Design and Development of a Quote Validation Tool for Arabic Scripts.

by

Abdulrhman Alshareef

A Master's thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements for the degree of

Master of Science
in
ELECTRONIC BUSINESS TECHNOLOGIES

School of Electrical Engineering and Computer Science
University of Ottawa

Abstract

Over the past decade, there has been a tremendous development in e-publishing tools. The Arab world tendency towards electronic publishing has facilitated the prosperity of Arabic e-publishing over the Internet. Likewise, it has enabled the ordinary user to deploy documents, letters, opinions, and ideas with freedom and ease of use. Although freedom of expression should be guaranteed to everyone, it may be used to disseminate false or distorted information. This may lead to the loss of ordinary user's confidence in e-content. However, the user's confidence in e-content will increase if the credibility of the content is emphasized. There are many factors that challenge this task including not only the rapidly growth of Arabic digital publishing, the absent from control over electronic content, and the lack of e-publishing regulations and laws, but also how to develop an efficient framework to confirm the digital content authenticity. Therefore, the need to monitor the credibility of Internet content while maintaining freedom of expression to its users has become an urgent matter of debate. A flexible framework needs to be developed that will overcome these issues and allow for a comprehensible and comfortable content validation environment that would satisfy the end users' desires. This thesis proposes a framework that serves to confirm fundamental text authenticity in Arabic scripts on the Internet. This framework will demonstrate the design and the development of new quotes verification algorithm and the necessary components of framework design, development and implementation based on Service Oriented architecture.

Acknowledgements

My sincere appreciation is intended for my academic supervisor Prof. Abdulmotaleb El Saddik, who has supported, guided, and encouraged me throughout my academic success. This thesis would not be a success without the assistance he has provided me.

I would also like to extend my gratitude to my beloved grandmother and parents for their ongoing support. Special thanks to my beloved wife and dear daughters (Shaden & Rusiyl) for their help, encouragement, and support during my graduate studies. Lastly, I thank my family and close friends who have supported me.

I would also like to acknowledge the financial assistance of the King Abdul-Aziz University (KAU).

Table of Contents

List of Tables

List of Figures

Chapter 1 : Introduction

Arabic is one of the fastest growing languages used over the Internet. According to recent statistics, the estimated total number of Arabic Internet users reached 347,002,991 users during 2010[1]. This indicates that the highest growth rate amongst utilized languages over the internet is Arabic by more than 2500%, compared to its nearest rivals (Russian, and Chinese). Coupled with some reports and official statistics in select Arab countries, these studies underline the enduring increase in the number of Arab users in the internet[2][3][4][5]. This also predicts an increase in Arabic websites' deployment. Moreover, this emphasizes the potential expanding of Arabic electronic content over the Internet. Generally, the number of Arabic websites is increasing[6]. This would point to a steady expansion in e-publishing in the Arab world in order to reach Arabic electronic content over the Internet[7] [8]. Therefore, the need to support the Arabic language on the Internet has become an urgent necessity.

1.1 Motivation

The tremendous development in e-publishing tools today encourages the e-publishing movement. Likewise, the Arab world is exploiting these e-publication benefits, along with Arab leaders' invitations to increase online publications [9]. The websites' movement to support the Arabic language is steadily growing. In addition to this, freedom of expression and flexibilities that are offered by websites on the Internet such as social-networking sites, blogs and other[10][11][12]have motivated Arab users to exploit the Internet in a smoother and more affordable manner. Continually, social-networking sites facilitate the deployment of Arabic electronic content. However on the other hand, it keeps the developer's stride to achieve the aspirations of Arab users [13][14][15][16].

Arabic websites and publications may contain quotes or texts from given Arabic books such as: the Quran, 1001 Nights, Panchatantra, and so forth. In essence, most Arab

authors may support their articles with quotes from the Quran or other religious texts. Hence, they may base their conclusions and opinions on quotes from these texts. Generally speaking, regular readers are incapable of observing the authenticity of these quotes incorporated by the author. Therefore, they are unable of distinguishing whether or not the quote is distorted.

However, to verify the quote's authenticity, users have to conduct extensive research on the original reference. This is difficult for the ordinary reader who is likely not familiar with these particular books and references, especially if the author includes original quotes without identifying their specific source or place at the original reference. For instance, the author may include original quotes without including the specific chapter, page or paragraph numbers. Continually, this verification process is often very time-consuming and requires much effort to authenticate the information. These factors combined promote the need to develop a mechanism which would facilitate a "Quote validation process", especially for the ordinary Internet user.

This proposed tool aims to assure that the original quotes are not tempered with. In other words, it seeks to compare quotes of the fundamental text with those of the original text and to confirm the match. This would then emphasize the authenticity of Arabic quotes. As a result, this procedure would increase the ordinary user's confidence in Arabic digital contents, which would in turn lead to higher a demand for Arabic websites. It would also support the initiative to enrich the Arabic digital content in the Internet. Lastly, it would help provide oversight and control on Arabic digital contents without compromising the freedom of expression. As well as, it will guard the quotes from distortion, modification, and erroneous translation that results in the destruction of the credibility of e-citation from Arabic books [17].

Although there has been a previous study which examined the credibility of Internet article content [17], this study was aimed towards English scholarly websites. The author discussed the websites' conduct to validate the articles' references as well its commitment to document these references. This procedure leads to validate the content of their

articles. It explained the websites' obligation to automate the articles' documentation; however it did not address the issue of the quotes' authenticity on the particular websites.

This thesis is motivated by the fact that these issues surrounding authenticity have limited the credibility of e-citation from given Arabic books, making it necessary to come up with a flexible solution to overcome this. Among various other properties, the following features will be chosen as an objective of the tool design and implementation: Quotation filtering algorithm, Quotation validation mechanism, Quotation nearest similar, Original reference database, Services-Oriented Architectures (SOA), web services, and cloud computing.

1.2 Objective and Contribution

The objective of this thesis is to improve typical information retrieval environments and enhance them by exploiting new technologies. For this purpose, the new technology features will be used in order to make them more comprehensible and standardized as much as possible. The lack of a database for most Arabic books such as the Quran, Saheah Albukari, 1001 nights, Panchatantra, etc., is the first challenge that needs to be addressed in order to develop a quote's validation tool. Another objective of this thesis is to implement a Standard Web Service (SWS) based on the Extensible Markup Language (XML) to standardize the tool. Having said that, it will enables websites, web applications, office application, and web browsers to take advantage of the validation tool. It will also reduce the difficulty of interoperability (the ability to exchange the information between diverse systems), especially with the Arabic language[18]. The final objective is to develop a validation tool based on Services-Oriented Architectures (SOA) organizational pattern, in order to facilitate the integration of the required function. This would be coupled with the use of cloud computing infrastructure which aims to support a real-time service delivery.

The main contribution of this thesis is the design and development of a quote's validation tool for Arabic e-publishing (QVT). QVT will be designed based on SOA organizational pattern, which gives the power of combined services as many as possible. The QVT will be developed based on an XML web service and will provide, flexible systems' integration, the capability of services' customization, time reduction, and will overcome issues surrounding Arabic character recognition. Coupled with the power of the cloud computing infrastructure, this will result in the reduction of maintenance, infrastructure and costs. Besides this, it will increase the capability of service delivery and remote access [19].

As previously mentioned, previous research in this field discussing electronic content credibility focused on English [17], [20–23] materials or plagiarism detection concepts[15]. In other words, the general search concept used was based on a word detect function[15], [16]where scholarly content was used to confirm this credibility. To the best of my awareness, the only study which has discussed the same concept has focused on weblogs[24]. Not only do they lack a reliable database and rely on the Internet, but this leads back to the same problem that this thesis addresses. However, in this thesis, a new mechanism is introduced which serves to confirm the quote authenticity in Arabic scripts based on the fundamental texts.

Finally, the new QVT mechanism will be implemented in order to verify the authenticity of the quotation based on fundamental text, while keeping it as affordable as possible for its users. QVT is designed in such a way that it minimizes maintenance costs, input errors, and delivery time.

1.3 Thesis Organization

In chapter 2, the thesis literature’s background is reviewed. Related work about Internet searches in general and Arabic searches in particular will be discussed. In chapter 3, QVT system's design will be presented through reviewing the system architecture and

the software design. In Chapter 4, QVT implementation's details will be provided. In chapter 5, the validation process will be demonstrated along with measurement data in order to validate QVT system. In chapter 6 a conclusion will be debated and future work will be discussed.

1.4 Publications

The scientific outcome of this thesis is listed below:

1. **Abdulrhman Alshareef** and Abdulmotaleb El Saddik, “A QURANIC QUOTE VERIFICATION ALGORITHM FOR VERSES AUTHENTICATION”, IEEE 8th International Conference on Innovations in Information Technology (IIT'12), AbuDhabi, Al-Ain, UAE, 2012.

Chapter 2 : Background and Related Work

2.1 Literature Background

In this chapter, the thesis' main components are illustrated. In order to develop a quote validation system, the system relies on the concept of text retrieval to investigate the intended quote. Thus, quote authenticity is the first term that will be reviewed in order to provide a framework for design and development of the proposed tool in this thesis. Text retrieval depends on the existence of a reliable database. The lack of aggregate, authoritative and accessible databases of given Arabic books is one of the reasons that prompted the existence of this thesis[21], [25]. The original reference database term will also be discussed. Since the basic framework of a quote validation tool is based on a web service, the technical essentials related to this technology will be mentioned.

2.1.1 Quote Authenticity

Quote authenticity is a set of policies and procedures that are necessary in order to validate the intended quotation. This process is one of the stages that are used by researchers in order to analyze information credibility on the Internet. Analysis process contains four stages of investigation to verify the information credibility of the web content [20]. Researchers in the field have addressed the following credibility issues for e-content:

- Content [17], [20], [22].
- Author's information [20], [22–24].
- System's design [20], [21], [24].
- Content authentication [17], [20], [23], [24].

Accordingly, emphasis on the validity of the quotations will be reflected in the confidence of Internet users [20], [21], [23], [24], [26].

The validation prototype contains two main components: 1) Text retrieval mechanism and 2) An original reference database. These components will be reviewed in terms of applying the criteria as well as the implementation of methodology to support the Arabic language due to its distinctive characteristics [27].

2.1.1.1 Text Retrieval

Text retrieval is the essential system used by researchers to locate any portion of information in a large-scale database. It is considered a part of the information retrieval system, which is particularly interested in searching within texts. Retrieving related information from a database in response to a user's request is the main concern within the information retrieval field [28]. However, the performance of these information retrieval systems needs to be evaluated in order to enhance them. Enhancing the performance will in turn lead to raising the accuracy of the extracted results. Evaluating the performance, precision and recall metrics are the common measurement tools used by researchers [28]. Although the use of these metrics may be considered practical, they are considered subjective due to the nature of this thesis. To enhance the systems accuracy, we need to do more than simply increase the precision ratio. Generally speaking, the accuracy measures how close the output result is to the expected result (actual value). The precision measures how close the results are to one another. In essence, it measures the possibility of repeating the same output when using the identical input. Therefore, to raise the efficiency of the system, the need to improve accuracy is essential while also considering the high precision.

Generally, text retrieval systems contain two main components: search algorithm and a document database. Figure 2.1 shows the basic searching architecture for text retrieval [29]. The service layer is primarily focusing on extracting the information from the indexed database in the storage layer. The user interface layer is responsible for communicating with the user in order to receive its queries and to display the results. Once any request has passed through the user interface layer, the service layer considers request queries as strings of words and tries to find them in an indexed database[28].

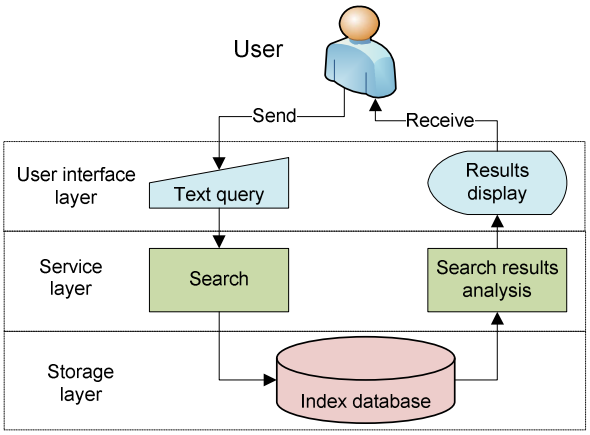

Figure 2.1 Text Retrieval Architecture [29]

Such systems contain certain techniques which assist in extracting results quickly. Full text search is one of the assistive techniques which is used in order to retrieve information as quick as possible. This was recognized in the 1990s[28]. In order to fulfil the user's request, the full-text search algorithm attempts to inspect every word in the stored documents to match the texts supplied by the user. Hence, the accuracy and precision of the information retrieved by this technique needs to be enhanced. There is a number of patterns that can be implemented with the full text searching technique to improve the accuracy of the retrieved information. "Phrase-based" and "regular expression" are distinct search patterns that may be used to enhance search performance.

In the phrase-based pattern, the search engine allows users to search for information based on an exact match between the given sentence or phrase and the retrieved one [30]. Using this pattern is essential to meet this thesis's objective which aims to match the fundamental text to the quote's authentic validation. This pattern is based on entire sentence matching rather than being restricted to keyword matching. These days, phrase-based searching is one of the most influential concepts in terms of optimizing the textual content. Considering the phrase search pattern leads to better accuracy than keyword search patterns in terms of this thesis's objective [31].

In computing, in regard of match, identify or discernment of any set of text, a regular expression is a popular method to accomplish this due to its ease and flexibility. A set of text could be characters, words, phrases, sentences or patterns of characters. A regular expression is a collection of formal and informal language (meta-characters) that is recognized by a regular expression compiler[31]. A regular expression compiler is part of a software program which aims to inspect a given text and discover information that match the desired pattern[32]. For this purpose, a regular expression occupies powerful meta-characters that can be used to indicate retrieval patterns with high precision.

The major purpose of the proposed system is to support the Arabic language due to it's distinctive characteristics. In order to raise the efficiency of the proposed system, it will focus on the power of the text retrieval system by operating a combination of phrase-based pattern searches as well as a regular expression pattern search. Thereupon, the system accuracy will be improved and the precision will improve.

2.1.1.2 Original Reference

The original reference is a relational database (RDB) that consists of a collection of genuine texts from the most famous fundamental Arabic books such as the Holy Quran, Saheah-Albukari, 1001 nights, Panchatantra, etc. A relational database is generally a collection of information or data related to each other through a logical connection that is established in formally defined tables. It will facilitate the process of accessing and retrieving the information from the database storage. Furthermore, a relational database has the advantage of being expanded even after it has been established without the need of any adjustment.

Original reference RDB is a set of schemas containing tables. Each schema represents an Arabic book's genuine script with its translations. Each table contains data fixed into predefined columns with their own unique attributes. For example, a digit column will only accept numbers and so forth. Each row contains a unique set of data

that is consistent with the attributes and classification of each column. For instance, a typical translation entry database would include a table that described the text in the original language with columns for original text id, book id, chapter id, and so forth. Another table would describe a text translation with columns for original text id, language id, translated text and so on.

Original reference RDB runs in a relational database management system (RDBMS) server. Continually, the RDBMS server provides multi-user access to a number of RDB. Structured query language (SQL) is a standard interactive language that will be used in order to handle the original reference RDB. SQL statements will be used to create queries to retrieve the information from the Original reference RDB.

The lack of an electronic aggregate of given Arabic books, in addition to the difficulty regarding access to paper copies converted to electronic files is the reasoning behind using the relational database system. , It also drove the trend of establishing the Original reference RDB.

