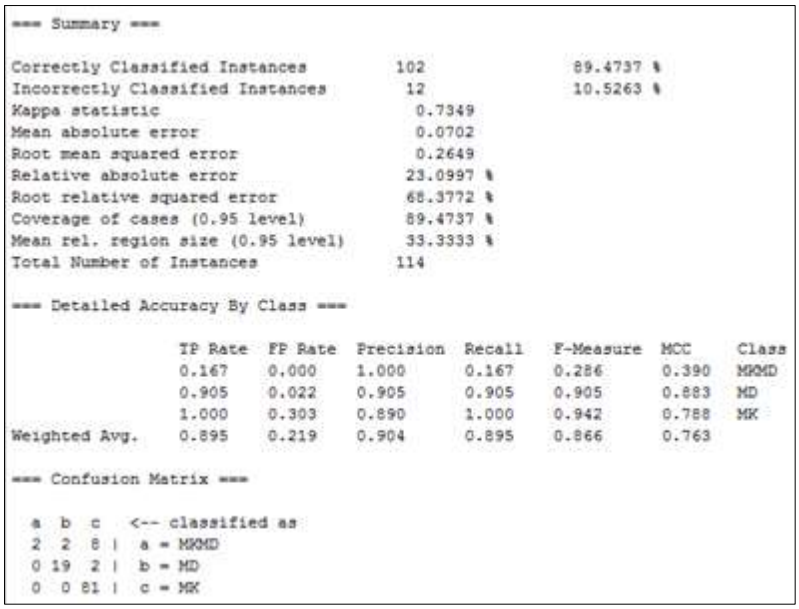
Figure 6.4 Word cloud for the most frequent 100 words in the holy Quran measured by TF-IDF (Alhawarat et al., 2015)

On the other hand, the same experiments were performed on selected parts of the holy Quran. This study achieved two main results: the first one is that the calculated frequency for each term in Quran depends on the partitioning methods used in the analyses. The second result reports that the frequent terms calculated based on TF are more suited to semantic search and clustering applications whereas those calculated based on TF-IDF are very useful in topic modelling.

6.2.2 Classification

Another machine learning approach is the classification algorithms that used in Quranic studies to classify Quran chapters as well as to retrieve similar and related verses. (Akour et al., 2014) exploited Support Vector Machine (SVM) algorithm, specifically LibSVM classifier in Weka, to classify Quran chapters into Makki and Madani chapters. In this experiment, stop words were removed from the whole Quran and the top 1000 4grams words were extracted based on the highest frequency. Next, the SVM matrix was built where the top 1000 4-grams represent the columns and the frequency of the particular 4gram in the particular 114 Quran chapters represent the rows. The three labels of classes were added in the last column as Makki (MK), Madani (MD), or both (MKMD). This information was collected from Islamic websites. Finally, the LibSVM classifier was used to evaluate the

classification accuracy. Figure 6.5 demonstrates the classification result which reaches 89% of correctly classified chapters and 10% of incorrectly classified chapters. Detailed accuracy information based on different measures appears also in this figure.



```
=== Summary ===

Correctly Classified Instances         102               89.4737 %
Incorrectly Classified Instances        12               10.5263 %
Kappa statistic                          0.7349
Mean absolute error                      0.0702
Root mean squared error                  0.2649
Relative absolute error                 23.0997 %
Root relative squared error             68.3772 %
Coverage of cases (0.95 level)          89.4737 %
Mean rel. region size (0.95 level)      33.3333 %
Total Number of Instances              114

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      Class
                 0.167    0.000    1.000      0.167    0.286      0.390    MKMD
                 0.905    0.022    0.905      0.905    0.905      0.883    MD
                 1.000    0.303    0.890      1.000    0.942      0.788    MK
Weighted Avg.    0.895    0.219    0.904      0.895    0.866      0.763

=== Confusion Matrix ===

  a  b  c   <-- classified as
  2  2  8 |  a = MKMD
  0 19  2 |  b = MD
  0  0 81 |  c = MK
```

Figure 6.5 The classification result using LibSVM classifier in Weka (Akour et al., 2014)

6.2.3 Information Retrieval

Due to the large amount of textual data which we need to store properly and access efficiently, looking for automatic information retrieval systems that achieve these goals is an obligatory task. For Quranic Arabic, this process is very critical because of the nature of text which needs more accuracy as well as reliability during the development of such systems.

One work was conducted by (Abdelnasser et al., 2014) to build a question answering system that takes an Arabic question related to Quran from the user as an input and retrieves the related verses to return the suitable answer using Quran and its interpretation books (Tafseer). The system consists of different online and offline modules as clarified in Figure 6.6.

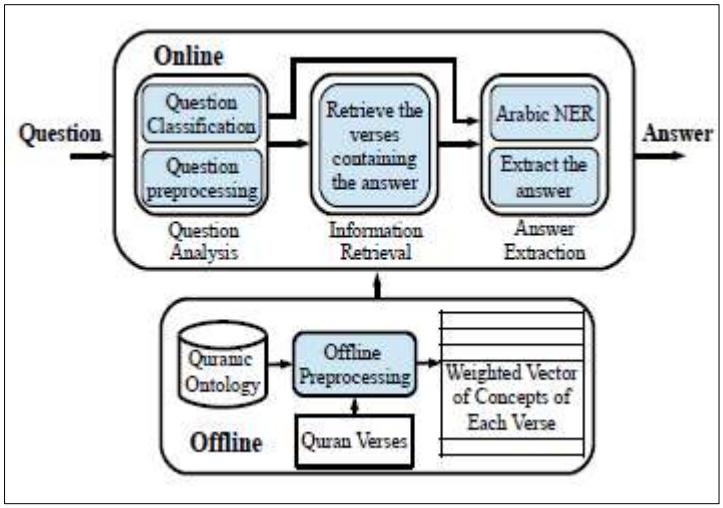
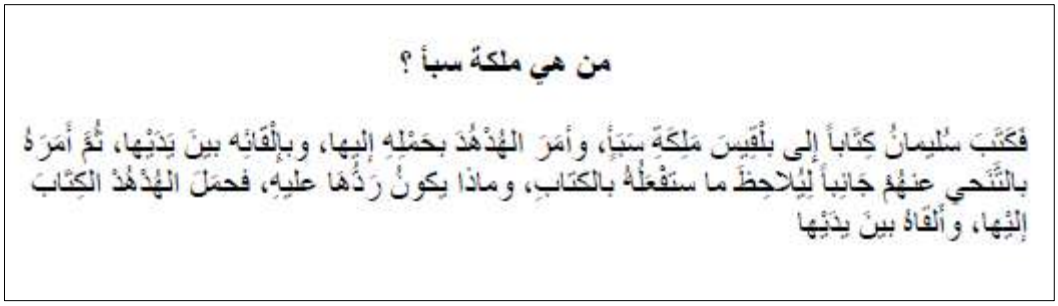


Figure 6.6 The Question Answering System for Quran (Abdelnasser et al., 2014)

The online modules include the question analysis module where the input question is preprocessed to extract the query that will be used in the information retrieval module. Then, the question is classified to get the type of the question and hence the type of the expected answer. The second online module is the information retrieval where the most semantically related verses were retrieved from Quran and Taffseer. Finally, the answer extraction module is responsible for defining the answer as a phrase. This is accomplished by identifying the named entities in the question and extracting the main features that help in ranking each candidate answer.

The offline modules consist of Quranic ontology of concepts which classifies the Quran verses according to their topics. Then, a weighted vector of concepts is generated for each verse in the Quran and the top scoring verses that are semantically related to the user question are passed to the information retrieval module. Figure 6.7 shows a sample of a user query and its output answer.


من هي ملكة سبأ ؟

فكَتَبَ سُليمانُ كِتَاباً إلى بِلْقِيسَ مَلِكَةِ سَبَأٍ، وأمَرَ الهُدْهُدَ بحَمْلِهِ إليها، وبإلْقَائِه بينَ يَدَيْها، ثُمَّ أمَرَهُ بالتَّنَحي عنهُمْ جَانِباً لِيُلاحِظَ ما ستَفْعَلُهُ بالكتَابِ، وماذا يكونُ رَدُّهَا عليهِ، فحمَلَ الهُدْهُدُ الكِتَابَ إليْها، وألْقَاهُ بينَ يدَيْها

Figure 6.7 Sample of the input question and the retrieved answer (Abdelnasser et al., 2014)

In the rest of this chapter, we present our different approach which is an analytical study that aims at mining the Arabic text of the holy Quran (Bentrcia et al., will appear in 2018). To the best of our knowledge, there is no research study that analyzed this sacred text the way it is done in this work. The main contribution is that, we combine statistical and grammatical methods to mine Quran. First, we exploit an efficient Arabic tool, which is AND conjunction, to extract the co-occurred words that are combined by AND conjunction and hence represent conjunctive phrases. Second, we propose a set of patterns that are used to extract the whole set of co-occurred words combined by AND. Moreover, we demonstrate various cases of the words that take different positions/orders in the conjunctive phrase. In particular, we show that different orders of one word yield different meanings and association measures. This study presents a totally novel approach since none of the existing methods illustrated the order concept of co-occurred words or even provided statistics about the different positions/ orders that co-occurred words had taken in Quran. Finally, we measure the value of the association relationship between the two co-occurred words in the conjunctive phrase using Pointwise Mutual Information method (PMI) (Church and Hanks, 1990) and the Sketch Engine tool function (Word Sketch Difference: the Word Sketch Difference help). These basic analyses can be exploited very efficiently to build Quranic ontologies by extracting semantic relations from the holy Quran and assigning precise properties and restrictions to them (Alvarez et al, 2007).

6.3 Analyzing the order of words in AND conjunctive phrases

The holy Quran is the last heavenly books that God revealed to the Prophet Muhammad, peace be upon him. It is divided into 114 chapters called Surah, of different size, and each chapter consists of several verses named Aya, which make a total of 6243 verses, and 77430 words (Dukes and Habash, 2010).