2.1.2 Introduction to Web Service

Web service is a standard principle that facilitates the approach of integration of web-based applications, even considering the variation in the platform infrastructure. Web service exploits open-source technologies to exchange data in an easy and affordable manner with regards to Internet protocol. Employing the web service will enhance interoperability between the proposed tool, and any web-based applications. XML, SOAP, and WSDL are the open standard technologies that are used by the web service (see Figure 2.2).

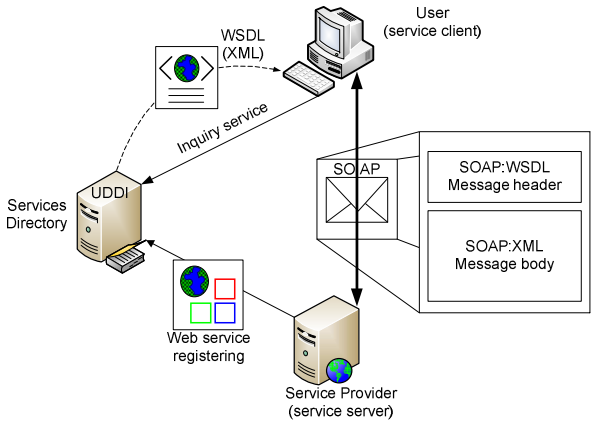

Figure 2.2 Web Service Architecture [33]

Extensible Markup Language (XML) is a simple markup language consumed as a metadata to tag given information. The simplicity of the XML comes from it being a machine-readable and human-readable language. As well, an XML is based on adjustable text that is derived from the Standard Generalized Markup Language, which is one of the standards adopted by ISO (ISO 8879:1986 SGML) [34]. It will emphasize the generality and usability of using the XML over the Internet between diverse systems. Web service is using the XML as a basic language for the representation of subjective data structures either for the SOAP or WSDL.

Simple Object Access Protocol (SOAP) is a special protocol for exchanging information in the form of a structured data package. A SOAP package consists of the XML message that is transmitted over Application Layer protocols (ALP's) between clients and servers. SOAP relies on ALPs for message negotiation and transmission. In essence, Web service is using SOAP as a means of transportation protocol to transmit the data to and from the service requester over Hypertext Transfer Protocol (HTTP) for most services [34].

The Web Services Description Language (WSDL) is used to explain the web service's operational attributes based on the XML language. A WSDL description

contains the data exchange operational details, such as how it can handle the SOAP messages in terms of the expected input parameters. Besides this, it describes the return's data structures in response to the service request. Currently, the development of software programs can easily recognize operational details through the description file. The necessary actions will then be accomplished automatically in order to handle SOAP messages [34].

Web services allow various applications from different platforms to integrate with each other without the need to adjust the code. Web services are an XML-based language that makes it effortless to integrate with any operating system or programming language. Despite the differentiation in the application platforms and programming languages, the application's interface can be developed by software developers to integrate web services with it. For instance, C# can communicate with Java, Mac applications can communicate with Windows applications. Furthermore, it facilitates web services customization based on application's requirements. This process results in the reduction of time needed and human errors made.

Generally, traditional client/server models provide the user with a graphical user interface (GUI). Although the client/server model is the main pattern of the web service, the web service distributes the operational attributes, data, and processes throughout a systematic interface over the Internet instead of using GUI. A web service can be developed based on the Services Oriented Architectures approach, which allows for easy adoption, customization, and adjustment. This is coupled with the use of cloud computing infrastructure to maintain an available, accessible, and convenience service.

2.1.2.1 Services Oriented Architecture

The Services-Oriented Architecture (SOA) is an organizational pattern which illustrates the approach to designing and developing services in order to fulfil the user requirements, with the consideration of the system's interoperability [35]. These services are software components with detailed features that can be used in any structure pattern.

These components allow remote access over a standard ALP in the Internet to exploit their capabilities. Potential integration is what distinguishes these services. This is due to the use of open standards for the development of such services, which promote the principle of services loosely coupling. Thus, it enhances the service's chance to be used again by any application or development platform [35] (see Figure 2.3).

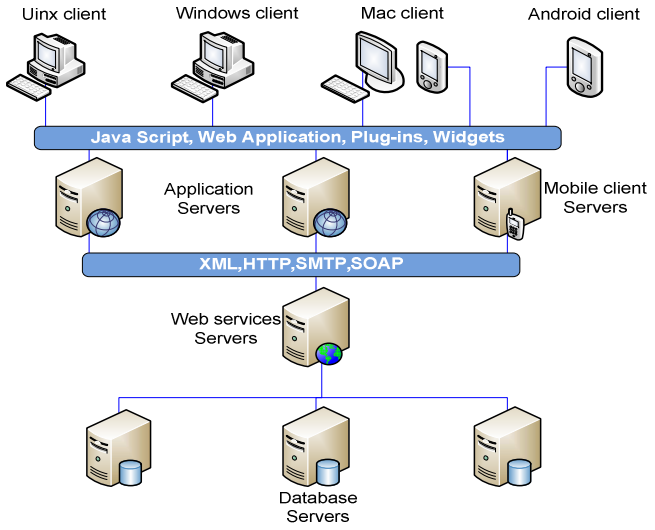

Figure 2.3 SOA Pattern

This approach facilitates the integration of the required services according to the program's strategy. This architecture allows development change and radical coding changes in the internal application infrastructure without altering the entire program [36].

2.1.2.2 Cloud Computing

Cloud computing is a computing service model which facilitates the products, services and solutions delivery in real time over the Internet. Currently, cloud computing is one of the most controversial topics in information technology. The use of cloud computing infrastructure to deliver services captures the attention of everyone, whether it is the general public or specialists.

Cloud computing offers efficient attributes to handle applications, platforms, web services, and data in a distributed, ubiquitous and global approach[37]. Cloud computing gives text retrieval systems the opportunity to expand. To promote proficient resources' utilization, this thesis will focus on the power of the web service by employing a combination of SOA pattern and Cloud computing infrastructure [38] (see Figure 2.4).

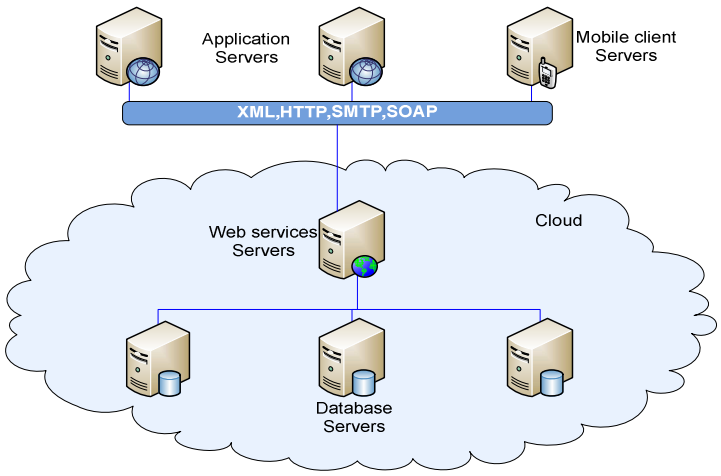

Figure 2.4 SOA Pattern Over the Cloud

Cloud computing's dynamic features offer the necessary development and flexibility to allow swift expansion in collection size. Besides this, it affords a potential program's improvement to serve the largest number of users with minimum administrative efforts [39]. In spite of this, information growth over the Internet is fast. The challenge to locate a related and a trustworthy source of information makes information retrieval systems such as QVT valuable.

The main objective of using cloud computing service is to reduce infrastructure range, maintenance work, software license, and administration oversight. Accordingly, it will support the trends of Green IT that aims to reduce cost, increase capability, and use the available service.

2.2 Related Work

2.2.1 Text Retrieval

Text retrieval, usually known as an information retrieval system that supports text-based searching, refers to systems that support data retrieving based on predetermined criteria. Most Internet search engines such as Google and Yahoo are using the information retrieval system's concept in building their search engines. In 2000, Pinkerton [28] promoted the use of full-text search in the search area. Pinkerton's work has raised the information retrieval system to a level which makes it easier for inexperienced users to find information quickly and easily through his assumption that Internet users are in the first place "naive and unsophisticated users". Additionally, he uses a full-text search based on a data indexing approach to swift the data searching with high precision.

A necessary and primary aspect in a text retrieval system is finding accurate information in a large-scale collection database with high precision. Patterson, Watters, & Shepherd [30] discussed the accuracy of text retrieval based on phrase-based search pattern in an academic study which addressed comparative evaluations of different text retrieval patterns. The study results showed promising accuracy improvement in retrieved results; these promised results are due to the utilization of the phrase-based search pattern [39]. The study showed that the use of a phrase search pattern can fail to retrieve results due to invalidity of the inputted sentence. In addition, the accurate results depend on the correctness of the inputs. Therefore, using this method will also help to validate the quotes entered in the proposed tool. It is possible users will fail to remember a quote correctly when they want to make sure it is accurate, so in this situation, the tool will show that the input quote is actually invalid based on the input and the user may need to retrieve some segments of the correct phrase. To do so, other methods will be used in order to search for the correct sentence close to the entered quote.

Zhenjun and Xiangyu (2009) [32] discussed the text retrieval systems for the Chinese language. Due to the uniqueness of the language, the authors have illustrated how to take advantage of regular expressions in order to search for texts to match them based on text patterns in the language. Likewise, Arabic is unique, so the use of regular expressions in text retrieval systems will help to retrieve the authentic sentence close to the entered quote. The capabilities of regular expressions give the text retrieval system the ability to search effectively. Regular expression methods give search engines the ability to customize search patterns based on the specific language's characteristics [32].

Software implementation architecture is the main aspect in the text retrieval systems development. Moreover, diverse development circumstances will change the software demands for each different component. Cutting, Pedersen, and Halvorsen [40] discussed the implementation architecture for the text retrieval system. In order to build systems quickly, developer must consider its potential ability to be exploited by other programs. It has the ability to be customized based on the circumstances of the development's strategies. The above authors illustrated the advantage of the object-oriented approach which is the essential prototype of the SOA. Using this approach facilitates the isolation of functional components and furthermore, it facilitates components combining to implement any system.

The SOA approach will be used to combine phrase-based and regular expressions in order to develop the text retrieval system. The combination of these patterns will in turn improve the text retrieval system. Thereupon, system efficiency will be enhanced as well.

2.2.2 Text retrieval in Arabic

Arabic is a Semitic language, characterized as being the most difficult language in the treatment of written and spoken. This is due to the grammatical, morphological, vocal and phonological characteristics of the language. The Arabic alphabet consists of 28 letters in contrast of Latin alphabets; and some of these characters may be demonstrated

in different forms. This is due to the characters' position within a word and the characters' vocal sound Such as Hamza on the letter Alif [14].

Over 1 billion Muslims around the world use Arabic as their main means of communication for prayer and scholarly religious conversation [13]. An Arabic word may consist of a grammatical pattern, and typically it is composed of one or more of roots, pronouns, prepositions, conjunctions, affixes and prefixes all in one morphological structure [41]. As an example, “waakalattha” is an Arabic word which comprises the following roots: (verb)”akalat”, pronoun “t”, conjunction letter “wa”, and noun “ha”. This Arabic word is translated to the English phrase consisting of four words: “and she ate it” (see Table 2.1) [42].

	English phrase	Arabic word	
Complete	and she ate it	waakalattha	وأكلتها
Detailed	and	wa	و
	ate	akalat	أكلـ
	she	t	ـتـ
	it	ha	ـها

Table 2.1 Arabic Word Morphological Structure

Arabic has a high degree of sentence structure variation. Unlike English, the characteristics of the Arabic language create potential problems in any Arabic text retrieval system [41], [42]. Moukdad and Large (2001) [42] had listed the most influence characteristics that may cause difficulties in Arabic Information Retrieval (AIR). This study was focused on the possibility of using English search engines to search in Arabic. However, the authors suggested that when using English search engines, the language modification is necessary to overcome the potential problems that arise from the Arabic language characteristics [27].

Al Ameed and et al. (2006) [41], founded a new approach to enhance Arabic search engines. Their study developed a prototype that demonstrated improved performance in

retrieving the data. However, the study developed the concept of a word-sense search while as it is explained, the phrase search pattern is the objective of this thesis. Ataa-Allah and et al. (2006) [14], discussed the performance of AIR systems. The study developed a model which aimed to improve the performance of AIR systems to consider the morphological and grammatical pattern rules. In a recent study that discusses Arabic plagiarism's concerns[15] ,that authors intend to develop a mechanism to detect plagiarism in Arabic document to support the Arabic language. This motivates us to elaborate a validation tool to confirm the authenticity of the Arabic quotes.

This adjustment in the text retrieval system is necessary to overcome potential difficulties that may occur due to specific Arabic language characteristics. An improvement will be proposed in the system accuracy and precision by considering the language syntactic rules in the system development by implementing rules to overcome the characteristics' issues.

2.2.3 Text retrieval Over the Cloud

Most of the text retrieval systems over the cloud are either representing a digital library or a search engine which benefit from the cloud computing infrastructure. Services such as database, application hosting, web hosting, e-mail, and virtual private network (VPN) are a valuable for individual users and small and medium enterprises (SME).

Teregowda, Urgaonkar and Giles (2010) [38] argued about the challenges that faced the digital library search engines. They illustrated "CiteSeer" as an example of a search engine in the digital library business and expected the expansion of a digital library in the database/the potential extra services to be implemented in order to improve the performance of the search engine and to encourage the authors to look for alternative infrastructure. They found that 'infrastructure virtualization" and "cloud computing" are acutely smart alternatives for the traditional search engines' infrastructure. The study's

results demonstrate that a cloud implementation of information retrieval and digital library systems may be a practical substitute for its sustained business and expansion.

Lagerspetz and Tarkoma (2011) [43] presented the advantages and disadvantages of mobile desktop search over cloud. Cloud computing provides a practical management of information data in a distributed, ubiquitous and common environment. Likewise, it facilitates the integration between several platforms, systems and applications. For instance, an Android health mobile system [37] was developed using cloud infrastructure to enable data save, update, and search over the cloud in an accessible way.

Cloud computing supports the exchange of data through a distributed and ubiquitous environment with the advantage of a global access. In the same way, it encourages the collaboration between various applications, platforms, and web services. The text retrieval systems may be considerably enhanced with cloud-based technologies.

Chapter 3 : System Design

In this chapter, the design and architecture of the Quotes Validation tool for Arabic script (QVT) will be introduced. At the beginning, the layered architecture of QVT will be demonstrated. In this part, the logical layers of QVT will be explained in each layer. Thereafter, QVT software design technique will be illustrated by the unified modeling language (UML) diagrams that include structure diagrams, behavior diagrams, and interaction diagrams.

3.1 QVT Architecture

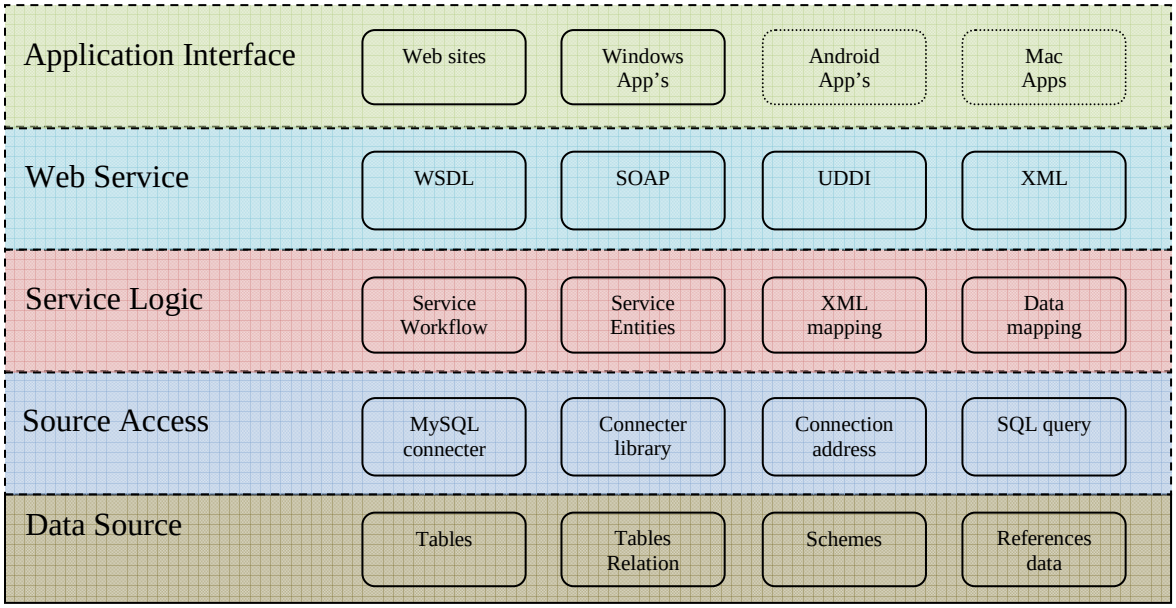

Figure 3.1 QVT System Architecture

Figure 3.1 illustrates the proposed QVT system architecture. QVT has five layers: (1) Application Interface (AI), (2) Web Service (WS), (3) Service Logic (SL), (4) Source Access (SA), and (5) Data Source (DS). Each layer offers various components to facilitate the development of interoperable services. The following section will review the components that were established in each layer and the layers' integration will also be clarified.

3.1.1 Application Interface (AI) Layer

The Application Interface (AI) layer is a hypothetical layer on the user surface within a QVT system based on SOA. It contains the system's user interface for diverse systems and platforms. It is considered as a means of communication between the users and the QVT system. Usually, users use different applications or platforms which can access the QVT system via AI layer. The independence of the AI layer allows the system to acquire various interfaces for different systems, whether it be in the programming language or application platforms.

The AI layer architecture was developed based on Model View Controller (MVC) architecture. Figure 3.2 illustrates the AI architecture. The AI layer generates two events (1) input and (2) output. The input event is allowing the users to request a service from the system. The output event is allowing the system to demonstrate the outcomes of the users' request. Either of both events must pass through the WS layer in order to achieve the user's request.

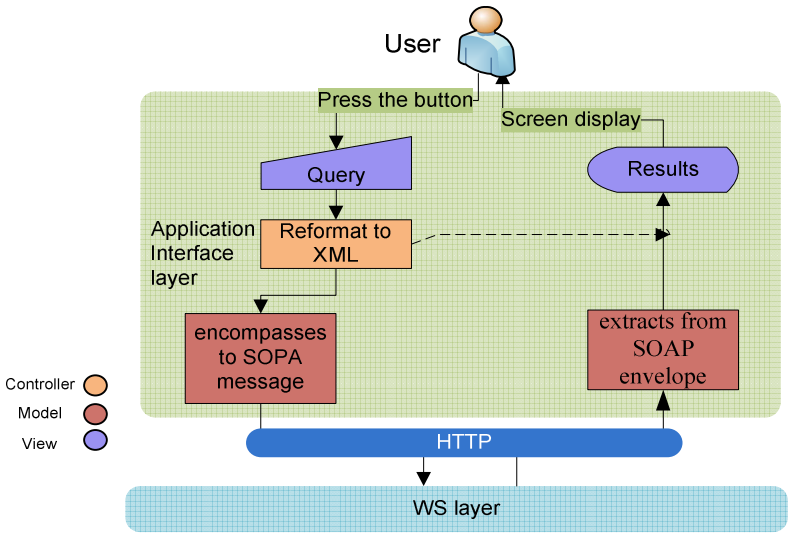

Figure 3.2 Application Interface Layer Architecture

In AI layer, the input event will be handled by the controller in order to convert it into the appropriate action. For this purpose, the controller's component reformats the input's request to XML language, which is understood by the SOAP model. Then, the SOAP model encompasses the request into an SOAP message which facilitates the message exchange. Subsequently, it is sent via the WS layer to the system core layers.

The results will be received via WS layer from the system core layers. Meanwhile, the SOAP model extracts the results from the SOAP envelope and is reformatted from XML language to the application's language. Later, the output's event will be demonstrated on the user's monitor.

3.1.2 Web Service (WS) Layer

The Web Service (WS) layer is a collection of principles and technologies that regulates the integration process between the service provider and the service requester. A web service uses functional technologies that make the process of information exchange easy and affordable, even considering the variation in the platform infrastructure. XML, SOAP, and WSDL are the open standard technologies that are used by the web service[33].

WS layer includes useful information for the developers to integrate the services. It acts as an intermediary to transfer requests from the service requester to the service provider and vice versa. Besides the open technologies, among the components of WS layer is the Universal Description Discovery and Integration (UDDI) which operates as a service directory. Usually, UDDI contains the service's information, the expected result and the invoking directions. In the following section, these components will be discussed in detail [35] (see Figure 3.3).

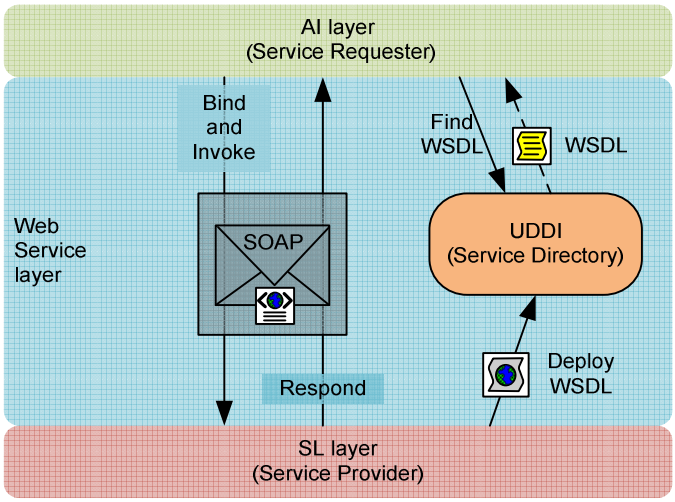

Figure 3.3 Web Service Layer Architecture

3.1.2.1 XML

Extensible Markup Language (XML) is a simple, usable markup language which is used to describe the data objects, process behaviors, and document information. XML is a variable machine-readable language based on the Standard Generalized Markup Language (SGML) [34]. The XML emphasizes the simplicity and universality through the use of any possible Unicode character for all international languages throughout the world.

The XML is the essential language for the representation of subjective data structures in the web service's components. As a rule, the web service's components generate and exchange a well-formed XML document (see Figure 3.4). Well-formed XML document are XML documents that have satisfied the XML specification rules. The following are some of the specification rules that need to be fulfilled [44]:

- **Document structures**: There is only one father element (root) that contains all of the child elements. Besides, it may begin with XML declaration statement.
- **Document characters**: The document must enclose only a legal Unicode character.

- **Markup and Content**: Special syntax characters such as "<" and "&" is prohibited unless they are used in markup description positions.
- **Tag**: "*start-tags*", "*end-tags*" *and* "*empty-element tags*" must be defined appropriately devoid of misplacement and overlapping.
- **Element**: The beginning and end of the element tag's name must be matched exactly due to it is case-sensitivity.
- **Tag's names**: Special characters such as ! , " , # , $, % , & , ' , (,) , * , + , , / , ; , < , = , > , ? , @ , [, \ ,] , ^ , ` , { , | , } , ~ are unacceptable . Furthermore, it is illegal to start the Tag's name with a dash (-), dot (.), space character, or numeric digit.

```
POST /QuranAuthenticity.asmx HTTP/1.1
Host: localhost
Content-Type: text/xml; charset=utf-8
Content-Length: length
SOAPAction: "http://QouteAuthenticity.org/AyahSearchReturnData"

<?xml version="1.0" encoding="utf-8"?>
<soap:Envelope xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
               xmlns:xsd="http://www.w3.org/2001/XMLSchema"
               xmlns:soap="http://schemas.xmlsoap.org/soap/envelope/">
  <soap:Body>
    <AyahSearchReturnData xmlns="http://QouteAuthenticity.org/">
      <SuraId>string</SuraId>
      <AyahId>string</AyahId>
    </AyahSearchReturnData>
  </soap:Body>
</soap:Envelope>
```

Figure 3.4 A Well-Formed XML Document.

3.1.2.2 SOAP

The Simple Object Access Protocol (SOAP) is an XML-based protocol for exchanging information in a manner of XML well-formed document. A SOAP package is actually transmitted over ALPs between the service requester and the service provider. The ALPs are the core channel for SOAP message cooperation and communication [34].

SOAP is considered as the keystone of the web service stack. In essence, it is composed of three elements: an envelope, a collection of encoding rules, and a remote procedure call representation. The SOAP envelope comprises the message contents and the description of service handling procedure. The SOAP encoding rules define a management approach to exchange the predefined data types. The SOAP remote procedure called representation is a convention that represents the service's request and response procedures [44].

WS layer usually handles two types of SOAP messages: requests and responses. Each message is a SOAP envelope. The SOAP envelope includes two main elements: the header and the body. The SOAP header is an elective element; it mostly encloses the service specific information about the SOAP message. The SOAP body is a compulsory element; typically it encloses the SOAP message content that is proposed to transmit it to the message's endpoint. Figure 3.5 illustrates the SOAP message's pattern for the QVT system.

(a)

```
POST /QuranAuthenticity.asmx HTTP/1.1
Host: localhost
Content-Type: application/soap+xml; charset=utf-8
Content-Length: length

<?xml version="1.0" encoding="utf-8"?>
<soap12:Envelope xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema"
    xmlns:soap12="http://www.w3.org/2003/05/soap-envelope">
  <soap12:Body>
    <AuthnReturnData xmlns="http://QouteAuthenticity.org/">
      <Quote>string</Quote>
    </AuthnReturnData>
  </soap12:Body>
</soap12:Envelope>
```

(b)

```
HTTP/1.1 200 OK
Content-Type: application/soap+xml; charset=utf-8
Content-Length: length

<?xml version="1.0" encoding="utf-8"?>
<soap12:Envelope xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema"
    xmlns:soap12="http://www.w3.org/2003/05/soap-envelope">
  <soap12:Body>
    <AuthnReturnDataResponse xmlns="http://QouteAuthenticity.org/">
      <AuthnReturnDataResult>
        <xsd:schema>schema</xsd:schema>xml</AuthnReturnDataResult>
    </AuthnReturnDataResponse>
  </soap12:Body>
</soap12:Envelope>
```

Figure 3.5 SOAP Message Pattern (a) Request (b) Response.

3.1.2.3 WSDL

The Web Services Description Language (WSDL) is used to explain the web service's functional attributes. Like SOAP, WSDL is based on an XML well-formed document. Ordinarily, the exchange functional instructions are described in the WSDL document. The expected input parameters and the response data structures are examples of the functional attributes that are included in the WSDL document. WSDL is one of the most important holders in the web service stack. Furthermore, it is considered as an instruction manual in any web service system [34]. These days, development software programs can easily initialize the functional attributes for web service in the WSDL document. Likewise, they can also realize the WSDL functional attributes automatically that are necessary to handle SOAP messages.

WSDL is used by the service's provider to deploy its services over the Internet. For this purpose, the service's provider publishes the WSDL document through the UDDI registry (this will be explained later on). On the other hand, the service's requester is attempting to discover the WSDL document for the services that he wanted within the UDDI registry [45].

3.1.2.4 UDDI

The Universal Description, Discovery and Integration (UDDI) registry is an XML-based independent platform that determines a common standard to detect and invoke web services. Individuals, SMEs, and large enterprises can benefit from this service to register their web services on the Internet and discover other web services. The Organization for the Advancement of Structured Information Standards (OASIS) has led initiatives to design, develop, and manage the UDDI registry.

UDDI was initially developed as a web service basic standard [46]. The objective of developing UDDI registry is to support the adaption of the integration technique in the

e-commerce industry. The registry was designed to interact with the SOAP messages to provide access to WSDL documents for the directory listed services.

3.1.3 Service Logic (SL) Layer

The Service Logic (SL) layer hosts the QVT logical entities, service workflow, communications functions and data mapping. The SL layer is the core tier for the QVT system as it controls the relationship between the applications and infrastructure layers. It encodes the logical functions that are responsible for making decisions about input validation and display formatting. Furthermore, it takes the responsibility of activating the functional entities based on the data received from the inquirer. In sum, the SL layer is the mastermind of the QVT system.

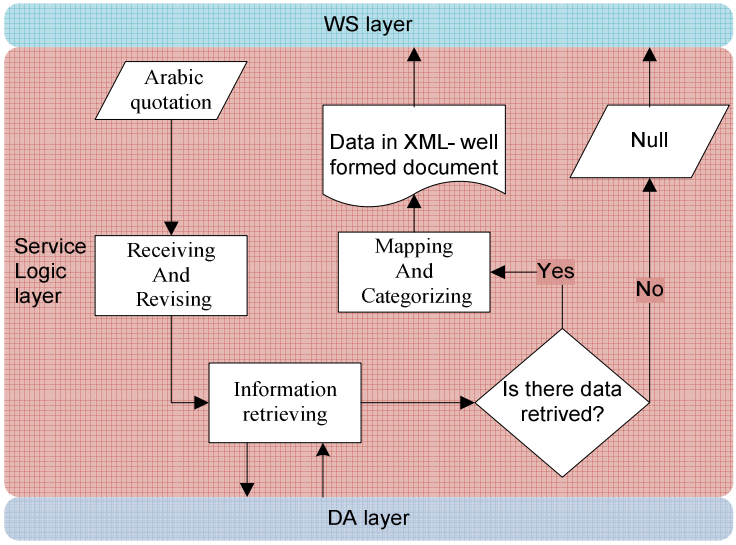

Figure 3.6 Service Logic Layer Architecture.

The QVT system is structured into three stages: (1) data receiving and revising, (2) information retrieving, and (3) data output mapping and categorizing (see Figure 3.6). In the first stage, the system takes care of the revision of the specific Arabic quote through noticing the morphological and syntactic details of the intended text based on Arabic language rules. In the second stage, the system works on retrieving the information from

the database of the given Arabic books. In the third stage, the system will assure that there is information retrieved based on the quote search. Therefore, the system will map the retrieved information and categorize it in order to facilitate the data accessibility, comprehensiveness, and utilization. These stages will be discussed in detail in the following section.

3.1.3.1 Data Receiving and Revising

Due to the distinctive nature of the Arabic language, most search algorithms fail to accomplish their task precisely and accurately. Arabic morphological and phonological characteristics are what limit the capabilities of traditional search engines to provide acceptable results. As it has been explained, the Arabic alphabet consists of 28 letters; and some of these characters are demonstrated in different forms. Continually, Arabic books usually use Arabic diacritics (Harakat) to clarify the phonological sounds of the character based on the grammatical rules. Some Arabic books are obtaining their own unique symbols such as the Quran which in turn creates additional obstacles to traditional search engines.