The Quranic text is very challenging to be studied because it is the word of God. Therefore, every word in the Quran counts a great deal and needs a solid knowledge of Arabic in general and the language of the holy Quran in particular. We have tested this fact during the conducting of this work where we found that each word in Quran reserves a specific position in the verse because of important reasons related to the interpretation of that verse (Al-Soyouti, 1973). More accurately, a word may precede an adjacent word because of a special care, the more care you pay for a word in Quran, the more precedence among words it has in the verse. For this reason, we face some words which precede adjacent words in many verses whereas they follow them in others. In the case of conjunctive phrases, mentioned in Chapter 5, Section 5.2.2, we can divide the two combined words based on their position/order in the conjunctive phrase into three main categories:

- Words that occurred in a specific order in the conjunctive phrase and repeated only one time in Quran. It occupies a high percentage of 81.47% of the total number of AND conjunctive phrases.
- Words that occurred in a specific order in the conjunctive phrase and repeated many times in Quran. It occupies a reasonable percentage of 18.62% of the total number of AND conjunctive phrases.

- Words that occurred in two different orders in the conjunctive phrase and repeated one/ many time(s) in the holy Quran. It occupies a small percentage of 3.43% of the total number of AND conjunctive phrases.

The three categories and their percentages are shown in Figure 6.8.

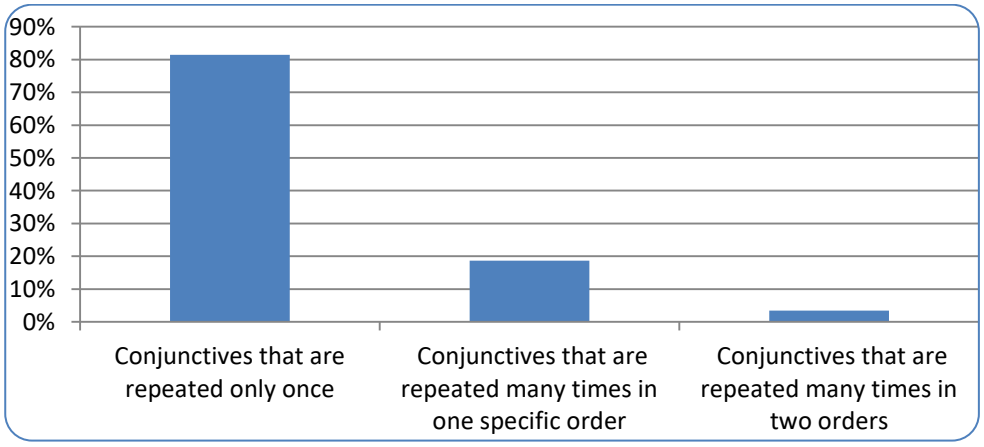


Figure 6.8 The three categories of word orders and their percentages

6.3.1 Words that have occurred in one specific order in the conjunctive phrase and repeated only once in the Quran

This set includes words that are combined together with AND conjunction and occurred together in that order only once in the holy Quran even if they are repeated many times separately. As elements of this set, we can find conjunctive phrases of proper nouns and nouns, as shown in Table 6.1.

Table 6.1 Sample of conjunctive phrases that occurred once in one specific order

TERM2		AND	TERM1		TYPE
Solomon	سُلَيْمَان	وَ	هَارُون	Aaron	Proper Nouns
Uzza	ٱلْعُزَّى	وَ	ٱللَّات	Lat	Proper Nouns
Cucumbers	قِثَّائِ	وَ	بَقْل	Green Herbs	Proper Nouns
Summer	صَّيْف	وَ	شِتَاء	Winter	Proper Nouns

al-Marwah	مَرْوَة	وَ	صَّفَا	as-Safa	Proper Nouns
Sheep	غَنَم	وَ	بَقَر	Cow	Proper Nouns
Morning Star	طَارِق	وَ	سَمَاء	Heaven	Nouns
No disputing	لَا جِدَالَ	وَ	لَا فُسُوقَ	No disobedience	Nouns
Diver	غَوَّاص	وَ	بَنَّاء	Builder	Nouns
Weaning	فِصَال	وَ	حَمْل	Gestation	Nouns
Bowls	جِفَان	وَ	تَّمَاثِيل	Statues	Nouns
Nor sleep	لَا نَوْمٌ	وَ	سِنَةٌ	Drowsiness	Nouns

6.3.2 Words that have occurred in one specific order in the conjunctive phrase and repeated many times in Quran

There are many Arabic and Islamic studies that talk about order in Quranic co-occurred words and explain the reasons that make a word precedes or follows an adjacent word in the verse (Abderrahman, 1987; Al-Samiraii, 2006). In the case of conjunctive phrases, we find a set of words that follow the same order many times in the Quran. This repetition could be considered as a sign for the existence of a relationship between these words. Table 6.2 illustrates some elements of this set.

Islamic scholars indicate many reasons for words precedence. One of them is the word preference. We find this, for example, in the phrase "**الذَّكَرَ وَالأُنْثَى**", "**Male AND Female**", in the verse 45 of An-Najm (The Star) chapter:

وَأَنَّهُ خَلَقَ الزَّوْجَيْنِ الذَّكَرَ وَالأُنْثَى

(*And that He creates the two mates - the male and female* -) [53:45]

The word male always precedes the word female because male exhibits some distinct features that female do not i.e. physical capabilities that make him stronger and more capable of performing some tasks that female cannot.

Another reason is word precedence in the sense of existence such as in the phrase " **إِسْحَاقَ وَيَعْقُوبَ**", "**Isaak AND Jacob**" where the prophet Isaak was born before his brother the prophet Jacob and the prophet Ishmael was born before his brother Isaak, as shown in the verse 84 of Al-An'am (The Cattle) chapter:

وَوَهَبْنَا لَهُ إِسْحَاقَ وَيَعْقُوبَ كُلًّا هَدَيْنَا وَنُوحًا هَدَيْنَا مِن قَبْلُ وَمِن ذُرِّيَّتِهِ دَاوُودَ وَسُلَيْمَانَ وَأَيُّوبَ وَيُوسُفَ وَمُوسَىٰ وَهَارُونَ وَكَذَٰلِكَ نَجْزِي الْمُحْسِنِينَ

(*And We gave to Abraham, Isaac and Jacob - all [of them] We guided. And Noah, We guided before; and among his descendants, David and Solomon and Job and Joseph and Moses and Aaron. Thus do We reward the doers of good*) [6:84]

Table 6.2 Sample of conjunctive phrases that occurred many times in one specific order

The Conjunctive Phrase		Frequency in Quran
'Judgment AND Knowledge'	'حُكْم 'و ' عِلْم '	4
' East AND west'	'مَشْرِق 'و ' مَغْرِب'	6
'Unseen AND the Witnessed'	'غَيْب 'و ' شَهَادَة '	10
'Guidance AND Mercy'	'هُدًى 'و ' رَحْمَة '	13
' Isaac AND Jacob'	'إِسْحَاق 'و ' يَعْقُوب'	10
'World AND Hereafter'	'دُنْيَا 'و ' اخِر '	16
' Male AND Female '	'ذَكَر 'و ' أُنْثَى'	4
' Night AND Day '	' لَيْل 'و' نَهَار'	21
'Protector NOR Helper'	' وَلِي ' و لَا 'نَصِير'	12
' Ishmael AND Isaac' '	إِسْمَاعِيل' و ' إِسْحَاق '	6
'Forgiveness AND Reward'	' مَّغْفِرَة ' و 'أَجْر '	6

Also, it appears clearly in the verse 39 of Ibrahim (Abraham) chapter:

الْحَمْدُ لِلَّهِ الَّذِي وَهَبَ لِي عَلَى الْكِبَرِ إِسْمَاعِيلَ وَإِسْحَاقَ إِنَّ رَبِّي لَسَمِيعُ الدُّعَاءِ

(*Praise to Allah, who has granted to me in old age Ishmael and Isaac. Indeed, my Lord is the Hearer of supplication*) [14:39]

A different reason is word precedence in the sense of time such as in the phrase " **الْمَشْرِقُ وَالْمَغْرِبُ**", "**East AND West**", where the day starts by the sunrise from east to west, as mentioned bellow in the verse 115 of Al-Baqarah (The Cow) chapter:

وَلِلَّهِ الْمَشْرِقُ وَالْمَغْرِبُ فَأَيْنَمَا تُوَلُّواْ فَثَمَّ وَجْهُ اللّهِ إِنَّ اللّهَ وَاسِعٌ عَلِيمٌ

(*And to Allah belongs the east and the west. So wherever you [might] turn, there is the Face of Allah. Indeed, Allah is all-Encompassing and knowing*) [2:115]

In addition, we find word precedence according to the development situation such as " **السَّمْعَ وَالْبْصَرَ**", "**Hearing AND Vision**" in the fetus where the evolution of hearing is completed before the evolution of vision which is delayed after the birth of the fetus. The verse 78 of An-Nahl (The Bees) chapter states this clearly:

وَاللَّهُ أَخْرَجَكُمْ مِنْ بُطُونِ أُمَّهَاتِكُمْ لَا تَعْلَمُونَ شَيْئًا وَجَعَلَ لَكُمُ السَّمْعَ وَالْأَبْصَارَ وَالْأَفْئِدَةَ لَعَلَّكُمْ تَشْكُرُونَ

(*And Allah has extracted you from the wombs of your mothers not knowing a thing, and He made for you hearing and vision and intellect that perhaps you would be grateful*) [16:78]

6.3.3 Words that have occurred in two different orders in the conjunctive phrase and repeated one/many time(s) in the holy Quran

One main application of word co-occurrences is to extract semantic relations that may exist between them (Bullinaria and Levy, 2007). In Arabic grammar, the association relation between two words in AND conjunctive phrase $word_1$ and $word_2$ is the same as the relation between $word_2$ and $word_1$, which is not the case in Quranic conjunctive phrases. Our contribution in this study is to reveal and discuss the differences between the two types of

association that may exist between $word_1$ and $word_2$, and $word_2$ and $word_1$ in the Quranic conjunctive phrases from the contextual meaning side and the association magnitude side.