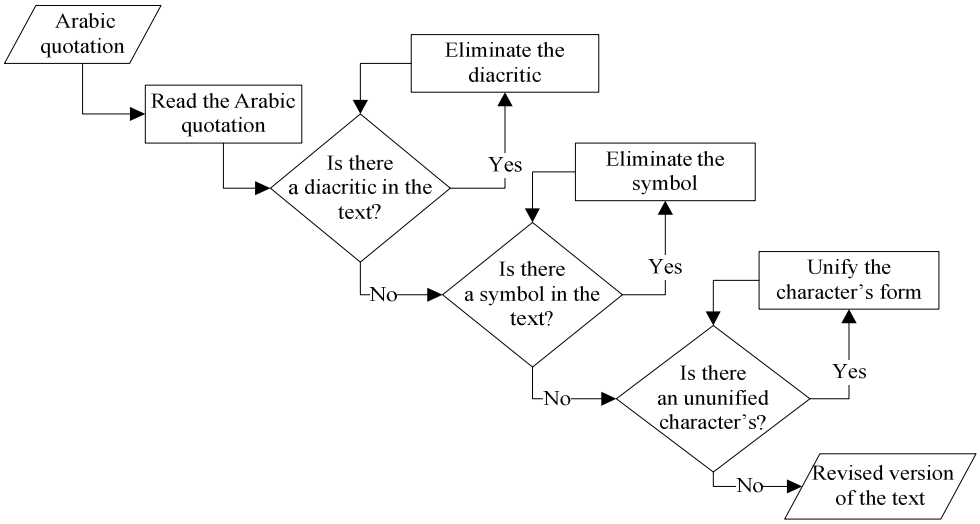

Figure 3.7 Data Receiving and Revising Flowchart Model.

To overcome these difficulties, the QVT system uses the receiving and revising stage, which aids to overcome these obstacles as much as possible (see Figure 3.7). In this stage, the system will review and revise the quotation entered by the user. As a result, the next stage will receive a revised version of the text, free from any recognized diacritics and symbols. Likewise, some of the characters that have no problems with form unification will be unified. To accomplish this stage, the entered quotation must go through three phases: (I) Eliminate Arabic diacritics, (II) Eliminate exclusive symbols, and (III) Unify the character's format.

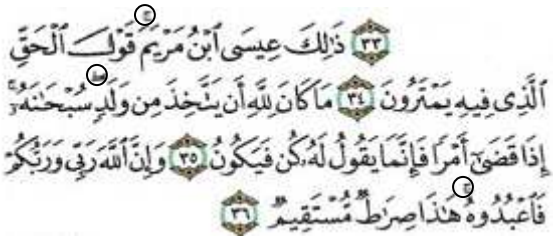

Figure 3.8 Example illustrates Quranic verses structure as well as diacritics and exclusive symbols placement.

Phase I. Eliminate Arabic Diacritics:

The Arabic language is rich in diacritics, which are used to recognize the vocal pronunciation. Arabic's diacritics consist of eight main characters that are shown in Table 3.1[47]. The majority of Arabic books apply those diacritics for each letter, which is what distinguishes the Arabic language from all others. Arabic books such as religious books and poetry texts rely on the use of Arabic's diacritics to distinguish between similar words (see Figure 3.8). Although the use of these diacritics enables the reader to distinguish between the meanings of the words, eliminating them whilst using search engines positively affects the search results. As stated earlier, these diacritics represent an obstacle to retrieve data and information from Arabic books, especially for traditional search engines. For that reason, the first phase aims to eliminate any recognized diacritics (if any) from the entered text.

Fett'ha	◌َ	Shed'dah	◌ّ
Thumm'ah	◌ُ	Tenween Fett'ha	◌ً
Kas'rah	◌ِ	Tenween Thumm'ah	◌ٌ
Su'koon	◌ْ	Tenween Kas'rah	◌ٍ

Table 3.1 Example Illustrates the Arabic Main Diacritics.

Phase II. Eliminate Exclusive Symbols:

Religious books such as the Quran and Hadith's books have their own unique symbols which distinguish them from other Arabic books and references. These symbols are used to facilitate understanding and reciting the text as they may use distinct personal or verbal semantics. For instance, they include signs related to where readers should stop, where they must continue, the possibility of stopping, the possibility of continuing and so on. Furthermore, they may include signs related to contestant's phrases such as (prayer upon the prophets or their companions). Table 3.2 shows an example of some of these symbols; as previously mentioned, these symbols limit the ability of traditional search engines to present accurate results. Therefore, the second phase aims to eliminate any recognized exclusive symbols (if any).

Continuing is better	◌ۖ	Stopping is better	◌ۗ
Must stop	◌ۘ	Must continue	◌ۙ
Peace be upon him	ﷺ	May God Be Pleased With Him	رضي الله عنه

Table 3.2 Some Examples of Exclusive Symbols.

Phase III. Unifying the Character's Format:

Arabic contains some characters that may be written in various forms. Although there are characters which are unable to accommodate particular forms of unification, the Arabic letter " أ " (Alif) is the only Arabic character that may be adjusted. Alif is the first character in the Arabic language that is written in diverse forms with Hamza "ء" and Mu'adah "~". The letter may be written in four different forms as follows: (آ, ا , إ , أ). Although each form has a different vocal sound, the need to reduce the search entering error and increase accurate results becomes necessary. The differentiation in the drawing of the particular character can limit traditional search engine capabilities to provide an accurate outcome. Hence, the third phase aims to unify the character drawing forms of the letter Alif to an abstract character " ا ".

3.1.3.2 Information Retrieval

The information retrieval stage is responsible for accomplishing the full-text-search technique in the original reference's database. For this purpose, the SOA approach will be used to combine phrase-based and regular expression patterns to develop the QVT system (see Figure 3.9). Since the system applies the SOA development methodology, the libraries in this stage will establish the connection between the SL layer and the SA layer. The system will identify the required function in order to retrieve the data. Thereupon, the operational class will be identified to utilize its function. Besides this, the system and class libraries will be consumed in order to run the system's functions. The information retrieval stage manages the logical function that identifies and applies the verification mechanism which consists of two phases: (I) quotation genuineness validation and (II) the nearest similar script retrieval.

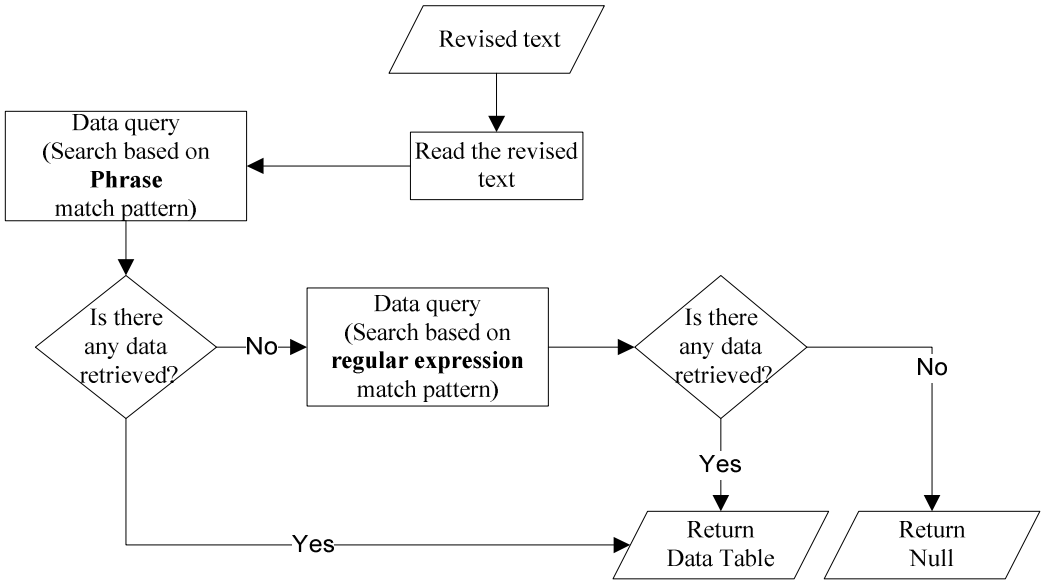

Figure 3.9 Information Retrieving Flowchart Model.

Phase I. Quotation Genuineness Validation:

In this phase, the system confirms the whether or not the text exists in the original reference's database without an alteration. The QVT system considers a complete unaltered quotation as a genuine quotation. The definition of a complete unaltered quotation is different from one reference to another. For example, a complete unaltered quotation in the Quran would be a Quranic verse. A Quranic verse is a verse that is literally independent from the previous verse and following verse. These verses are separated with a numeric interval that determines the beginning and the end of the given verse (see Figure 3.8). Similarly, a complete unaltered quotation in Hadith's books is a one complete Hadith and so on. Therefore, the validation of Arabic quotations is based on the original reference's interval structure. In order to explain this thesis' objectives and results, the Quran will be considered as an original reference case study for the remaining of the thesis.

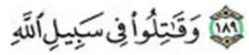

Figure 3.10 Example illustrates an incomplete Quranic verse.

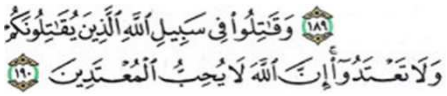

Figure 3.11 Example illustrates a complete Quranic verse.

The phrase-based matching pattern is the approach that is used to search for a complete unaltered quotation. Matching the complete sentence with each verse in the database is the phase procedure that accomplishes this search. If a match sentence is located in the database, the data will be returned. Alternatively, the quotation will be taken to the next phase in the case of the match sentence being undiscovered.

Phase II. The Nearest Similar Script Retrieval:

In this phase, since a match sentence is undetected in the first phase, QVT will conduct an extensive search in order to detect the nearest similar verse to the entered quotation. The nearest similar verse is a Quranic verse that has the slightest difference in comparison to the entered quotation. The QVT system exploits the regular expression characters to develop the process query. In essence, the regular expression pattern achieves a more accurate data search query. A regular expression manipulates powerful meta-characters; these characters will be handled in a manner to indicate retrieval data based on predetermined patterns with high precision. In the case of a similar sentence being obtained, a table including a list of similar verses will be returned. Otherwise, the quotation will be considered distorted.

In order to retrieve the data from the SA layer, a data reader will be used. A data reader is an object in the ADO.NET library to retrieve a read-only stream of data from a database. Outcomes are returned as the query executes that can be customized based on the program display scheme. The next stage is responsible for mapping and formatting the data in a manner which facilitates the reading, transmitting, and exchanging of the data.

3.1.3.3 Data Mapping and Categorizing

The Data mapping and categorizing stage is responsible for mapping data to assist the information exchange over the WS layer. Data mapping is the process of generating data elements to harmonize data between two different data patterns. The mapping process is one of the most important steps out of the data integration tasks. The data table technique is employed to accomplish the mapping process. A data table is a central object in the ADO.NET library that displays information based on a schema pattern. A data table is used to represent a data set. The data set will include the retrieved data for the similar verse such as full text, chapter name, and verse number. This information will be displayed in predefined data columns that facilitate the data mapping process.

3.1.4 Source Access (SA) Layer

The Source Access (SA) layer is a set of classes and functions for reading from and writing in a database. It hosts the QVT communications' function and the SQL queries. The SA layer provides the services necessary for communicating and retrieving the data from the DA layer. It generally organizes the relationship between the QVT and the database. Besides, it encodes the SQL queries that are responsible for retrieving the information from original references' database. Lastly, it assumes the responsibility of starting the connection to the cloud database.

The QVT system uses the SQL language in order to implement the information retrieving process from the SL layer. The SQL keywords such as "LIKE" and "REGEXP"

are used to implement the phrase-based matching and regular expression patterns respectively. Both keywords will be associated with the "WHERE" clause to implement the SQL queries. The search functions “LIKE” executes a search process based on a “per-character” match technique. Therefore, it may generate more accurate results than other possible comparison patterns [39], [48]. Figure 3.12 shows a portion of the SQL code that retrieves information from the Quran database with one condition. The condition is that they retrieve the information only when a match detected.

```
1   SELECT AyahId, SurahName , AyahText, SurahId
2   FROM quran
3   WHERE  AyahText
4   LIKE '  Revised Quotation '
```

Figure 3.12 Phrase-based Pattern SQL Code.

The search function "REGEXP” executes a search process based on a predefined pattern search approach. The regular expressions powerful meta-characters may generate more accurate results. Figure 3.13 shows a portion of SQL code that retrieves information from the Quran database with several conditions. To retrieve the information, at least one of these conditions must be met. In the case of achieving more than one condition, the data will be retrieved as well. Meta-characters such as “[[:<:]] [[:>:]]” and “.” are all powerful characters that are implemented to overcome Arabic language obstacles and search difficulties. “[[:<:]] [[:>:]]” character is stand as text boundaries. It matches the beginning and end of the revised quotation[48]. Each dot represents one digital character “8 bits." Each Arabic character represents two digital characters “16 bits." Since supporting the Arabic language is the objective of the QVT, the system will use two dots “..” in order to represent one Arabic character. As previously mentioned, the Arabic language is characterized by using pronouns as suffixes and prepositions, and conjunctions as prefixes [41]. The two dots “..” character can be used to overcome the suffix and prefix difficulties.

```
SELECT AyahId, SurahName , AyahText, SurahId
    FROM quran
    WHERE
AyahText REGEXP '[[:<:]]Revised Quotation [[:>:]]'
OR AyahText REGEXP '[[:<:]]..Revised Quotation [[:>:]]'
OR AyahText REGEXP '[[:<:]]....Revised Quotation [[:>:]]'
OR AyahText REGEXP '[[:<:]]Revised Quotation..[[:>:]]'
OR AyahText REGEXP '[[:<:]]Revised Quotation....[[:>:]]'
OR AyahText REGEXP '[[:<:]]Revised Quotation......[[:>:]]'
OR AyahText REGEXP '[[:<:]]..Revised Quotation....[[:>:]]'
```

Figure 3.13 Regular Expression Pattern SQL Code.

3.1.5 Data Source (DS) Layer

The Data Source (DS) layer comprises of the original reference (RDB). A collection of genuine texts for the most famous Arabic books such as the Quran, Saheah-Albukari, 1001 nights, Panchatantra, etc. are stored in the DS layer. Original reference (RDB) is a collection of schemas established in formally defined tables. Each schema represents an Arabic book's genuine script and its translations. Original reference RDB will be developed using the MySQL database management system. MySQL is an open-source database management system that offers a useful meaning to manage the original reference RDB.

Original reference RDB runs in a MySQL cloud database server. The cloud server provides remote multi-user access to the reference RDB. Besides, it facilitates the services' delivery in real time over the Internet. It also offers efficient attributes to handle the data in a distributed, ubiquitous and global approach [37]. Using the cloud service will overcome the problem of duplicating the original reference DB in each device that aims to run the QVT system. It will reduce the application storage capacity when installed on the device to give the user the advantage of using them along with other applications. It gives the QVT system the chance to expand and offers a global accessibility to the original reference DB without having been copied or moved from place to another. The dynamic features offered by the cloud provider allow for fixable development expansion. As a result; infrastructure range, maintenance work, software license, and management control will all be reduced.

3.2 QVT Software Design

3.2.1 Structure Diagrams

Structure diagrams emphasize the system modulation plan. The package diagram, the component diagram, and the class diagram are the structure diagrams that will be illustrated in order to identify the systems modulation organization.

3.2.1.1 Package Diagram

QVT implementation is based on linking the web services with the cloud database. The low-level communication functions will be handled using C#. The Microsoft.Net library will be utilized for the User Interface (UI). Furthermore, Microsoft proposes the ASP.NET library which facilitates the web service's development.

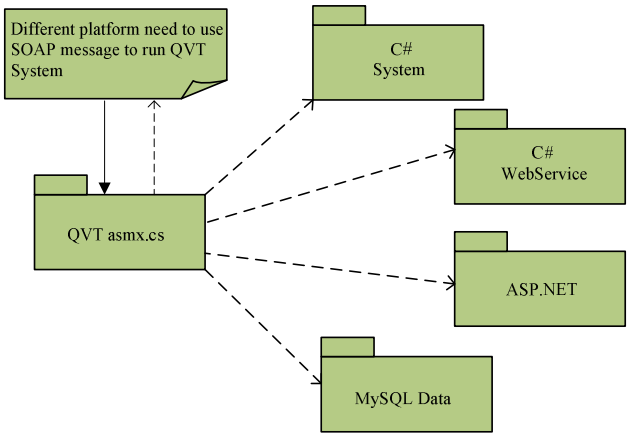

Figure 3.14 QVT Package Diagram

ASP.NET library offers components which allow the web service to be called from different application’s platforms. Since the cloud database will be developed using the MySQL database management system, the database MySQL client library will be installed and used to establish the correlation between the web service and the cloud database.

The QVT asmx.cs file is the main package of the system. It needs to run the main C# standard libraries including the C# system library, Web Services library, ASP.NET library, and the MySQL data library as shown in Figure 3.14. The QVT web service features will be implemented using the ASP.NET package. The design of the QVT depends on the separation between the user interface and the web service, while maintaining the web service and the database access file combined. Thus, the system accessibility and compatibility will be maintained. Along with this, programming code integrity and clarity will be increased.

3.2.1.2 Component Diagram

The QVT system utilized multiple components. The important components according to the system operational processes are the user interface components, the QVT service's component, and the database component. However, the main components have to communicate with each other. To accomplish this task, secondary components facilitate this communication and the information exchange.

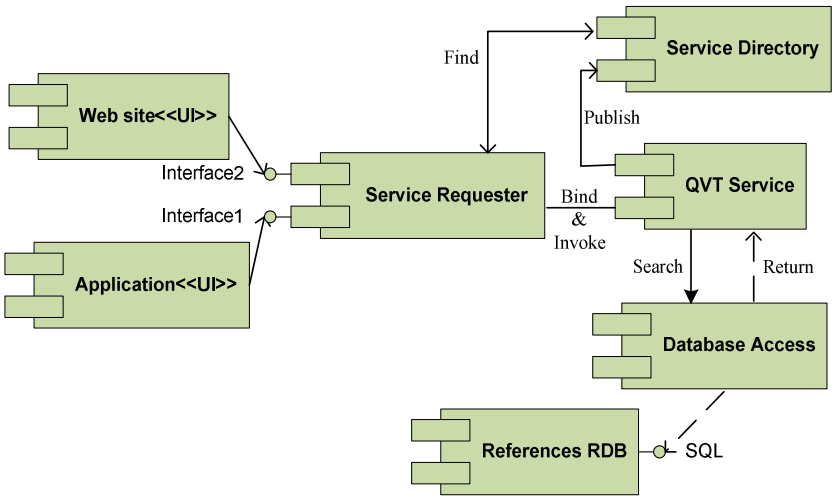

Figure 3.15 QVT Components Diagram

The QVT service component is the main component of the system. It needs to establish communication with the other components such as the user interface component and the database component as shown in Figure 3.15. Having said that, we must consider the task of the secondary components; these facilitate the integration of other components. For example, the database access component contains the SQL queries as well the cloud database server's address. This component will be used to aid the QVT service in order to access the database quickly and easily.

3.2.1.3 Class Diagram

The QVT system used C# functional classes. The important classes according to their functionalities are application layer class diagram, the QVT service class diagram, and data access class diagram. Each class diagram will be briefly described.

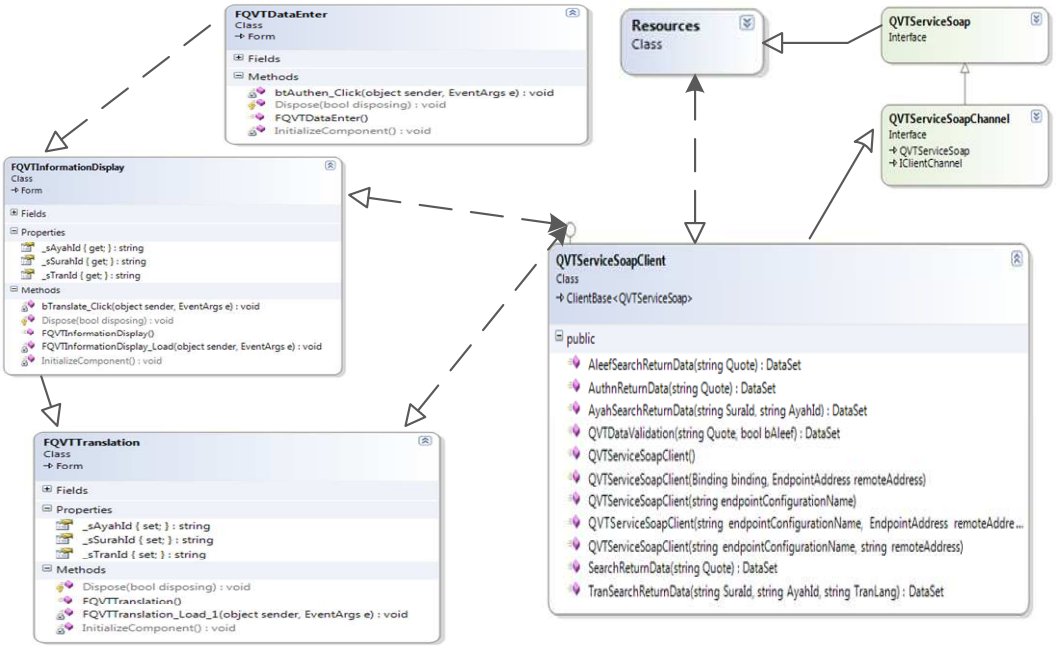

Figure 3.16 QVT Windows Interface Class Diagram.

Figure 3.16 shows the application layer class diagrams of the QVT windows application. The main class is the *QVTServiceSOAPClient* which represents the

application client that creates the session and exchanges the information between the user interface and the web service. The *FQVTDataEnter* class represents the main DOT.NET interface window for the user to enter the quotation wished to be validated. The *FQVTInformationDisplay* class represents the interface window that displays the service outputs. It will receive the data from the *FQVTDataEnter* and run the QVT web service to retrieve the information and then display the results. The *FQVTTranslation* class represents the interface window that shows the quote's translation. Like the prior class, it will receive the data from the *FQVTInformationDisplay* and run the QVT web service in order to retrieve the translation information and show the results. The *Recourses* class is used to bind the windows application with the web service. The *QVTServiceSOAPChannel* class will be exploited by the *QVTServiceSOAP* class to establish the connection to and from the QVT web service.

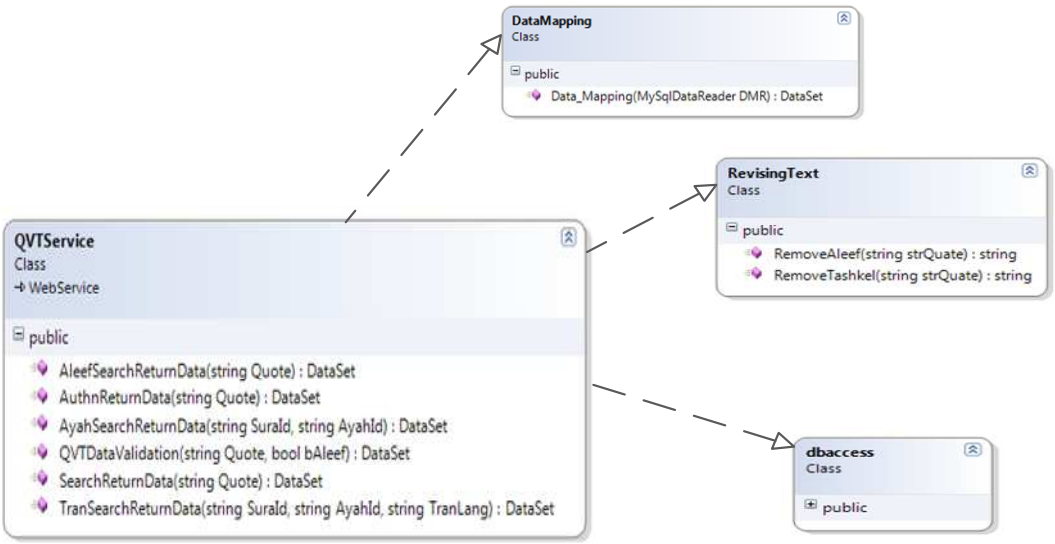

Figure 3.17 QVT Service Class Diagram

Figure 3.17 demonstrates the QVT service class diagrams. The QVT enables the information retrieving process as well as the quotation revision process. If a user requests a data search within the database; these two processes will be started. The *QVTServices* is the main system class that handles the search function on the SL layer. Additionally, it

manages the connection with the *DBaccess* class. The *RevisingText* is the responsible class which applies the revision procedure to the entered quotation. *DataMapping* is the responsible class which manages the data mapping process on the retrieved information from the database. The Class's component will accomplish the services logical search procedure as mentioned earlier.

Figure 3.18 shows the data access class diagrams. The main class in this diagram is *DBAccess*. Its primary role is to generate the SQL query command to retrieve the necessary information from the database. The SA Layer uses the *QueryExecuter* class as its server connection executer. The *QueryExecuter* class comprises the server address for the database cloud service. It is responsible for establishing the data connection, executing the SQL query, and terminating the data connection after it's completed.

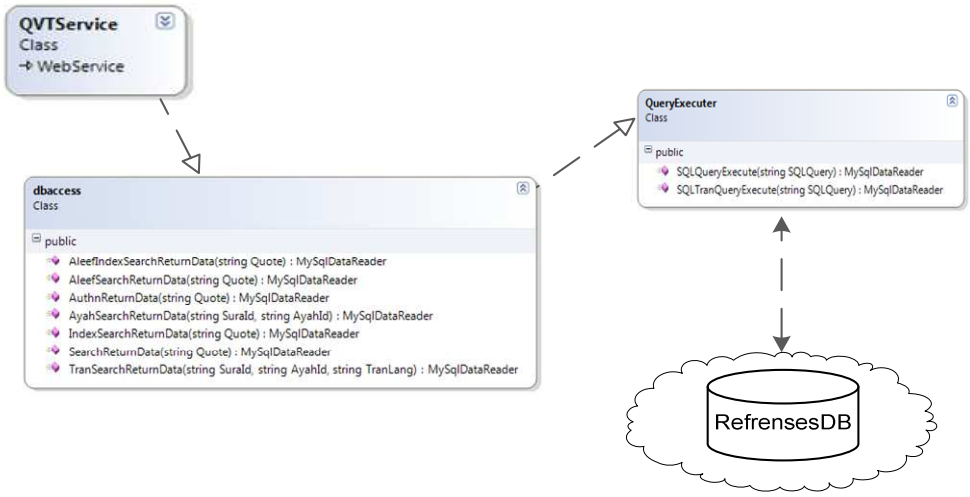

Figure 3.18 Data Access Class Diagram

3.2.2 Behavior Diagrams

Behavior diagrams emphasize the systems activity conducting plan. The activity diagram is a behavior diagram that will be used to describe the systems functionality events.

3.2.2.1 Activity Diagrams

Figure 3.19 provides a high level view of the major activity provided by the QVT system. There are three main activities: (1) request of events, (2) transmitting of events and (3) fulfillment of events. Request events contain the necessary actions to request a verification service from the AI layer. Transmitting events contain the necessary actions to transfer the SOAP message to and from the SL layer over the Internet. Fulfillment events contain the essential logical function to accomplish a text search process in the SL layer. The QVT verification system offers the following operational functions:

i. Acquisition of the Arabic quotation;
ii. Generate a SOAP message;
iii. Integrate the Arabic quote into the SOAP message;
iv. Transmit the SOAP message, including the quotation over the Internet;
v. Receive the SOAP message including the quotation over the Internet;
vi. Extract the Arabic quotation from the SOAP message;
vii. Convert quotation to a string in order to facilitate the system reorganization;
viii. Revise the quotation string to simplify the data search;
ix. Complete the data search based on a revised quotation string;
x. Retrieve the information; and
xi. Map the data.

Figure 3.20 shows the quotation revision activity which is an essential step in order to smooth the information retrieval in the QVT system. The process of revision applies to any quotation in order to overcome the language obstacles.

As exposed in the diagram, the system reads the Arabic quotation as a string based on a DOT.NET configuration library. The system will then store the quote string in a primary string parameter. Afterward, it will request the revision function to start the revising loop process which filters the quotation from the characters that cause the search difficulties. The revision function will obtain the string parameter, including the Arabic

quotation script. In sum, the revision function starts the revising loop in order to process the quote string.

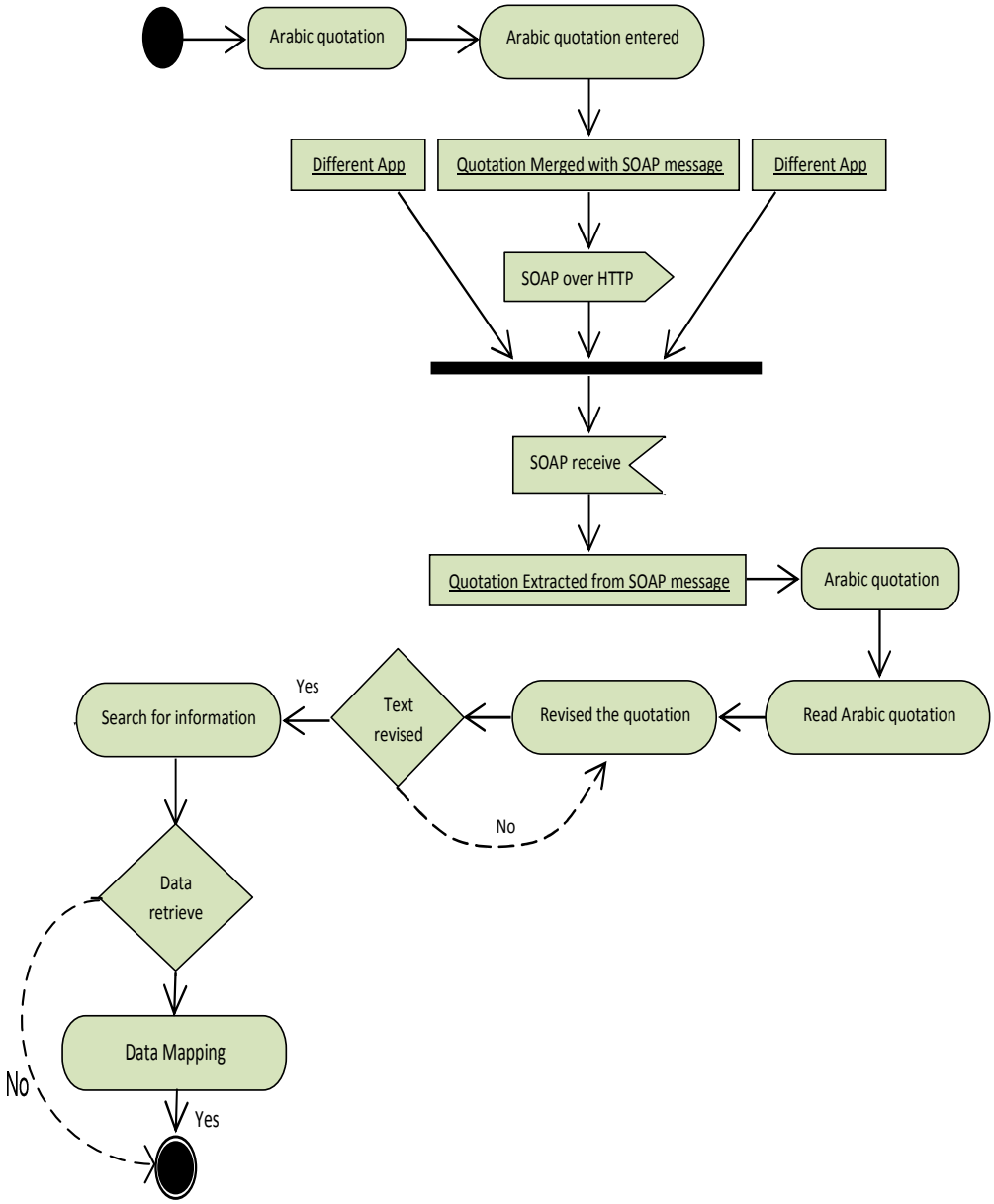

Figure 3.19 Core Activity Diagram

The revision loop process contains three phases, and each phase will remove its characters group. An important feature of this process is that each phase will keep

revising the quotation text until all of the predefined characters are removed from the quote string.