There are no extra or meaningless words in the Quran; on the contrary, there exist words which have more than one meaning based on their positions in the verse. Moreover, the order which a word follows in a verse may also influence its interpretation. In the case of conjunctive phrases, we find a set of words that follow two different orders one/ many time(s) in the Quran such as the examples of Table 6.3.

Whether a specific word precedes or follows its adjacent word is based on the context of the verse where they occur (Al- Masiri, 2005). For example, in the phrase "**الْأَرْضَ وَالسَّمَاوَاتِ**", "**Earth AND Heavens**", the word 'Earth' precedes the word 'Heavens' because earth is created before heavens, as illustrated in the verse 4 of Taha (Ta-Ha) chapter:

تَنزِيلًا مِّمَّنْ خَلَقَ الْأَرْضَ وَالسَّمَاوَاتِ الْعُلَى

(*A revelation from He who created the earth and highest heavens*) [20:4]

However, in more than 100 verses, we find the word 'heavens' comes before 'earth' because of its huge space and great creation. An example is the verse 77 of An-Nahl (The Bees) chapter:

وَلِلَّهِ غَيْبُ السَّمَاوَاتِ وَالْأَرْضِ وَمَا أَمْرُ السَّاعَةِ إِلَّا كَلَمْحِ الْبَصَرِ أَوْ هُوَ أَقْرَبُ إِنَّ اللَّهَ عَلَىٰ كُلِّ شَيْءٍ قَدِيرٌ

(*And to Allah belongs the unseen [aspects] of the heavens and the earth. And the command for the Hour is not but as a glance of the eye or even nearer. Indeed, Allah is over all things competent*) [16:77]

Another example of words which have occurred in two different orders is the phrase " **الْجِنَّ وَالْإِنسَ**" "**Jinn AND Mankind**" in the verse 56 of Adh-Dhariyat (The Winnowing Winds) chapter, we found that the word 'Jinn' precedes 'Mankind' because Jinn are created before Mankind.

وَمَا خَلَقْتُ الْجِنَّ وَالْإِنسَ إِلَّا لِيَعْبُدُونِ

(*And I did not create the jinn and mankind except to worship Me*) [51:56]

Moreover, in the verses where there is a kind of challenging in movement and speed, we also find Jinn before Men because of their supernatural ability, as presented in the verse 33 of Ar-Rahman (The Beneficent) chapter:

يَا مَعْشَرَ الْجِنِّ وَالْإِنسِ إِنِ اسْتَطَعْتُمْ أَن تَنفُذُوا مِنْ أَقْطَارِ السَّمَاوَاتِ وَالْأَرْضِ فَانفُذُوا لَا تَنفُذُونَ إِلَّا بِسُلْطَانٍ

(*O company of jinn and mankind, if you are able to pass beyond the regions of the heavens and the earth, then pass. You will not pass except by authority [from Allah*]) [55:33]

However, in some verses, such as the verse 88 of Al-Israa (The Night Journey) chapter, God asked Men before Jinn to create Quran because it is a challenge for them first and foremost:

قُل لَّئِنِ اجْتَمَعَتِ الْإِنسُ وَالْجِنُّ عَلَىٰ أَن يَأْتُوا بِمِثْلِ هَٰذَا الْقُرْآنِ لَا يَأْتُونَ بِمِثْلِهِ وَلَوْ كَانَ بَعْضُهُمْ لِبَعْضٍ ظَهِيرًا

(*Say, "If mankind and the jinn gathered in order to produce the like of this Qur'an, they could not produce the like of it, even if they were to each other assistants.*") [17:88]

On the other side, in order to find the difference in the association values between the two words in the conjunctive phrase, we apply Pointwise Mutual Information method (PMI) to measure how much information one word can give about the other one which occurs with it. This method is derived from information theory and widely proposed to find semantic relations between either adjacent words that occur together frequently or trigger pairs, which are long distance word pairs.

Table 6.3 Sample of conjunctive phrases that occurred one/ many time(s) in Quran in two orders

The Conjunctive Phrase		Frequency	(PMI) Method	Word Sketch Difference Function
'Heavens AND Earth'	'سَمَاء ' و'أَرْض '	2	1.1827	5.4
' Earth AND Heavens '	'أَرْض ' و 'سَمَاء '	148	5.0673	10.2
' Thamud AND ‘Ad '	'ثَمُود ' و 'عَاد '	5	7.8071	9.0
' ‘Ad AND Thamud'	'عَاد ' و 'ثَمُود '	1	5.9997	6.7
' Warner AND 'Bearer of glad tidings '	'نَذِير ' و 'بَشِير '	5	8.1641	10.0
'Bearer of glad tidings AND Warner'	'بَشِير ' و 'نَذِير '	2	7.1641	9.8
'Jinn AND Mankind'	'جِنّ ' و 'إِنس '	3	7.5626	9.1
'Mankind AND Jinn'	'إِنس ' و 'جِنّ '	9	9.2996	10.8
'Harm NOR Benefit'	'ضَر ' ولَا 'نَفْع '	3	9.4106	9.9
'Benefit NOR Harm'	' نَفْع' ولَا ' ضَر '	4	10.1476	10.4

$$PMI(x,y) = \log \frac{\mathrm{p(x,y)}}{\mathrm{p(x)p(y)}} \quad (6.1)$$

where *p(x, y)* is the probability that the two words *x* and *y* occur together in the same verse, *p(x)* is the probability that word x occurs alone in that verse, and the same for *p(y).*

From Table 6.3, we can notice the difference in the association values between the two combined words with a different order in the conjunctive phrase. The phrase “بَشِير و نَذِير” “Bearer of glad tidings AND Warner” has an association value of 7.1641 whereas the phrase “نَذِير و بَشِير” “Warner AND Bearer of glad tidings” has 8.1641. High PMI value indicates a high degree of association relationship between the words and vice versa. Moreover, high frequent pairs of words have high association values compared to those with low frequency.

In addition, to validate the first approach we use another method which is the word sketch difference function available in the Sketch Engine tool (Kilgarriff et al., 2014). This function is used to compare any two words in their lemma form by displaying those patterns and combinations that the two words have shared in common or differentiated by. Besides that, there are four numbers next to each pattern; the first two show the frequency of co-occurrence with the first and the second word, whereas the last two show the salience scores for the pattern with both words (the Word Sketch Difference help). As an example, we compare the two phrases "ثَمُود و عَاد" "Thamud AND 'Ad" and "عَاد و ثَمُود" "'Ad AND Thamud" using Word Sketch Differences as shown in Figure 6.9 and Figure 6.11.

Figure 6.10 illustrates the impact of the word order in the conjunctive phrase on the association score. The first column in the figure lists the whole words in the corpus that combined with 'عاد' or 'ثمود', or both. The second column lists the frequency of the occurrence of each word in the first column with the word 'ثمود' whereas the third column lists the frequency of the occurrence of each word in the first column with the word 'عاد'. The fourth column reflects the association score of the relation between the words listed in the first column and the word 'ثمود' while the fifth column reflects the association score of the relation between the words listed in the first column and the word 'عاد'. It is clear that the word 'ثمود' in the first column comes to the right of the word 'عاد' in this form only once in the holy Quran with an association value of 6.7. However, when the same word 'ثمود' comes to the left of the same word 'عاد', the association value increases to 9.0 with a frequency of 5, as depicted in Figure 6.12.

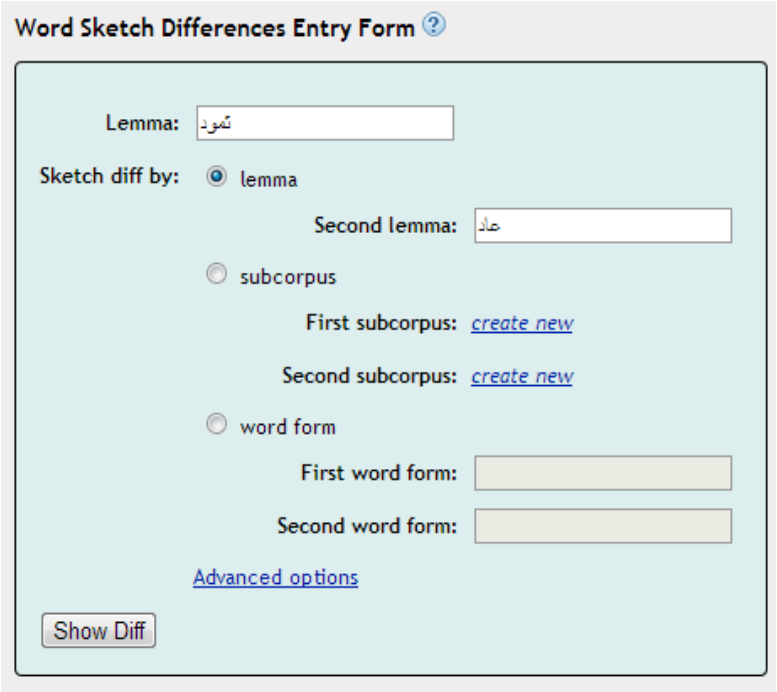


Figure 6.9 Word Sketch differences entry form for the phrase “Thamud AND ‘Ad”

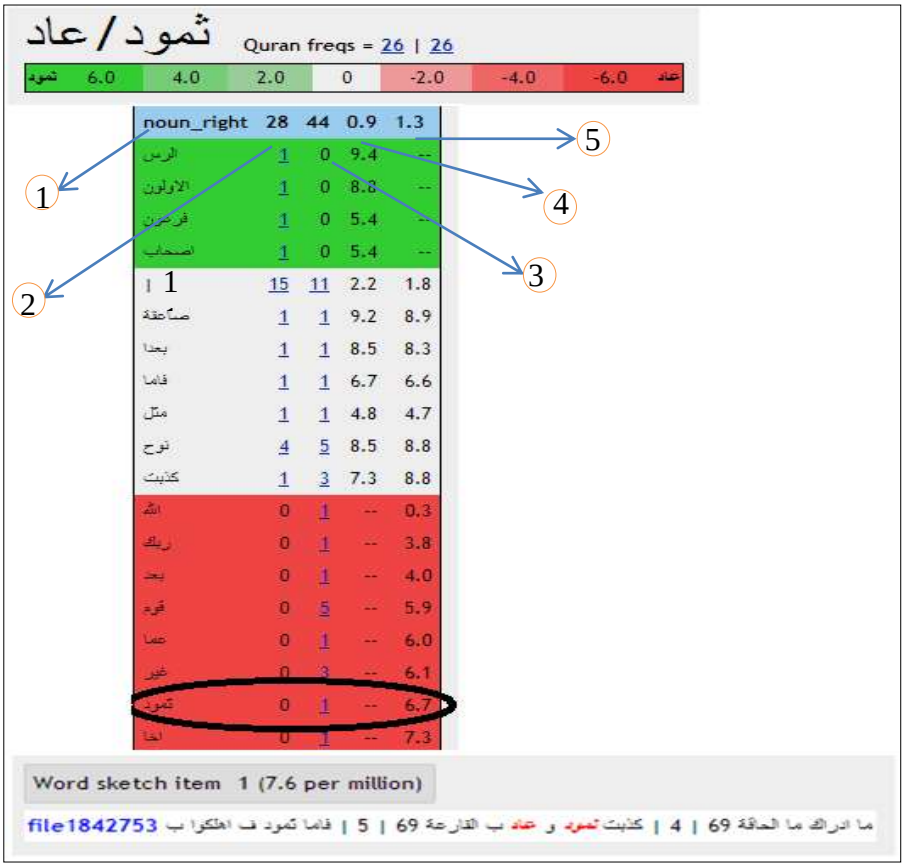


Figure 6.10 The association score and frequency of the phrase “Thamud AND ‘Ad”

Figure 6.11 Word Sketch differences entry form for the phrase "'Ad AND Thamud"

Figure 6.12 The association score and frequency of the phrase "'Ad AND Thamud"

6.4 Summary

In this work, we have performed an analytical study on the Arabic conjunctive phrases, namely AND conjunction, extracted from the Quranic Arabic corpus. This research is very useful for religious scholars, scientist, and linguists because it shows the linguistic miracle of the holy Quran based on scientific evidences. Efficiently, we have analyzed the order of the

two words that form the conjunctive phrase and its effect on the contextual meaning of the Quranic verse where they have occurred and the association relationship between them. We have reported three different cases: words that have occurred in a specific order in the conjunctive phrase and repeated only once in the Quran, words that have occurred in a specific order in the conjunctive phrase and repeated many times in the Quran, and words that have occurred in two different orders in the conjunctive phrase and repeated one/many time(s) in the holy Quran. In the future, we plan to explore a wider range of Quranic co-occurred words rather than the AND conjunctive phrases and test different association measurements.

7 Measuring Similarity between Quran Chapters

7.1 Introduction

A similarity measure is a function which computes the degree of similarity between a pair of documents. This computational task is very complex and is fundamental in Information retrieval and natural language processing applications such as text automatic clustering (Willett, 1988) and search engines (Strehl et al., 2000). The immense need for this task increases with the huge collections of documents in digital libraries and repositories that should be categorized efficiently, and over web networks where users search for relevant documents related to a specific query. However, for Arabic documents, this remains a challenging issue due to the morphological and complicated structure of Arabic language and the shortage in resources and tools that support Arabic (Habash, 2010).

In this study, we retrieve similar words/phrases from a sample of Arabic documents, represented by the chapters of the holy Quran, using lexical matching method. The rationale behind selecting Quran is that it is the most sophisticated Arabic books with linguistic and religious values, and a comprehensive resource of Arabic vocabulary. Moreover, we explore three different similarity metrics, which are cosine, Jaccard, and correlation distances, to find out the degree of similarity between any two Quranic chapters through measuring the distance between them. This almost ranges between zero and one; a distance value of one indicates that the two chapters are totally different whereas the distance value of zero indicates that the chapters are identical. Furthermore, we generate and validate manually a precise and comprehensive set of stop words from Arabic vowelized Quran. Removing these stop words from the chapters increases the efficiency of information systems that deal with Quran. Compared to other few studies on Quran, this work is the first one that retrieves similar words and phrases from two Quranic chapters written in Arabic script besides measuring the

similarity value using different metrics. These contributions would be widely exploited in many linguistic and religious studies.

The rest of the chapter is organized as follows: in Section 7.2, the methodology is described including the data set and the proposed approaches. Finally, Section 7.3 presents conclusion and future work.

7.2 Measuring Similarities

7.2.1 Data Set

We conduct our experiment using a collection of documents represented by Quran chapters written in Arabic script. The holy Quran is divided into 114 chapters (Surah) with 6236 verses (Ayah) of different lengths, and 77430 words (Dukes and Habash, 2010). Each chapter has a specific name (title) illustrates the main topic that the chapter is talking about.

The concept of similarity between Quran chapters or even between phrases (verses) within a chapter is very popular and shows the importance of semantics that the phrase holds (Al-Abbad, 2002). For instance, in Ash-Shu`ara' chapter (The Poets) where the following phrase was repeated eight times within the chapter:

" إِنَّ فِي ذَٰلِكَ لَآيَةً وَمَا كَانَ أَكْثَرُهُمْ مُؤْمِنِينَ وَإِنَّ رَبَّكَ لَهُوَ الْعَزِيزُ الرَّحِيمُ"

"Surely, in this there is a sign yet most of them do not believe. Your Lord, He is the Almighty, the Most Merciful."

On the other hand, there are so many examples of phrases shared by different chapters such as the following:

" وَيَقُولُونَ مَتَى هَٰذَا الْوَعْدُ إِن كُنتُمْ صَادِقِينَ "

"They ask: 'If what you say is true, when will this promise come? "

We can find this phrase in Yunus (Jonah), Al-Anbiya' (The Prophets), An-Naml (The Ant, The Ants), Saba' (Sheba), Ya seen (Ya Seen), and Al-Mulk (The Dominion, Sovereignty, Control).

The next sections demonstrate how to extract similar phrases from any two Quran chapters, and measure the similarity value using three different metrics.

7.2.2 Lexical-based Similarity

We develop lexical-based similarity method, which depends on simple matching algorithm, to extract similar words and phrases from Quran chapters. This process starts by comparing words from the first chapter with words from the second chapter looking for similar words and their positions in the two chapters. This process is terminated by building a dictionary of keywords (similar words), shared between the two selected chapters, and their frequencies. Next, for each keyword in the dictionary, we initiate a searching process that looks in both chapters for a match between words that follow each of the extracted keywords. If more than one word is returned by this algorithm, then there are similar phrases between the two chapters. Otherwise, keywords are returned. The pseudo code for the lexical-based similarity algorithm is depicted bellow.

Algorithm: Similar words and phrases between two Quran chapters
Input: Two Quran chapters written in Arabic
Output: List of similar words and pharses shared between the two chapters

Steps:

1. **for** each word i in chapter i **do**
   for each word j in chapter j **do**
   if word i is equal to word j
   save word i in simWord
   save the index of the occurrence of word i in chapter i in index 1 and the index of the occurrence of word j in chapter j in index 2
   end if
   end for
   end for
2. **for** i=1: no of simWord **do**
   while size of index 2 not reached **do**
   compare simWord i in chapter 1(index 1_i) and chapter 2 (index $2_{i,j}$)
   while they are equal **do**
   save simWord i in simPhrase
   move to compare the following words of simWord i in the both chapters
   end while
   move to the next occurrence j=j+1
   end while
   end for
3. return simWord and simPhrase

End algorithm

Although simple matching algorithm is time consuming especially for large documents but for Quran chapters it is very suitable since the longest one, which is Al-Bakarah (The Cow), does not exceed 6144 words, and hence a reasonable number of comparison operations between words.

We present some examples of large, medium, and short documents that share similar words and phrases, as depicted in Table 7.1.

Table 7.1 Sample of similar phrases shared between two Quran chapters

Finding similarity between two documents	**Similar phrases Samples**	**Frequency**
Al-Bakarah and Al Imran The Cow and The family of Imran	1-'بِسْمِ"اللَّهِ"الرَّحْمَنِ"الرَّحِيمِ"الم' In the Name of Allah, the Merciful, the Most Merciful AlifLaamMeem	2
	2-'لَا"رَيْبَ"فِيهِ' There is no doubt	3
	3-'إِنَّ"اللَّهَ"عَلَى"كُلِّ"شَيْءٍ"قَدِيرٌ' Allah has power over all things	4
	4-'وَمَا"اللَّهُ"بِغَافِلٍ"عَمَّا"تَعْمَلُونَ' Allah is not inattentive of what you do	5
	5-'مِلَّةَ"إِبْرَاهِيمَ"حَنِيفًا"وَمَا"كَانَ"مِنَ"الْمُشْرِكِينَ' The Creed of Abraham, the upright one. He was not among the idolaters.	2
	6-'إِبْرَاهِيمَ"وَإِسْمَاعِيلَ"وَإِسْحَاقَ"وَيَعْقُوبَ"وَالْأَسْبَاطِ"وَمَا"أُوتِيَ"مُوسَى"وَعِيسَى' Abraham, Ishmael, Isaac, Jacob, and the tribes; to Moses and Jesus	2
	7-'اللَّهِ"وَالْمَلَائِكَةِ"وَالنَّاسِ"أَجْمَعِينَ"خَالِدِينَ"فِيهَا"لَا"يُخَفَّفُ"عَنْهُمُ"الْعَذَابُ"وَلَا"هُمْ"يُنْظَرُونَ' Allah, the angels, and all people. They are there (in the Fire) for eternity neither shall the punishment be lightened for them; nor shall they be given respite.	2
	8-'إِنَّ"فِي"ذَلِكَ"لَآيَةً"لَكُمْ"إِنْ"كُنْتُمْ"مُؤْمِنِينَ' That will be a sign for you if you are believers.	2
	9-'وَثَبِّتْ"أَقْدَامَنَا"وَانْصُرْنَا"عَلَى"الْقَوْمِ"الْكَافِرِينَ' Make us firm of foot and give us victory against the nation of unbelievers.	2
	10-'اللَّهُ"لَا"إِلَهَ"إِلَّا"هُوَ"الْحَيُّ"الْقَيُّومُ' Allah, there is no god except He, the Living, the Everlasting	2
Al-Muzzammil and Al-Insan The Enfolded One and The Human	1-'وَاذْكُرِ"اسْمَ"رَبِّكَ' Remember the Name of your Lord	2
	2-'إِنَّ"هَذِهِ"تَذْكِرَةٌ"فَمَنْ"شَاءَ"اتَّخَذَ"إِلَى"رَبِّهِ"سَبِيلًا' This is indeed a Reminder. Let whosoever will take the Path to his Lord.	2
	3-'إِنَّ"اللَّهَ' Allah is	2
Al-Falaq and Al-Nas The Daybreak and Mankind	1-'بِسْمِ"اللَّهِ"الرَّحْمَنِ"الرَّحِيمِ"قُلْ"أَعُوذُ"بِرَبِّ' In the Name of Allah, the Merciful, the Most Merciful Say: 'I take refuge with the Lord of	2
	2-'مِنْ"شَرِّ' From the evil	5