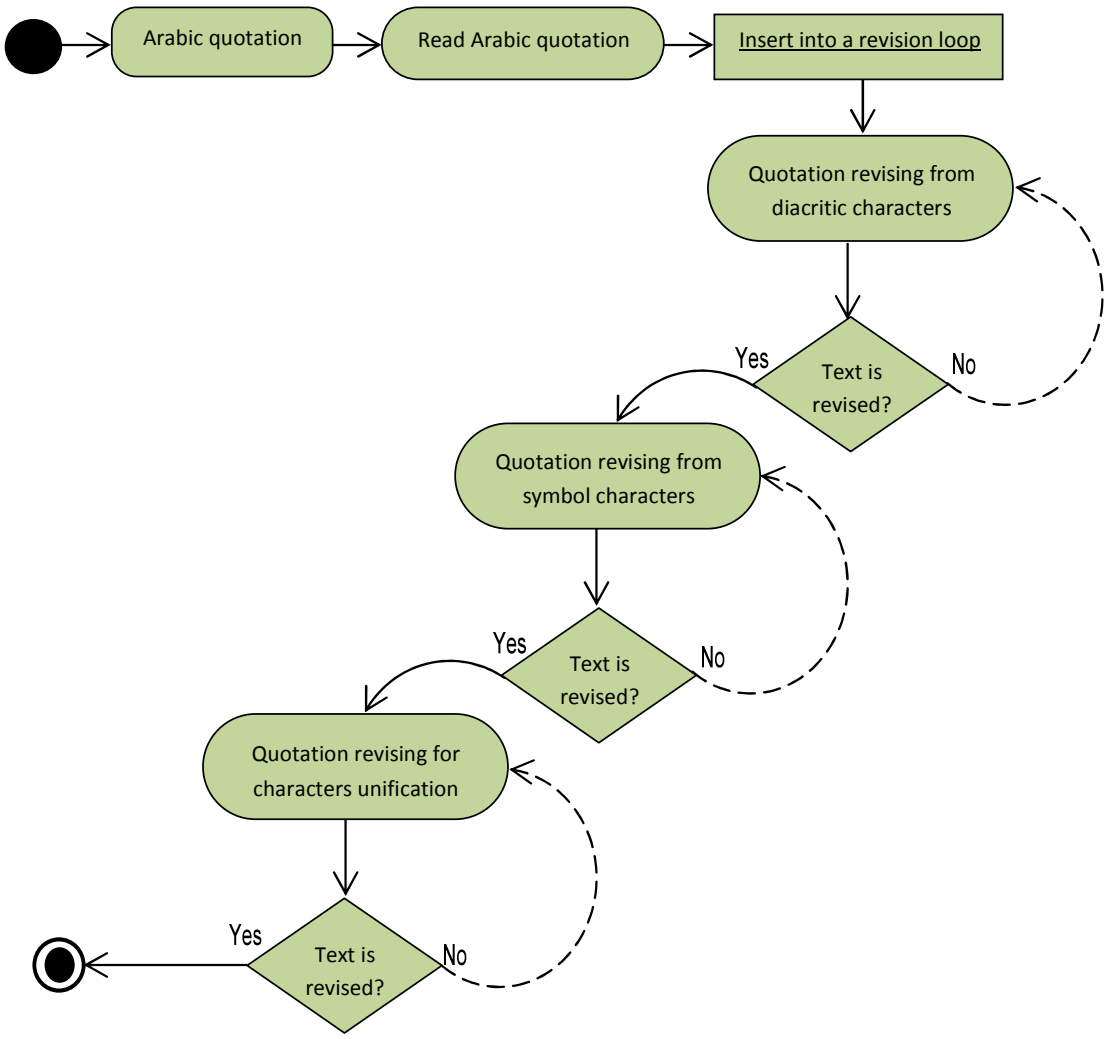

Figure 3.20 Activity Diagram for the Quotation Revising

The QVT revision approaches rely on an advanced character revising algorithm. Figure 3.21 presents a pseudo code algorithm which clarifies the essential steps that are considered in sequence to revise quotation strings when a user requests a search service. The algorithm output will then be taken to the next step in the QVT system which is the information retrieval query.

The Input:
Entered Arabic quotation
The Output:
The revised quotation text

```
1   Initialize an empty number parameter names it Letters_Counter and sets it at zero.
2   Initialize an empty string parameter names it Quote_Script.
3   Read the sentence.
4   Set the read sentence in Quote_Script parameter
5
6   Count the letters in Quote_Script.
7   Set the letters' count in Letters_Counter
8
9   IF the Letters_Counter is greater than Zero.
10      FOR each letter in the sentence.
11          IF the letter contains a diacritic character.
12              Remove the diacritic character.
13            ENDIF
14          IF the letter contains a symbol character.
15              Remove the symbol character.
16            ENDIF
17          IF the letter is equal " آ " or " أ " or " إ "
18              Replace the letter with " ا ".
19            ENDIF
20          Subtract one from Letters_Counter.
21      END FOR
22  ELSE
23    RETURN the revised string.
```

Figure 3.21 QVT Revising Pseudo Code

3.2.3 Interaction Diagrams

Interaction diagrams emphasize the control streams among the system component. Interaction is a division of the behavior diagrams that illustrates the systems workflow in terms of function, control and data among the systems structure. The sequence diagram is an interaction diagram that will be used to describe the systems workflow.

3.2.3.1 Sequence Diagrams

In general, all applications perform the same following sequence to initialize the connection between the application GUI and the web service at the beginning:

- The web service registers to the UDDI directory;
- The application looks for a web service at the UDDI directory;
- The UDDI discloses the web service WSDL document;
- The application requests to bind a web service based on the WSDL information;
- The web service confirms the binding;
- The application invokes the web service; and
- The web service retrieves the information.

The interactions between the application and the web service system are demonstrated by the sequence diagram in Figure 3.22. In essence, the web service starts the connection sequence diagram. This is done by registering the WSDL document information in the UDDI directory. Later, any application can look for the wanted web service to invoke it and connect to it. After the connection is established, the application may send a request enclosed in a SOAP message to the web service in order to retrieve the information.

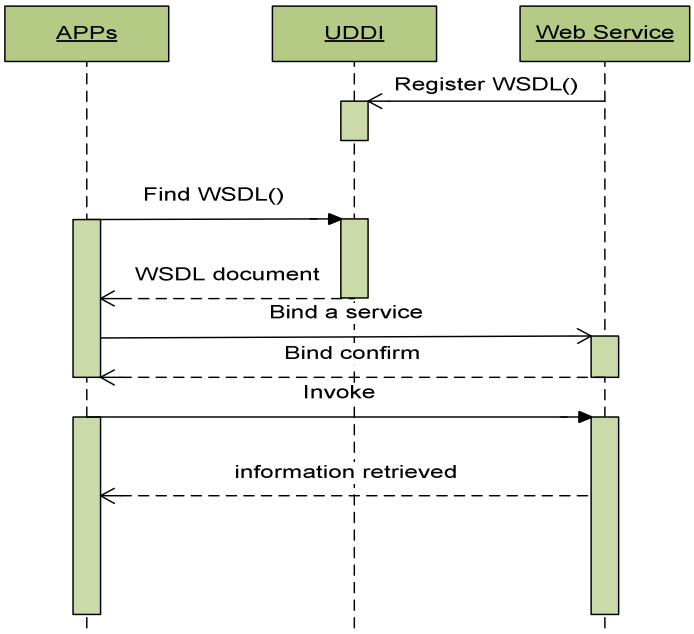

Figure 3.22 Web Service Connection Sequence Diagram

The sequence diagram of quotation validation interactions that occurs between the user and the QVT system are represented by Figure 3.23. The “application interface” component in this diagram is a theoretical class representing all of the application platforms that connect to the systems web service. The quotation validation scenario will generate the following sequences:

I. User request to validate a quotation via the application interface;
II. The application interface receives the request and then resends it using a SOAP message to the QVT system through a pre-established connection;
III. The QVT system revises the quotation and generates the SQL query;
IV. The QVT system connects to the references DB and applies the SQL query;
V. The references DB returns the information (if any) to the QVT system; and
VI. The QVT system resends the retrieved information to the application interface using a SOAP message. Then, the application interface in turn displays the results to the end user.

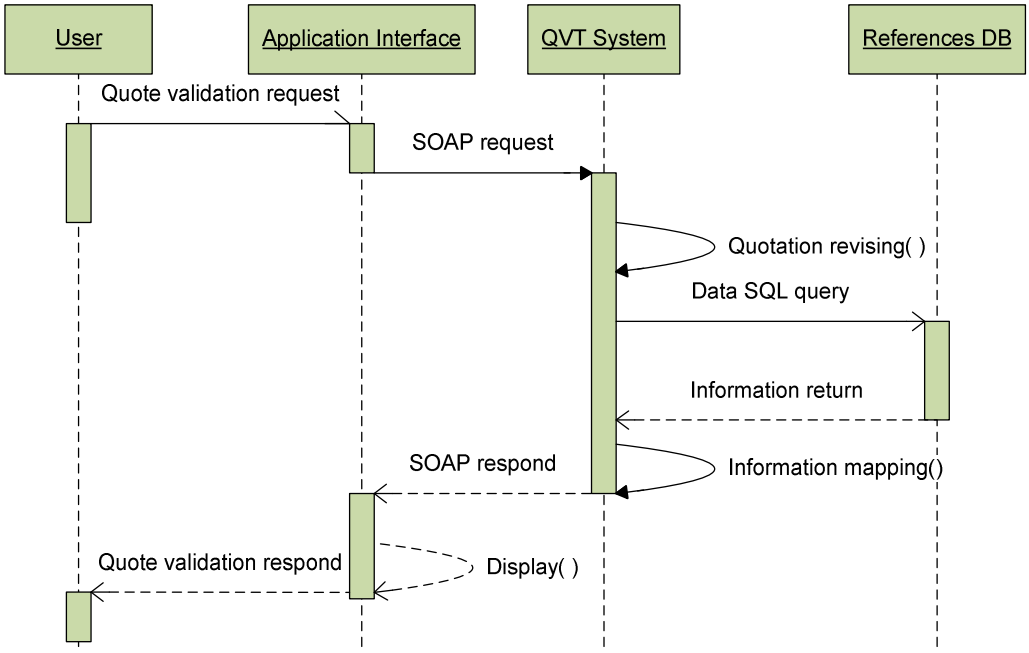

Figure 3.23 Quotation Validation Sequence Diagram

Chapter 4 : Implementation

The QVT system approach is based on the concept of the web service. As described in chapter 3, the QVT approach facilitates the systems integration to various application platforms. As well, the use of cloud computing will support the services real time delivery in order to enhance the service accessibility. The following sections provide the implementation details of the QVT web service and the Graphical User Interface (GUI). Continually, the implementation of the Database-as-a-Service (DBaaS) will be illustrated by taking advantage of a Xeround [49] cloud database service.

4.1 QVT Web Service

The QVT web service was implemented with the Visual Studio C Sharp project based ASP.NET technology. It extends all of the properties of the web service library. As mentioned earlier, the web service properties will provide developers with different platforms and the capability to discover the service and bind it to their applications. If the default GUI is the website, then the QVT web service will be integrated to the website services as a reference service. By the same token, any fundamental platform can bind the QVT web service. It supports website platforms, windows applications, mobile applications, web browsers, and other platforms.

As shown in Figure 4.1, the QVT web service provides multiple search functions which can be utilized based on the developer application's design. Besides this, it provides descriptions for each function. However, this interface is usually hidden for the end user. In the following section, the QVT web service features will be reviewed in brief. The tag #1 shows the web services name, while tag # 2 refers to the WSDL service description link where it can be imported automatically as will be demonstrated later. Lastly, Tag # 3 indicates the function name whereas tag # 4 describes the functions objective.

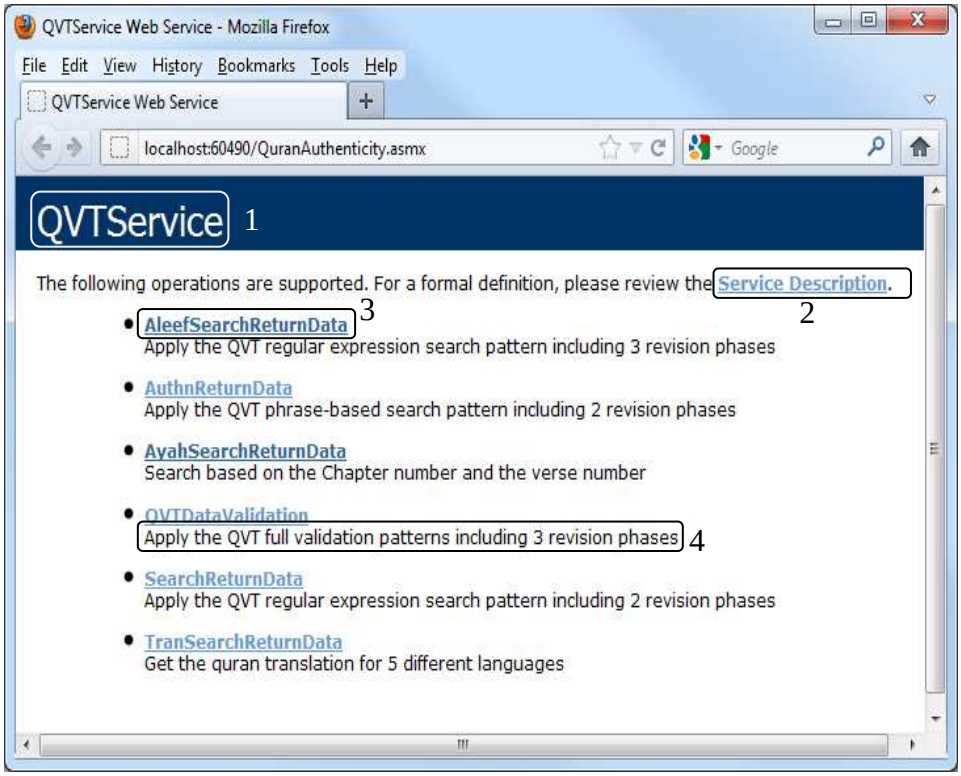

Figure 4.1 QVT Web Service Interface

At present, the QVT web service provides the following search functions:

- **QVT data validation**: Apply the QVT full validation patterns, including three revision phases.
- **Authentication function**: Apply the QVT phrase-based search pattern, including two revision phases.
- **Search function**: Apply the QVT regular expression search pattern, including two revision phases.
- **Search function with Alif unification**: Apply the QVT regular expression search pattern, including three revision phases.
- **Search function based on verse & chapter number**: Search based on the chapter number and the verse number.
- **Search function for quotation translation**: Get the Quran translation for five different languages.

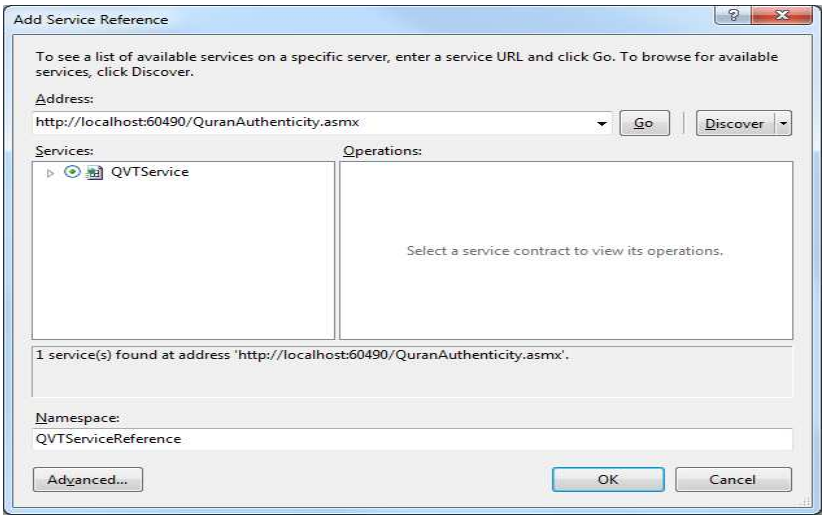

Figure 4.2 QVT Web Service Automatic Binding.

As shown in Figure 4.2, the QVT web service can be integrated automatically by the system developing software through the web service address. The application's developer can add the web service as a service reference which facilitates functions invoking which are offered by the web service. Figure 4.3 represents a screenshot for one of the QVT web service factions. As displayed, screenshots demonstrate the factions required parameters as well as the SOAP request and response message attributes.

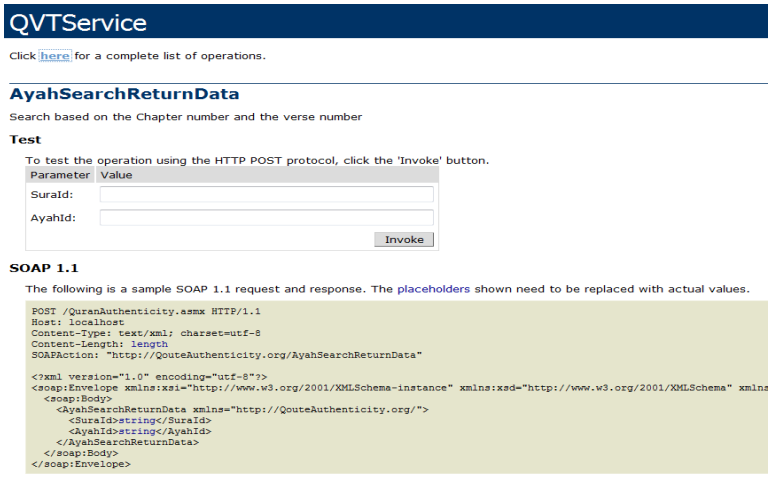

Figure 4.3 Screenshot for Ayah Search Function.

4.2 Graphical User Interface (GUI)

Using the QVT web service shown in figure 4.1, developers can integrate it into multiple platform application interfaces. Web service features may be inherited to any application platform. Additionally, the software's developer may develop a multiple GUI from the same web service. In the following screenshot, the implementations of two different interfaces are demonstrated.

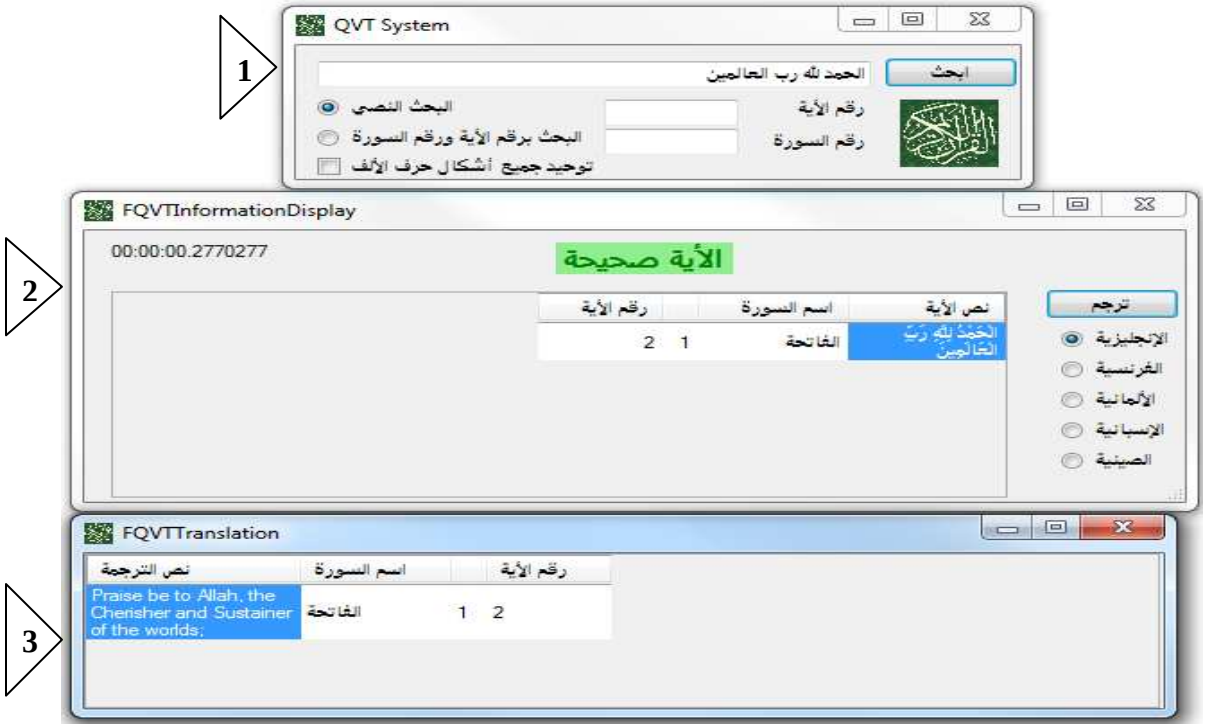

Figure 4.4 Screenshot for Windows GUI

4.2.1 Windows GUI

Figure 4.4 shows a screenshot for a windows GUI of the system. As displayed, window GUI consists of three windows in the Arabic language. Each window completes the tasks assigned to it; the first window is obtaining information from the user and analyses it. The results are then displayed in the second window; the second windows

main task is to show the output results. It also performs the translating task for the chosen text to one of the following languages: English, French, German, Spanish or Chinese. Lastly, the third windows main concern is to demonstrate the verse translation based on the language selected by the user.

4.2.2 Web GUI

Figure 4.5 shows a screenshot for a web GUI of the system. As displayed, the web GUI consists of the system's tasks in one window with the Arabic language. The systems tasks can be completed within the same website browser. The main task is obtaining information from the user and analyzing it. Then, the results will be displayed on the data table with a select option. The second data table main task is to demonstrate the verse translation based on the language selected by the user. The translating task is accomplished in following languages: English, French, German, Spanish or Chinese.

Figure 4.5 Screenshot for Web GUI

4.3 DBaaS Cloud Service

Database-as-a-Service (DBaaS) is a cloud computing service that offers database storage and database management services for developers over cloud. Service providers such as Xeround take care of providing these services which increase high accessibility for the web service database. Initial service installation, auto scaling, backup's scheduling and many more functions are all services that are offered by the DBaaS service provider.

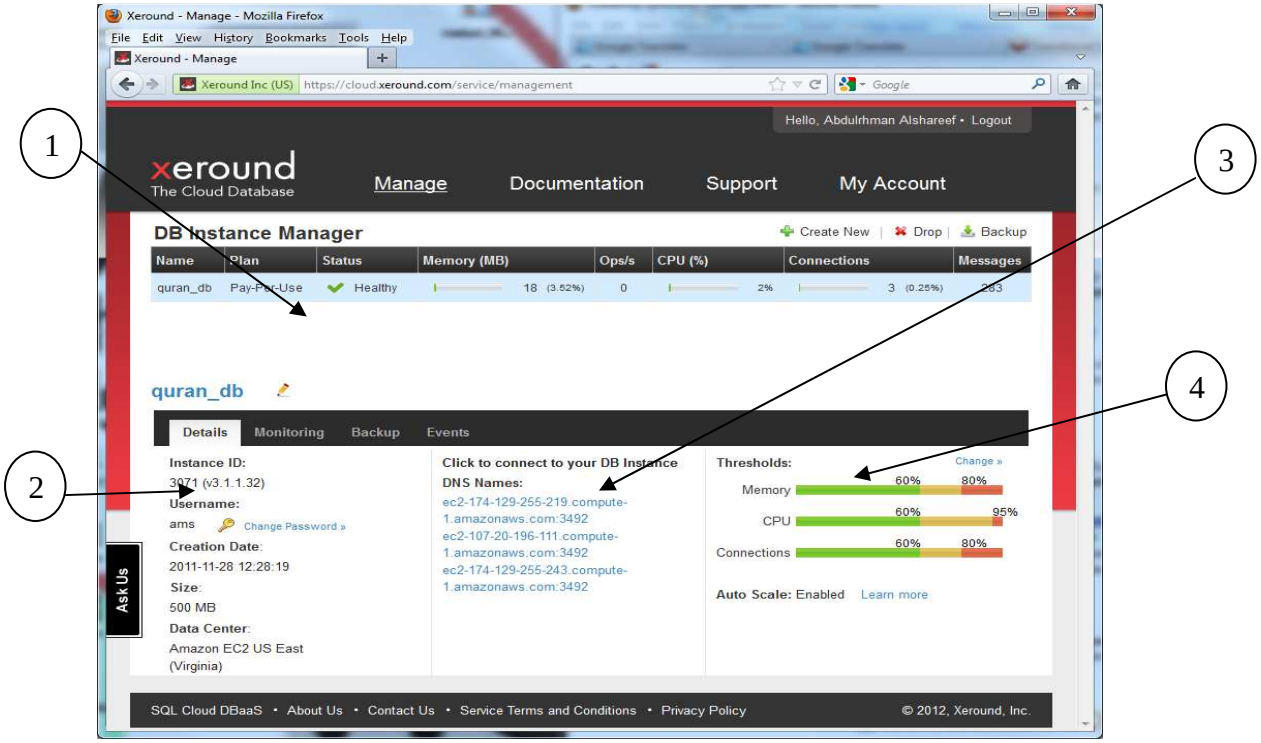

Figure 4.6 Screenshot of the DBaaS by Xeround

Figure 4.6 illustrates a screenshot of the DBaaS by Xeround. Xeround's main account page provides access to multiple functions that are offered by the DBaaS service. As demonstrated, Tag 1 indicates to the database management zone, which contains the active database information in the users account. Besides this, it also offers services to manage databases such as creating a new database, deleting a database, or backing up an existing database. Such services assist the developers to expand their databases or

increase the data capacity which facilitates the fulfillment of the end-user requests. Tag 2 indicates the account management zone, which contains the user and account information. Tag 3 indicates the database connect information zone, which contains the database DNS names' information. It offers access to the database information and table in order to manage the database content (see Figure 4.7). Tag 4 indicates the database status indicators zone, which contains visual indicators that monitor the database status and the server status. Although cloud DBaaS offers various services that are not listed formerly, this thesis has reviewed the major functions.

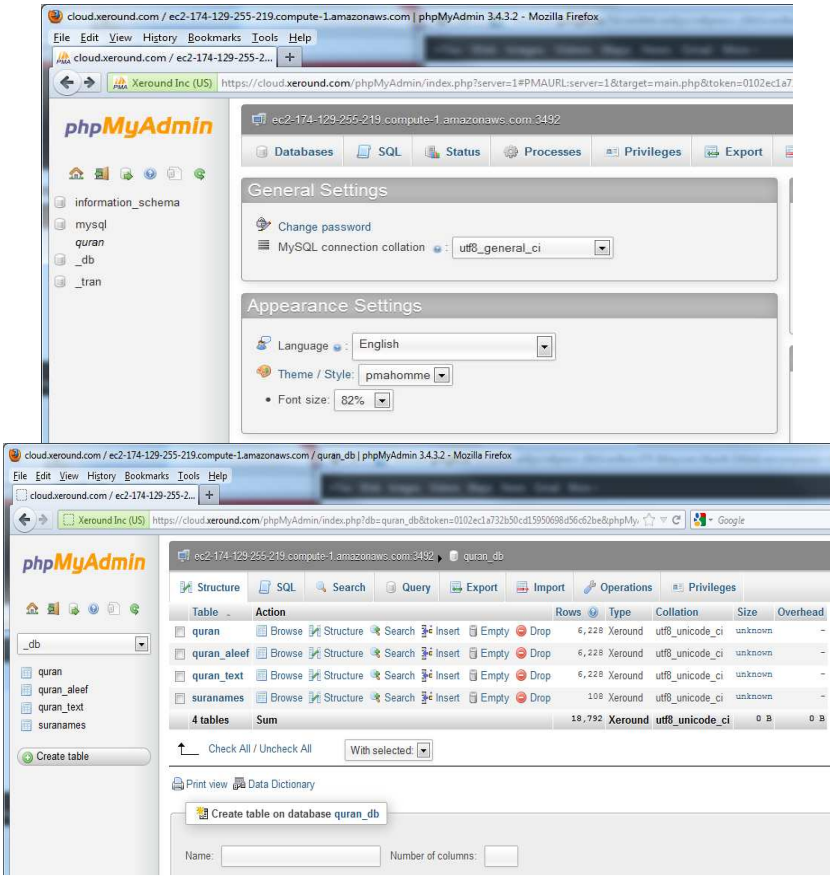

Figure 4.7 DBaaS Database Content Management Zone

Chapter 5 : Evaluation and Results

To evaluate the performance of the proposed system, we tested QVT with different evaluation phases. The evaluation process will accomplish two assessment phases: a validity assessment and a comparison assessment. In the following section, the characteristics of each phase will be we described.

5.1 QVT Validity Assessment

In this phase, the accuracy of the phrase based search algorithm will be examined. The Arabic quotation validation is confirmed, when the quotation is discovered in the original reference database based on the proposed algorithm. A cross-sectional validity assessment will be conducted in order to emphasize the effectiveness of the proposed algorithm. The objective of this experiment is to measure the impact of the proposed validation mechanism on the accuracy of Arabic text retrieval systems.

A probability unrestricted sample will be used in the validity assessment. The probability unrestricted sample is a random sample that was selected from the study target groups, which are texts from given Arabic books. However, due to circumstances of obtaining the samples, the sample will be limited to the available study group. The available study group is a set of random samples that have been identified from the Quran, a reference that is subject to be examined.

This assessment will help to determine the accuracy of the proposed system in order to identify the correct and incorrect data. In addition, the program will aid in verifying the systems precision when it is used with diverse samples. The independent variable will be the tested quotations, and the dependent variable will be the assessment's output correctness. The assessment will be controlled by the validation's procedure. The dependent variable is the studied situation that is likely to change whenever the

independent variable is distorted. The control variable prevents influences from changing the independent variables impact on the dependent variable.

This assessment will contain six groups of data. The proposed algorithm will be applied to each group, so that it will compare the texts in each group with the available database. Each collection has different characteristics. Subsequently, dataset groups will be reviewed as well their characteristics:

- **First group**: This group contains a dataset of 100 random quotations. Each quotation has been electronically selected at random from the original database, using the MySQL random select statement. Each quotation includes its complete unique diacritics and Quran's symbols.
- **Second group**: This group contains a dataset of 150 random quotations. Each quotation has been electronically selected at random from the original database, using the MySQL random select statement. Each quotation includes its complete unique diacritics and Quran's symbols.
- **Third group**: This group contains a dataset of 200 random quotations. Each quotation has been electronically selected at random from the original database, using the MySQL random select statement. Each quotation includes its complete unique diacritics and Quran's symbols.
- **Fourth group**: This group contains a dataset of 100 random quotations. Each quotation has been electronically selected at random from the original database, using the MySQL random select statement. Each quotation includes its complete unique diacritics and Quran's symbols. The dataset is divided into two clusters:
    - The first cluster represents 70% of the sample, containing the undistorted quotations.
    - The second cluster represents 30% of the sample, containing the distorted quotations.
- **Fifth group**: This group contains a dataset of 100 random quotations. Each quotation has been electronically selected at random from the original database,

using the MySQL random select statement. Each quotation includes its complete unique diacritics and Quran's symbols. The dataset is divided into two clusters:

- The first cluster represents 60% of the sample, containing the undistorted quotations.
- The second cluster represents 40% of the sample, containing the distorted quotations.