In this work, vowelization is an important component in Arabic Quranic script which we keep to preserve the meaning of phrases and to distinguish between words in general and identical words with different vowelization in particular. Table 7.2 shows some samples of chapters that share similar words, and their frequencies.

Table 7.2 Sample of similar words shared between two Quran chapters

Finding similarity between two documents	**Similar words Samples**		**Frequency**
An-Nisa' and Al-Ma'idah The Women and The Food	'اللَّهِ'	Allah	88
	'شَيْءٍ'	Thing	16
	'الَّذِينَ'	Those	100
	'عَلِيمٌ'	The Knower	6
	'الْمَوْتُ'	The Death	5
	'آمَنُوا'	Believe	45

	'النَّاسَ' 'النَّاسِ' 'الشَّيْطَانُ' 'كَفَرُوا' 'الصَّلَاةَ' 'مَرْيَمَ' 'السَّمَاوَاتِ'	People People Satan Disbelieve Prayer Mary Heavens	9 9 4 21 9 14 11
Ash-Shura and Al-Ahqaf The Consultation and The Dunes	'حم' 'مِنْ' 'الْأَرْضِ' 'الْجَنَّةِ' 'عَلَى' 'الدُّنْيَا' 'الظَّالِمِينَ' 'مُسْتَقِيمٍ' 'الْقِيَامَةِ' 'الْحَقَّ'	HaMeem Than Earth Paradise On Earthly life Harm doers Straight Resurrection Truth	2 51 9 3 13 3 6 2 2 3
Ad-Dhariyat and At-Tur The Wind That Scatter and The Mount	'اللَّهِ' 'اللَّيْلِ' 'السَّمَاءِ' 'قَوْمٌ' 'طَاغُونَ'	Allah Night Sky People Insolent	6 2 3 3 2

7.2.3 Statistical-based Similarity

This part of work starts by pre-processing the text of each Quran chapter. This includes three major steps:

7.2.3.1 Removing Stop Words

Arabic language is very rich due to its vocabulary and grammar. Hence, a large number of common and frequent words exist in Arabic texts and have no significant semantic relation to the context in which they occur and cannot be used alone to index and distinguish documents. Little studies were conducted to generate stop words for Arabic language (El-Khair, 2006; Alhadidi and Alwedyan, 2008). However, the resulting lists of stop words are neither comprehensive nor precise enough because they do not consider vowelization in their approaches. In this work, we add a very important contribution by generating and validating manually a set of stop words for Arabic, vowelized Quran text. It includes prepositions, conjunctions, nouns, and articles. Verbs with all their variations are not included since Quran is a sacred text and verbs may hold important semantics which we should not remove. Such

examples of stop words that have the same letters but different vowelization and then different meanings:

لِمَا لَمَّا لَمَا -أَلاَّ أَلَا -أَنَّ أَنْ - بِكُمُ بِكُمْ - أَفَمَنْ أَفَمِنْ

Moreover, each stop word is vowelized according to its position in the sentence such as these examples:

'بَعْضُهُمْ' 'بَعْضَهُ' 'بَعْضَهَا' 'بَعْضُهَا' 'بَعْضَكُمْ' 'بَعْضِكُمْ' 'بَعْضُكُمْ' 'بَعْضَ' 'بَعْضٍ' 'بَعْضُ' 'بَعْضِ' 'بَعْضًا'

'بَعْضُنَا' 'بَعْضَهُمْ'

A sample of the extracted stop words from Arabic vowelized Quran is demonstrated in Figure 7.1.

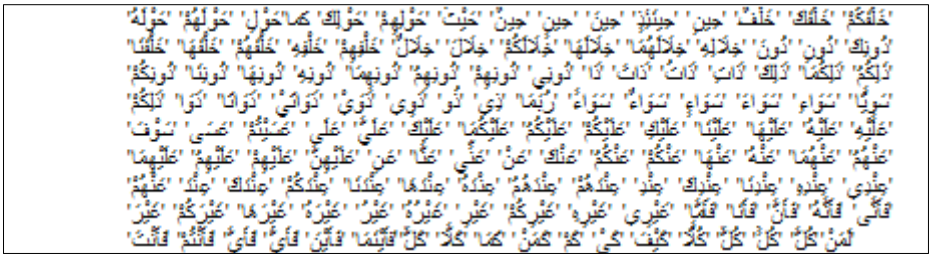


Figure 7.1 Sample of vowelized stop words that exist in Quran

7.2.3.2 Computing Vector Space Model

The purpose of this step is to convert textual data to numeric values. After removing stop words from Quran chapter, we use vector space model to transform each chapter into a vector where its elements are the numeric values of words that form the chapter. Each word has a specific value called weight which shows its importance in the chapter. There is an efficient statistical method to calculate weights called Term Frequency Inverse Document Frequency (tf.idf) (Salton and McGill, 1983):

$$\mathbf{w}_{ij} = \mathbf{tf}_{ij} \cdot \mathbf{log}\left(\frac{\mathbf{N}}{\mathbf{df}_i}\right) \qquad (7.1)$$

where $w_{i,j}$ is the weight of word i in document j, $tf_{i,j}$ is the number of occurrences of word i in document j, N is the total number of documents in the corpus (Quran chapters), and df_i is the number of documents containing word i. A high weight in tf.idf is reached by a high word

frequency in a given document and a low document frequency of the word in the corpus; the weights hence tend to filter out common words which are less discriminative.

This tf.idf formula is very useful to represent Quran chapters; each word in the chapter is converted to an equivalent positive number (weight) using words' frequency and importance in the chapter: $d_k = (w_{1,k}, w_{2,k}, \ldots, w_{t,k})$.

However, absent words are assigned a value of zero. This computation yields to a vector of numbers which can be employed in many different applications that rely on vectors comparison such as information retrieval and search engines. In this work, we computed 114 vectors of around 17000 elements since the number of Quran chapters is 114. Then, we used the resulting vectors to compute the similarity between two Quran chapters by comparing and measuring the distance between them. This process is explained in details in the next section.

7.2.3.3 Measuring Similarity Distances

Measuring similarity reflects the degree of closeness or separation (distance) of the target documents and generally refers to measuring the lexical similarity between two documents, which can be exploited very efficiently to estimate the semantic similarity. Selecting a good metric to represent the value of similarity is a crucial issue for many information retrieval systems that should be robust, fast, and precise. In order to find similarity distance between Quran chapters, we test three different metrics: cosine distance, Jaccard distance, and correlation distance (Siegmund, 1998). These metrics take as inputs the resulting vectors which were computed in Section 7.2.3.2, and provide a distance value of zero if the two chapters are identical, or a value between 0 and 1 otherwise.

7.2.3.3.1 Cosine distance

Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. Vectors with the same orientation have a value of one regardless their magnitude. Also, cosine similarity is particularly used in positive space and ranges between [0, 1]. Identical chapters have a value of 1 and less than one otherwise. A and B are two vectors that represent Quran chapters A and B, respectively. We can compute cosine similarity based on formula (7.2):

$$Cosine\ similarity\ =\ \cos\theta = \frac{A.B}{\|A\|\|B\|} \tag{7.2}$$

$$Cosine\ distance\ =\ 1 - cosine\ similarity$$

7.2.3.3.2 Jaccard distance

Also called Jaccard coefficient and we use it to measure the similarity between Quran chapters by counting the number of shared words between the two chapters and dividing on the number of words that are present in either of them. Jaccard Similarity value ranges between [0, 1]. A and B are two vectors that represent Quran chapters A and B, respectively. We can compute Jaccard similarity based on formula (7.3):

$$\text{Jaccard similarity} = \frac{|A \cap B|}{|A \cup B|} \tag{7.3}$$

$$Jaccard\ distance = 1 - Jaccard\ similarity$$

7.2.3.3.3 Correlation distance

Also it is known as Pearson's correlation coefficient. It returns a similarity value bounded between [-1, 1], computed using formula (7.4), where A and B are two vectors that represent Quran chapters A and B, respectively, *cov* is their covariance, and σ is the standard deviation. We can compute correlation similarity as follows:

$$Correlation\ similarity = \frac{cov(A,B)}{\sigma A\ \sigma B} \quad (7.4)$$

$$correlation\ distance = 1 - correlation\ similarity$$

To conduct our experiment, we select two Quran chapters that we want to find how similar they are. Then, we apply the above three metrics to their corresponding vectors. Table 7.3 demonstrates three examples of Quran chapters that are of large size, medium size, and short size.

Table 7.3 Distance between Quran chapters using three similarity metrics

Quran chapter size	**Quran chapter name**	**Cosine distance**	**Jaccard distance**	**Correlation distance**
Large	Al-Anfal and At-Tawbah (The Spoils of War and The Repentance)	0.8125	0.9710	0.8426
Medium	An-Naml and Al-Ankabut (The Ant and The Spider)	0.8811	0.9538	0.9054
Short	At-Takweer and Al-Infitar (The Folding Up and The Cleaving Asunder)	0.9413	0.9391	0.9442

We should observe that, for each selected pair of chapters, the three similarity metrics have comparable values and the resulting distances between Quran chapters are very large due to the diverse topics discussed in each chapter; in some chapters, a whole topic is discussed only in very few verses or even in only one verse.

7.3 Summary

Measuring similarity between two documents improves the performance of information retrieval systems, especially if they deal with documents written in challenging languages such as Arabic. In this work, we propose an approach to extract and measure similarity between Arabic Quran chapters. First, we use a lexical based approach represented by a simple matching algorithm to retrieve similar words and phrases from the chapters.

Furthermore, we test three different metrics to measure the similarity: cosine, Jaccard, and correlation distances which yield to comparable values. Besides that, we generate and validate manually a set of stop words from Arabic vowelized Quran. Taking vowelization into account increases the number of stop words but gains the efficiency of target systems. The proposed approaches achieved an excellent success rate. In the future work, we plan to extract and measure similarity between Quran verses and test more similarity metrics.

8 Conclusion and Future Work

8.1 Conclusions

Because Quran is the holy book of Muslims and is the main source for understanding their religion, there is an immense need to information systems that rely on Arabic Quranic text to present a precise and comprehensive knowledge about Quran to the world. However, developing such systems is still a challenging task due to the nature of Arabic writing, the semantic ambiguity of words, the shortage in resources and tools that support Arabic, and the religious nature of Quran text which needs a careful mining.

In this work, we proposed three different approaches that deal with Quranic Arabic text and rely on statistics and Arabic grammar in order to extract knowledge from Quran. In the first two approaches, we have used Quranic Arabic Corpus (QAC), which we have converted to Arabic script, to extract the whole set of AND conjunctive phrases that include nouns, proper nouns, and adjectives. In the first approach, we have extracted three types of semantic relations that exist in AND conjunctive phrases. Moreover, we have validated and measured their strength using statistical techniques. In the second approach, we have conducted an analytical study that reveals the rationale behind the different orders of words in the conjunctive phrase. In the third approach, we have used the Arabic text of the holy Quran to find and measure similarities between any two chapters. All these contributions are very useful to build ontologies for Quran and improve other scientific, religious, and linguistic studies.

An overview about extracting knowledge from Quran is presented in Chapter 1. We have discussed the major issues that make developing information systems for Arabic Quran very

challenging. The main contribution of this thesis has been the development of three novel approaches that seek analyzing Quran to produce Quran ontology, text mining study, and computational text similarity analysis.

Chapter 2 introduces several topics related to the holy Quran for example, its revelation, the classification of its chapters based on revelation location, and the different themes that Quran talks about. Quranic Arabic Corpus as well is introduced and a detailed description of its contents has been provided.

Chapter 3 reports in details the main previous approaches and techniques developed in the field of Arabic text mining in general and the Quranic text in particular.

In Chapter 4, we have highlighted various theoretical topics related to ontologies. We have presented the layer cake of any ontology which includes the ontological elements besides the common exploited approaches in their development. Also, a description of the main ontology tools and languages has been reported.

A novel approach that aims at enriching the automatic construction of Quran ontology is detailed in Chapter 5. We exploited Arabic grammar to capture semantic relations that exist in AND conjunctive phrases. Furthermore, we utilized statistical techniques namely correlation coefficient, Student t-test, and testing hypothesis to validate and measure the strength of the extracted relations. This aids domain experts to estimate and validate their final decisions very efficiently. Finally, we categorized manually those relations into Antonyms, Gender, and Class.

Chapter 6 discussed the concept of order between words that combined by AND conjunction. In particular, we conducted an analytical study about words that take different positions/orders in the conjunctive phrase. We demonstrated that different orders of one word yield different meanings and association measures. Finally, we measure the value of the association relationship between the two words in the phrase using Pointwise Mutual Information method (PMI) and the Sketch Engine tool function (Word Sketch Difference).

One text mining application was described in Chapter 7, where we tried to find similarities between any two chapters of Arabic Quran. We extracted similar words, phrases, verses, and their frequencies from the two chapters looking for estimating their semantic similarity. Moreover, we explored three different similarity metrics, which are cosine, Jaccard, and correlation distances, to find out the degree of similarity between any two Quranic chapters. Finally, we generated and validated manually a precise and comprehensive set of stop words from Arabic vowelized Quran.

Chapter 8 concludes the proposed work in this thesis and includes a summary of the whole chapters along with the main contribution points added to the field of Quranic Arabic mining. The future directions towards improving the contents of this work are also addressed.

8.2 Discussion and Future Directions

The field of text mining is very huge. Any interested researcher can add new features or improve previous achieved approaches especially for Arabic text, which still an immature field. In future, we are intending to reach new scopes by trying the following:

1. Develop an efficient classifier to categorize the extracted semantic relations automatically.

2. Arabic language is very rich because of its vocabulary and grammar that it includes. One of our novel approaches extracted semantic relations that AND conjunction holds and we would like to test other different conjunctions such as OR and BUT.
3. Explore a wider range of Quranic co-occurred words rather than the AND conjunctive phrases and test different association measurements.
4. Extract and measure similarity between Quran verses and test more similarity metrics.
5. Integrate the developed approaches into applications for improving information retrieval on the Web.

9 References

- Abbas, N. (2009). Qurany: *A Tool to Search for Concepts in the Quran*, MSc Research Thesis. School of Computing. University of Leeds, Leeds, UK.
- Abdelnasser, H., Mohamed, R., Ragab, M., Mohamed, A., Farouk, B., El-Makky, N., and Torki, M. (2014). Al-Bayan: an Arabic question answering system for the Holy Quran. Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), pp. 57-64, Doha, Qatar.
- Abderrahman, A. (1987). *Al-Iajaz Albayani li Al-Quran* (In Arabic), Cairo, Egypt, Dar Al-Maaref.
- Abulaish, M., and Dey, L. (2007). Biological Relation Extraction and Query Answering from MEDLINE Abstracts Using Ontology-Based Text Mining. Data & Knowledge Engineering, vol. 61, no. 2, pp. 228-262.
- Adhima, M. (1972). *Derasat li Osloob Al-Quraan Al-Karim*. In Arabic. Cairo, Egypt: Dar Al-Hadith.
- Akour, M., Alsmadi, I., and Alazzam, I. (2014). MQVC: Measuring Quranic Verses Similarity and Sura Classification Using N-Gram. WSEAS Transactions on Computers, vol. 13, pp. 485-491.
- Al-Abbad, A. (2002). *Ayat Mutashabihat Al-Alfadh fi Al-Quran Al-Karim* (In Arabic). Riyadh, Saudi Arabia: Dar Al-Fadhila.
- Al-Arfaj, A. and Al-Salman, A. (2015). Ontology Construction from Text: Challenges and Trends. International Journal of Artificial Intelligence and Expert Systems (IJAE), vol. 6, no. 2, pp. 15-26.
- Al-Dargazelli, S. (2004). Statistical Studies of Holy Quran. [Online]: http//www.quranicstudies.com/printout104.html
- Al-Ghalayini, M. (2007). *Jamea Al-Dorous Al-Arabiya*. In Arabic. Cairo, Egypt: Dar Al-ghad Al-Jadid.
- Alhadidi, B. and Alwedyan, M. (2008). Hybrid Stop-Word Removal Technique for Arabic Language. Egyptian Computer Science Journal, vol. 30, no. 1, pp. 35-38.
- Alhawarat, M., Hegazi, M., and Hilal, A. (2015). Processing the Text of the Holy Quran: a Text Mining Study. International Journal of Advanced Computer Science and Applications (IJACSA), vol. 6, no. 2, pp. 262-267.

- Al-Kabi, M., Al-Belaili, H., Abul-Huda, B., and Wahbeh, A. H. (2013). Keyword Extraction Based on Word Co-Occurrence Statistical Information for Arabic Text. ABHATH AL-YARMOUK: "Basic Sci. & Eng.", vol. 22, no. 1, pp. 75- 95.
- Al-Kabi, M., Kanaan, G., Al-Shalabi, R., Nahar, K., and Bani-Ismail, B. (2005). Statistical Classifier of the Holy Quran Verses (Fatiha and YaSeen Chapters). Journal of Applied Sciences, vol. 5, no. 3, pp. 580-583.
- Al-Masiri, M. (2005). *Dalalat Al-Takdim wa Al-Taakhir fi Al-Quran Al- Karim, Dirassa Tahliliya* (In Arabic), Cairo, Egypt, Maktabat Wahba.
- Alrabiah, M., Alhelewh, N., Al-Salman, A., and Atwell, E. (2014). An Empirical Study on the Holy Quran Based On a Large Classical Arabic Corpus. International Journal of Computational Linguistics (IJCL), vol. 5, no. 1, pp. 1-13.
- Alrabiah, M., Al-salman, A., and Atwell, E. (2014). A New Distributional Semantic Model for Classical Arabic. 2nd International Conference on Islamic Applications in Computer Science and Technology (IMAN 2014), Amman, Jordan.
- Alrehaili, S. M., and Atwell, E. (2014). Computational ontologies for semantic tagging of the Quran: A survey of past approaches. In LREC 2014 Proceedings, European Language Resources Association.
- Al-Samiraii, F. (2006). *Al-Taabir Al-Qurani* (In Arabic), Amman, Jordan, Dar Ammar.
- Al-Soyouti, J. (1973). *Al-Itkan fi Oloum Al-Quran* (In Arabic), Beirut, Lebanon, Almaktaba Althakafiya.
- Al-Taweel, M. (2009). *Significance of conjunctions and the impact thereby on differences Muslim Scholars (Foqaha)*. In Arabic. Master Dessertation. An-Najah National University. Nablus. Palestine.
- Alvarez, F. J., Vaquero, A., Sáenz, F., and de Buenaga, M. (2007). Semantic Relation Modeling and Representation for Problem-Solving Ontology-Based Linguistic Resources: Issues and Proposals. The 9th International Conference on Enterprise Information Systems (ICEIS), pp. 59-70, Madeira, Portugal.
- Al-Yahya, M., Al-Khalifa, H., Bahanshal, A., Al-Odah, I., and Al-Helwah, N. (2010). An Ontological Model for Representing Semantic Lexicons: An Application on Time Nouns in the Holy Quran. The Arabian Journal for Science and Engineering, vol. 35, no. 2, pp. 21-35.
- Al-Yahya, M., Al-Malak, S., and Aldhubayi, L. (2016). Ontological Lexicon Enrichment: The Badea System for Semi-automated Extraction of Antonymy Relations from Arabic Language Corpora. Malaysian Journal of Computer Science, vol. 29, no. 1, pp 56-73.

- Al-Zujaji, A. (1984). *Horouf Al-Maani*. In Arabic. Beirut, Lebanon: Muassasat Al-Ressala.
- Aman, M., Bin md Said, A. B. A. S., Kadir, S. J. A., and Baharudin, B. (2017). a review of studies on ontology development for islamic knowledge domain. Journal of Theoretical & Applied Information Technology, vol. 95, no. 14, pp. 3303-3311.
- Arpirez, J. C., Corcho, O., Fernandez-Lopez, M., and Gomez-Perez, A. (2001). WebODE: a scalable ontological engineering workbench. In the 1st International Conference on Knowledge Capture (KCAP_01), ACM Press, pp.6–13, Victoria.
- Attia, M, Rashwan, M., Ragheb, A., Al-Badrashiny, M., Al-Basoumy, H., and Abdou, S. (2008). A compact Arabic lexical semantics language resource based on the theory of semantic fields. The 6th International Conference on Natural Language Processing (GoTAL 2008), pp. 65-76, Gothenburg, Sweden.
- Banchs, R. E. (2013). *Text Mining with MATLAB*, New York, USA, Springer.
- Baqai, S., Basharat, A., Khalid, H., Hassan, A., and Zafar, S. (2009). Leveraging Semantic Web Technologies for Standardized Knowledge Modeling and Retrieval from the Holy Qur'an and Religious Texts. In Proceedings of the 7th International Conference on Frontiers of Information Technology (p. 42). Pakistan.
- Baroni, M., and Bisi, S. (2004). Using co-occurrence statistics & the web to discover synonyms in a technical language. In Proceedings of the 4th International Conference on Language Resources and Evaluation, vol. 5, pp. 1725-1728.
- Bentrcia, R., Zidat, S., and Marir, F. (2017). Extracting Semantic Relations from the Quranic Arabic Based on Arabic Conjunctive Patterns, Journal of King Saud University - Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2017.09.004
- Bentrcia, R., Zidat, S., and Marir, F. (2018). An analytical study on the holy Quran based on the order of words in Arabic AND conjunction. Will appear in The Malaysian Journal of Computer Science, vol. 31, no. 1.
- Blei, D., Ng, A., and Jordan, M. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, vol. 3, pp. 993-1022.
- Blei, D., Griffiths, T., Jordan, M., and Tenenbaum, J. (2004). Hierarchical Topic Models and the Nested Chinese Restaurant Process. Advances in Neural Information Processing Systems 16.
- Bloehdorn, S. and Hotho, A. (2004). Text classification by boosting weak learners based on terms and concepts. In Proceedings of the 4th IEEE International Conference on Data Mining (ICDM), pp. 331-334.

- Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. In Proceedings of the Biennial GSCL Conference, vol. 156, pp. 31-40.
- Brank, J., Grobelnik, M., and Mladenic, D. (2005). A survey of ontology evaluation techniques. In Proceedings of the conference on data mining and data warehouses (SiKDD 2005).
- Brewster, C., Jupp, S., Luciano, J., Shotton, D., Stevens, R. D., and Zhang, Z. (2009). Issues in learning an ontology from text. BMC Bioinformatics, vol. 10, no. 5, pp. 1-20.
- Buitelaar, P., Olejnik, D., and Sintek, M. (2004). A protégé plug-in for ontology extraction from text based on linguistic analysis. In Proceedings of the 1st European Semantic Web Symposium (ESWS).
- Buitelaar, P., Cimiano, P., and Magnini, B. (2005). *Ontology Learning from Text: Methods, Evaluation and Applications*, IOS Press, Amsterdam, ISBN: 1-58603-523-1.
- Bullinaria, J. A., and Levy, J. P. (2007). Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study. Behavior Research Methods, vol. 39, no. 3, pp. 510-526.
- Church, K. W., and Hanks, P. (1990). Word association norms, mutual information, and lexicography. Computational linguistics, vol. 16, no. 1, pp. 22-29.
- Cimiano, P., and Volker, J. (2005). Text2Onto – a framework for ontology learning and data driven change discovery. In proceeding of 10th International Conference on Applications of Natural Language to Information Systems NLDB 2005, Spain, pp. 227–238.
- Cimiano, P. (2006). Ontology Learning and Population from Text: Algorithms, Evaluation and Application. New York, USA: Springer.
- Corcho, O., and Gomez-Perez, A. (2000). A roadmap to ontology specification languages. In Proceedings of EKAW'00, France.
- de Winter, J. C. (2013). Using the Student's t-test with extremely small sample sizes. *Practical Assessment, Research & Evaluatio*n, vol. 18, no. 10, pp. 1-12.
- Domingue, J., Motta, E., and Corcho Garcia, O. (1999). *Knowledge Modelling in WebOnto and OCML: A User Guide*.
- Dost, M. K. B., and Ahmad, M. (2008). Statistical profile of Holy Quran and symmetry of Makki and Madni surras. Pakistan Journal of Commerce and Social Sciences, vol. 1, no. 1, pp. 1-16.

- Dukes, K., and Buckwalter, T. (2010). A dependency treebank of the Quran using traditional Arabic grammar. The 7th International Conference on Informatics and Systems (INFOS), pp. 1-7, Cairo, Egypt.
- Dukes, K. and Habash, N. (2010). Morphological annotation of Quranic Arabic. The 7th International Conference on Language Resources and Evaluation (LREC), pp. 2530-2536, Valletta, Malta.
- Dukes, K., Atwell, E. (2012). LAMP: A Multimodal Web Platform for Collaborative Linguistic analysis. In Proceedings of the Language Resources and Evaluation Conference (LREC), pp. 3268-3275, Turkey.
- Dukes, K., Atwell, E., and Habash, N. (2013). Supervised Collaboration for Syntactic Annotation of Quranic Arabic. Language Resources and Evaluation, vol. 47, no. 1, pp. 33-62.
- Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, vol. 19, no. 1, pp. 61-74.
- Elabd, E., Alshari, E., and Abdulkader, H. (2015). Semantic Boolean Arabic Information Retrieval. The International Arab Journal of Information Technology (IAJIT), vol. 12, no. 3, pp. 311-316.
- El-Khair, I. A. (2006). Effect of Stop Words Elimination for Arabic Information Retrieval: A comparative Study. International journal of Computing & Information Sciences, vol. 4, no. 3, pp. 119-133.
- Evans, R. (2003). A framework for named entity recognition in the open domain. In Proceedings of the Recent Advances in Natural Language Processing (RANLP-2003), pp. 137-144.
- Farghaly, A., Shaalan, K. (2009). Arabic Natural Language Processing: Challenges and Solutions. *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 8, no. 4, pp. 1-22.
- Fortuna, B., Grobelnik, M., and Mladenic, D. (2007). Ontogen: semi-automatic ontology editor", in Proceedings of the 2007 conference on Human interface: Part II, pp. 309-318.
- Frantzi, K., and Ananiadou, S. (1999). The c-value/ nc-value domain independent method for multiword term extraction. Journal of Natural Language Processing, vol. 6, no. 3, pp. 145–179.
- Genesereth, M., and Fikes, R. (1992). *Knowledge interchange format. Technical Report*, Logic-92-1, Computer Science Department, Stanford University.

- Gomaa, W., and Fahmy, A. (2013). A survey of text similarity approaches. International Journal of Computer Applications, vol. 68, no. 13, pp. 13-18.
- Gruber, T. (1993). A Translation Approach to Portable Ontology Specification, vol. 5, no. 2, pp. 199-220.
- Guarino, N., Vetere, G., and Masolo, C. (1999). Ontoseek: Content-based Access to the Web, *IEEE Intelligent Systems*, vol. 14, no. 3, pp. 70-80.
- Habash, N. (2010). *Introduction to Arabic Natural Language Processing*, California, USA, Morgan & Claypool Publisher.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. (2009). The WEKA Data Mining Software: An Update. SIGKDD Explorations, vol. 11, no. 1.
- Hamam, H., Ben Othman, M. T., Kilani, A., Ben Ammar, M., and Ncibi, F. (2015). Data Mining in the Quran Using Aspects and Dependencies. The 3rd International Conference on Islamic Applications in Computer Science and Technologies (IMAN 2015), Konya, Turkey.
- Hand, D. J., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining (Adaptive Computation and Machine Learning)*, Massachusetts, USA, MIT Press.
- Hearst, M. A. (1992), Automatic acquisition of hyponyms from large text corpora. In Proceedings of the 14th International conference on Computational linguistics, vol. 2, pp. 539-545, Nantes, France.
- Islam, A., and Inkpen, D. (2006). Second Order Co-occurrence PMI for Determining the Semantic Similarity of Words. The International Conference on Language Resources and Evaluation (LREC), pp. 1033–1038, Genoa, Italy.
- Kalyanpur, A., Parsia, B., Sirin, E., Grau, B. C., & Hendler, J. (2006). Swoop: A web ontology editing browser. *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 4, no. 2, pp. 144-153.
- Kanaan, G., Al-shalabi, R., and Sawalha, M. (2003). Full automatic Arabic text tagging system. In proceedings of the International Conference on Information Technology and Natural Sciences, pp. 258-267.
- Kass, R. E., Eden, U. T., and Brown, E. N. (2014). *Analysis of Neural Data*. New York, USA: Springer.
- Khoja, S. (2001). APT: Arabic part-of-speech tagger. In Proceedings of the Student Workshop at NAACL, pp. 20-25.

- Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., and Suchomel, V. (2014). The Sketch Engine: ten years on. Lexicography, vol. 1, no. 1, pp. 7-36.
- Lassila, O., and Swick, R. (1999). Resource description framework (RDF) model and syntax specification. W3C Recommendation.
- Lee, C. S., Kao, Y. F., Kuo, Y. H., and Wang, M. H. (2007). Automated Ontology Construction for Unstructured Text Documents. Data & Knowledge Engineering, vol. 60, no. 3, pp. 547–566.
- Lin, D., and Pantel, P. (2001). Induction of semantic classes from natural language text. In Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 317-322.
- Lin, D., and Pantel, P. (2001). Dirt - Discovery of Inference Rules from Text. In Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 323-328.
- Liu, K., Hogan, W. R., and Crowley, R. S. (2011). Natural Language Processing Methods and Systems for Biomedical Ontology Learning. Journal of Biomedical Informatics, vol. 44, no. 1, pp. 163-179.
- Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, vol. 9, pp. 2579-2605.
- Maedche, A., and Staab, S. (2000). Discovering Conceptual Relations from Text. In Proceedings of ECAI 2000, IOS Press, Amsterdam.
- Maedche, A., and Staab, S. (2001). Ontology Learning for the Semantic Web. IEEE Intelligent Systems, vol. 16, no. 2, pp.72-79.
- Maniraj, V., and Sivakumar, R. (2010). Ontology languages- A review. International Journal of Computer Theory and Engineering, vol. 2, no 6, p. 887.
- Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. International journal of lexicography, vol. 3, no. 4, pp. 235-244.
- Momtazi, S., Khudanpur, S., and Klakow, D. (2010). A comparative study of word co-occurrence for term clustering in language model-based sentence retrieval. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT 10), pp. 325-328, California, USA.
- Myatt, J.G. (2007). *Making Sense of Data, a Practical Guide to Exploratory Data Analysis and Data Mining.* New Jersey, USA: John Wiley & Sons.

- Nassourou, M. (2011). A knowledge-based hybrid statistical classifier for reconstructing the chronology of the Quran, accepte in WEBIST/WTM 2011, The Netherlands.
- Navigli, R., Velardi, P., and Gangemi, A. (2003). Ontology Learning and its Application to Automated Terminology Translation. Intelligent Systems, IEEE, vol.18, no. 1, pp. 22–31.
- Nirenburg, S., and Raskin, V. (2004). *Ontological semantics*. Mit Press.
- Noy, N., Fergerson, R., and Musen, M. (2000). The knowledge model of protege-2000: combining interoperability and flexibility. In 12th International Conference in Knowledge Engineering and Knowledge Management (EKAW_00), Lecture Notes in Artificial Intelligence, vol. 1937, pp. 17-32, Springer, Berlin.
- Panju, M. H. (2014). *Statistical extraction and visualization of topics in the qur'an corpus*, Master's thesis, University of Waterloo, Ontario, Canada.
- Punuru, J., and Chen, J. (2011). Learning non-taxonomical semantic relations from domain texts. Journal of Intelligent Information Systems, vol. 38, no.1, pp.191-207.
- Reinberger, M.-L., and Spyns, P. (2005). Unsupervised text mining for the learning of dogma-inspired ontologies. In Buitelaar, P., Cimiano, P., and Magnini, B., editors. Ontology Learning from Text: Methods, Applications and Evaluation, number 123 in Frontiers in Artificial Intelligence and Applications, pp. 29-43. lOS Press.
- Saad, S., Salim, N., Zainal, H., & Noah, S. A. M. (2010). A framework for Islamic knowledge via ontology representation. 2010 International Conference on Information Retrieval & Knowledge Management (CAMP) (pp. 310–314). IEEE.
- Sabou, M., Wroe, C., Goble, C., and Mishne, G. (2005). Learning Domain Ontologies for Web Service Descriptions: an Experiment in Bioinformatics. In Proceedings of the 14th International World Wide Web Conference (WWW2005), Chiba, Japan.
- Safeena, R., and Kammani, A. (2013). Quranic Computation: A Review of Research and Application. Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences (NOORIC 2013), pp. 203-208, Madinah, Saudi Arabia.
- Saif, A. M., and Aziz, M. J. (2011). An Automatic Collocation Extraction from Arabic Corpus. Journal of Computer Science, vol. 7, no. 1, pp. 6-11.
- Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. Communications of the ACM, vol. 18, no. 11, pp. 613–620.

- Salton, G., and McGill, M. (1983). *Introduction to modern information retrieval*. McGraw-Hill, New York.
- Schiitze, H. (1993). Word space. Advances in neural information processing systems, vol. 5, pp. 895-902.
- Sharaf, A., and Atwell, E. (2012). QurAna: Corpus of the Quran annotated with Pronominal Anaphora. The 8th International Conference on Language Resources and Evaluation (LREC 2012), pp. 130-137, Istanbul, Turkey.
- Sharaf, A., and Atwell, E. (2012). QurSim: A corpus for evaluation of relatedness in short texts. The 8th International Conference on Language Resources and Evaluation (LREC 2012), pp. 2295-2302, Istanbul, Turkey.
- Shoaib, M., Yasin, M. N., Hikmat, U. K., Saeed, M. I., and Khiyal, M. S. H. (2009). Relational WordNet model for semantic search in Quran. 2009 International Conference on Emerging Technologies, IEEE, pp. 29-34.
- Siddiqui, M. A., Faraz, S. M., and Sattar, S. A. (2013). Discovering the Thematic Structure of the Quran using Probabilistic Topic Model. Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences (NOORIC 2013), pp. 234-239, Madinah, Saudi Arabia.
- Siegmund, B. (1998). *Data Analysis, Statistical and Computational Methods for Scientists and Engineers.* New York, USA: Springer.
- Sireteanu, A. N. (2013). A Survey of web ontology languages and semantic web services. Annals of the Alexandru Ioan Cuza University-Economics, vol. 60, no. 1, pp. 42-53.
- Strehl, A., Ghosh, J., and Mooney, R. (2000). Impact of similarity measures on web-page clustering. In Workshop on artificial intelligence for web search (AAAI 2000), vol. 58, p. 64.
- Su, X., and Ilebrekke, L. (2002). A comparative study of ontology languages and tools. In International Conference on Advanced Information Systems Engineering, pp. 761-765, Springer Berlin Heidelberg.
- Sure, Y., Angele, J., and Staab, S. (2003). OntoEdit: Multifaceted inferencing for ontology engineering. Data Semantics, vol.1, pp. 128-152.
- Ta'a, A., Abidin, S. Z., Abdullah, M. S., Ali, B., and Ahmad, M. (2013). Al-Quran themes classification using ontology. Proceedings of the 4th International Conference on Computing and Informatics ICOCI 2013, pp. 383-389, Malaysia.

- Tashtoush, Y. M., Al-Soud, M. R., AbuJazoh, R. M., and Al-Frehat, M. (2017). The noble quran Arabic ontology: Domain ontological model and evaluation of human and social relations. In *the 8th International Conference on Information and Communication Systems (ICICS)*, pp. 40-45, IEEE, Irbid, Jordan.
- Thabtah, F., Gharaibeh, O., and Al-Zubaidy, R. (2012). Arabic text mining using rule based classification. Journal of Information & Knowledge Management, vol. 11, no. 1, pp. 1-10.
- Ul Ain, Q., Basharat, A. (2011). Ontology driven Information Extraction from the Holy Quran related Documents. The 26th Institution of Electrical and Electronics Engineers Pakistan Students' Seminar (IEEEP 2011), Karachi, Pakistan.
- Velardi, P., Navigh, R., Cuchiarelli, A., and Neri, F. (2005). Evaluation of OntoLearn, a methodology for automatic population of domain ontologies. In Buitelaar, P., Cimiano, P., and Magnini, B., editors, Ontology Learning from Text: Methods, Applications and Evaluation, number 123 in Frontiers in Artificial Intelligence and Applications, pp. 92-106. lOS Press.
- Verma, J. P. (2013). *Data Analysis in management with SPSS Software*. New Delhi, India: Springer.
- Völker, J., Hitzler, P., and Cimiano, P. (2007). Acquisition of OWL DL Axioms from Lexical Resources. Proceedings of the 4th European conference on The Semantic Web, pp. 670-685.
- Weiss, S. M., Indurkhya, N., Zhang, T., Damerau, F. J. (2005). *Text Mining Predictive Methods for Analyzing Unstructured Information*. New York, USA: Springer.
- Willett, P. (1988). Recent trends in hierarchic document clustering: a critical review. Information Processing and Management: an International Journal, vol. 24, no. 5, pp. 577-597.
- Xu, J., and Croft, W. B. (2000). Improving the Effectiveness of Information Retrieval. ACM Transactions on Information Systems (TOIS), vol. 18, no.1, pp. 79-112.
- Yauri, A. R., Kadir, R. A., Azman, A., and Murad, M. A. A. (2012). Quranic-based Concepts: Verse Relations Extraction Using Manchester OWL Syntax. International Conference on Information Retrieval & Knowledge Management (CAMP), IEEE, pp. 317-321.
- Youn, S., & McLeod, D. (2006). Ontology development tools for ontology-based knowledge management. In Encyclopedia of E-Commerce, E-Government, and Mobile Commerce, pp. 858-864. IGI Global.
- Zakariah, M., Khan, M. K., Tayan, O., and Salah, K. (2017). Digital Quran Computing: Review, Classification, and Trend Analysis. Arabian Journal for Science and Engineering, vol. 42, no. 8, pp 3077–3102.

- Zavitsanos, E., Paliouras, G., and Vouros, G. (2006*). Ontology Learning and Evaluation: A survey*. Technical report, DEMO-(2006-3), NCSR Demokritos, Athens, Greece.