- **Sixth group**: This group contains a dataset of 100 random quotations. Each quotation has been electronically selected at random from the original database, using the MySQL random select statement. Each quotation includes its complete unique diacritics and Quran's symbols. The dataset is divided into two clusters:
  - The first cluster represents 55% of the sample, containing the undistorted quotations.
  - The second cluster represents 45% of the sample, containing the distort quotations.

Distorted quotations have been distorted or manipulated manually by adding, deleting or modifying the script. Accordingly, the accuracy of the proposed system will be confirmed when the assessment's output is perfect. It will be considered perfect when the assessment's outputs are identical to the inputs.

5.2 QVT Comparison Assessment

In this phase, a comparison assessment will be developed to evaluate the phrase based and regular expression algorithms. In this experiment, a probability restricted sample will be utilized in the comparison assessment. The probability restricted sample is a sample restricted by predefined specification. This assessment will contain three groups of data. Each collection has different characteristics as will be explained in the following section:

- **First group**: This group contains a dataset of 20 quotations. Quotations were selected from the available study group, which are texts from the Quran. These selected quotations will be restricted by the following specifications:

- Each quotation must contain a complete Quranic verse.
- Each quotation must contain complete diacritic characters.
- Each quotation must contain at least one of Quran's symbols.

- **Second group**: This group contains a dataset of 10 phrases. Phrases were selected from the available study group. These selected phrases will be restricted by the following specifications:
  - Five phrases should consist of two words per phrase.
  - Five phrases should consist of three words or more per phrase.
- **Third group**: This group contains a dataset of 10 words. Words were selected from the available study group. These selected words will be restricted by the following specifications:
  - Five words should consist of three letter roots.
  - Five phrases should consist of more than three letter roots.

These groups will go through a comparison assessment. Each quotation, phrase, and word will be tested with each targeted environment. This assessment compares the retrieved results between the proposed tool QVT and five traditional Quran search engines. Due to the features' similarities with the selected environments, the results of this comparison will be significant in determining the effectiveness of the proposed tool.

To evaluate a text retrieval system, the accuracy of the information retrieved is one of the most essential aspect as it is the major objective of any text retrieval system. Once the accuracy is determined, the next aspect which affects the effectiveness of the information retrieved is the systems precision. The main objective a text retrieval system is to achieve the maximum rate of precision.

To measure the accuracy of the data, the search results will evaluate the correctness of the retrieved results based on the assumption that the outputs must be identical to the inputs. To measure the precision the proposed tool, the search results will be evaluated based on the number of correct retrieved results from the actual accepted results.

5.3 Test Results

5.3.1 Validity Assessment Findings

The information's accuracy is the best measurement technique to guarantee the efficiency of information retrieval systems. In order to measure the information's accuracy with different data query, QVT is applied to a large number of different quotes that are selected randomly. To accomplish this, a simple windows application has been developed. The search results will be evaluated based on the assumption that the outputs must be identical to the inputs. Figure 5.1 shows screenshots for the evaluation test application. As illustrated, the application allows the possibility to choose the group to be tested. Besides this, it demonstrates the number of the correct output quotations, the distorted output quotations, and the total number of the tested quotations. These numbers will be useful for the comparison process. The comparison will take place between the input facts which have been predefined, and the output results which will be obtained from the application after executing the proposed algorithm.

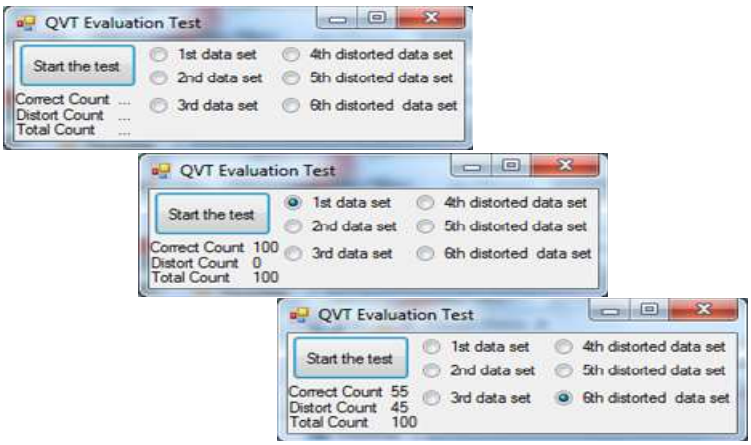

Figure 5.1 Screenshot for the Evaluation Test Application.

Table 5.1 shows the comparison outcomes based on the input facts and the output results. As demonstrated, the table shows the comparison results for six different groups which had their own characteristics. Each row represents a group of dataset. The

quotations input enumeration column represents the actual input data, which is considered as a fact. The quotations output enumeration column represents the output information which is acquired from the evaluations test application and is needed for evaluation (see Figure 4.7). The systems accuracy will be calculated based on the following equation:

$$System\ accuracy\ percentage = \frac{The\ number\ of\ retrieved\ results}{The\ number\ of\ actual\ input} * 100$$

The bare numbers represent the actual number of quotes earthier for the input or output. The numbers placed in brackets within the output enumeration column represent the percentage of the systems accuracy. The accuracy percentage is calculated for each cell by comparing them with their counterparts in the input column enumeration.

Group Number	Quotations' input enumeration			Quotations' output enumeration		
	Correct	Distorted	Total	Correct	Distorted	Total
First dataset	100	N/A	100	100 (100%)	N/A	100 (100%)
Second dataset	150	N/A	150	150 (100%)	N/A	150 (100%)
Third dataset	200	N/A	200	200 (100%)	N/A	200 (100%)
Fourth dataset	70	30	100	70 (100%)	30 (100%)	100 (100%)
Fifth dataset	60	40	100	60 (100%)	40 (100%)	100 (100%)
Sixth dataset	55	45	100	55 (100%)	45 (100%)	100 (100%)

Table 5.1 QVT Validity Assessment Outcomes

The QVT validity assessment outcomes were developed based on a large-scale sample. Diversity in the sample facilitates the systems accuracy verification. Despite the difference in sample size and content, final results are similar. Thus, the systems precision will be considered very high as well as its accuracy.

5.3.2 Comparison Assessment Findings

In brief, this section will illustrate the comparison assessment between the developed QVT tool and the top five traditional Quran search engines that are retrieved by Google[1] [50–54]. The features' likeness between the chosen engines facilitates the efficiency's evaluation of the proposed tool.

Tested quotations	QVT	Muslim web [51]	Quran prospector [52]	Ketab Allah [50]	Alawfa [54]	Quran complex [53]
فَاسْجُدُوا لِلَّهِ وَاعْبُدُوا ۩	1	0	0	0	0	0
ذَٰلِكَ الْكِتَابُ لَا رَيْبَ ۛ فِيهِ ۛ هُدًى لِّلْمُتَّقِينَ	1	0	0	0	3524	0
كَلَّا ۖ لَيُنۢبَذَنَّ فِي الْحُطَمَةِ	1	0	0	0	1565	0
كَلَّا لَا تُطِعْهُ وَاسْجُدْ وَاقْتَرِبْ ۩	1	0	0	0	3097	0
كَلَّا ۖ بَل لَّا تُكْرِمُونَ الْيَتِيمَ	1	0	0	0	3313	0
كَلَّا ۖ بَلْ ۜ رَانَ عَلَىٰ قُلُوبِهِم مَّا كَانُوا يَكْسِبُونَ	1	0	0	0	3312	0
وَقِيلَ مَنْ ۜ رَاقٍ	1	0	0	0	2971	0
السَّمَاءُ مُنفَطِرٌ بِهِ ۚ كَانَ وَعْدُهُ مَفْعُولًا	1	0	0	0	1527	0
أُولَٰئِكَ عَلَىٰ هُدًى مِّن رَّبِّهِمْ ۖ وَأُولَٰئِكَ هُمُ الْمُفْلِحُونَ	2[2]	0	0	0	4070	0
خَتَمَ اللَّهُ عَلَىٰ قُلُوبِهِمْ وَعَلَىٰ سَمْعِهِمْ ۖ وَعَلَىٰ أَبْصَارِهِمْ غِشَاوَةٌ ۖ وَلَهُمْ عَذَابٌ عَظِيمٌ	1	0	0	0	2319	0
أَلَا إِنَّهُمْ هُمُ الْمُفْسِدُونَ وَلَٰكِن لَّا يَشْعُرُونَ	1	0	0	0	4048	0
وَأَنفِقُوا فِي سَبِيلِ اللَّهِ وَلَا تُلْقُوا بِأَيْدِيكُمْ إِلَى التَّهْلُكَةِ ۛ وَأَحْسِنُوا ۛ إِنَّ اللَّهَ يُحِبُّ الْمُحْسِنِينَ	1	0	0	0	4604	0
ثُمَّ أَفِيضُوا مِنْ حَيْثُ أَفَاضَ النَّاسُ وَاسْتَغْفِرُوا اللَّهَ ۚ إِنَّ اللَّهَ غَفُورٌ رَّحِيمٌ	1	0	0	0	4809	0
أُولَٰئِكَ لَهُمْ نَصِيبٌ مِّمَّا كَسَبُوا ۚ وَاللَّهُ سَرِيعُ الْحِسَابِ	1	0	0	0	1052	0
وَعَدَ اللَّهُ الَّذِينَ آمَنُوا وَعَمِلُوا الصَّالِحَاتِ ۙ لَهُم مَّغْفِرَةٌ وَأَجْرٌ عَظِيمٌ	1	0	0	0	2572	0
أَعَدَّ اللَّهُ لَهُمْ عَذَابًا شَدِيدًا ۖ فَاتَّقُوا اللَّهَ يَا أُولِي الْأَلْبَابِ الَّذِينَ آمَنُوا ۚ قَدْ أَنزَلَ اللَّهُ إِلَيْكُمْ ذِكْرًا	1	0	0	0	3330	0
إِلَّا امْرَأَتَهُ قَدَّرْنَا ۙ إِنَّهَا لَمِنَ الْغَابِرِينَ	1	0	0	0	1871	0
إِنْ تَحْرِصْ عَلَىٰ هُدَاهُمْ فَإِنَّ اللَّهَ لَا يَهْدِي مَن يُضِلُّ ۖ وَمَا لَهُم مِّن نَّاصِرِينَ	1	0	0	0	5341	0
وَيَخِرُّونَ لِلْأَذْقَانِ يَبْكُونَ وَيَزِيدُهُمْ خُشُوعًا ۩	1	0	0	0	0	0
فَتَوَلَّ عَنْهُمْ ۘ يَوْمَ يَدْعُ الدَّاعِ إِلَىٰ شَيْءٍ نُّكُرٍ	1	0	0	0	1058	0

Table 5.2 Quotations Comparison Assessment Outcomes [3]

[1] Search was done by Google using the following Arabic sentence "البحث في القرآن ".
[2] Verse is repeated twice with the same script in two different Surahs.
[3] Sites were arranged in a table based on its rank in Google.

Table 5.2 shows the comparison outcomes between the proposed tools and the top five traditional Quran search engines based on an exact match. Thus, the comparison was executed using 20 Quranic quotations as clarified previously (see Section 5.2). Each quotation included its own diacritics and symbols at the test time. Results exposed in the table represent the number of outcomes retrieved by each search engine. As a rule, the expected result should be no more than one. Hence, the results prove that the proposed tool has a highly accurate outcome. In contrast, the traditional Quran search engines either retrieve no results [50–53] or retrieve a huge number of results [54]. This is due to the fact that the proposed tool considers the special Quranic symbols as described above unlike the traditional search engines. Besides this, the phrase based search pattern that is developed into the proposed tool reduces the results total to the minimum accurate outcome.

Tested phrases	**Accepted results**	**QVT**	**Muslim web** [51]	**Quran prospector** [52]	**Ketab Allah** [50]	**Alawfa** [54]	**Quran complex** [53]
"سبح لله"	5	5/ **5** (100%)	5/ **5** (100%)	5/ **5** (100%)	5/ **5** (100%)	5/ **1877** (0.3%)	3/ **3** (99.9%)
"ضرب الله مثلا"	7	7/ **7** (100%)	7/ **7** (100%)	7/ **7** (100%)	7/ **7** (100%)	7/ **1778** (0.4%)	4/ **4** (99.8%)
"شجرة طيبة"	1	1/ **1** (100%)	1/ **1** (100%)	1/ **1** (100%)	1/ **1** (100%)	1/ **24** (4.2%)	0/ **0** (95.8%)
"خلق السماوات والأرض"	29	29/ **29** (100%)	29/ **29** (100%)	29/ **29** (100%)	28/ **28** (99.7%)	29/ **354** (8.2%)	28/ **28** (99.7%)
"كتب الله"	5	5/ **5** (100%)	5/ **5** (100%)	5/ **5** (100%)	5/ **5** (100%)	5/ **1769** (0.3%)	5/ **5** (100%)
"عيسى ابن مريم"	16	16/ **16** (100%)	16/ **16** (100%)	16/ **16** (100%)	16/ **16** (100%)	16/ **84** (19.0%)	12/ **12** (95.5%)
"لمن المرسلين"	5	5/ **5** (100%)	5/ **5** (100%)	5/ **5** (100%)	5/ **5** (100%)	5/ **211** (2.4%)	5/ **5** (100%)
"ليجمعنكم إلى يوم القيامة"	2	2/ **2** (100%)	2/ **2** (100%)	2/ **2** (100%)	2/ **2** (100%)	2/ **776** (0.3%)	2/ **2** (100%)
"آيات الله"	42	42/ **42** (100%)	42/ **42** (100%)	42/ **42** (100%)	42/ **42** (100%)	42/ **1927** (2.2%)	19/ **19** (98.8%)
"يا أيها الذين آمنوا"	89	89/ **89** (100%)	88/ **88** (99.9%)	89/ **89** (100%)	0/ **0** (95.2%)	89/ **1850** (4.8%)	89/ **90** (99.9%)
Overall accuracy		**100%**	**99.9%**	**100**	**99.5%**	**4.2%**	**98.9%**

Table 5.3 Phrases Comparison Assessment Outcomes

As shown in table 5.3, the table gives the comparison outcomes for 10 Arabic phrases that have been tested with six different environments. Each row represents an Arabic phrase. The bare numbers represent the number of total correct retrieved results.

The bold bare numbers represent the number of total retrieved results. The numbers placed in brackets represent the percentage of the environments accuracy. The accuracy percentage was calculated for each cell by comparing them with the accepted result for each phrase. The accepted result is the accurate results which are proven through a human review in our assessment test. The number of incorrect retrievals is the inaccurate results that have been retrieved; it is not identical to the tested phrase. The environments accuracy will be determined based on the following equation:

$$Environment\ accuracy\ percentage = \frac{CR + IN}{CR + IN + CN + IR} * 100$$

Where CR = total number of correct retrieved results, IN = total number of incorrect not retrieved results, CN = total number of correct not retrieved results, IR = total number of incorrect retrieved results.

As shown in table 5.4, the table demonstrates the comparison results for 10 Arabic words that have been tested with six different environments. Prior accuracy equations will be applied.

Tested Words	**Accepted results**	**QVT**	**Muslim web** [51]	**Quran prospector** [52]	**Ketab Allah** [50]	**Alawfa** [54]	**Quran complex** [53]
ويل	38	38/ **48** (85%)	38/ **68** (56%)	38/ **68** (56%)	38/ **67** (57%)	38/ **68** (56%)	14/ **14** (65%)
مدين	10	10/ **13** (89%)	10/ **27** (37%)	10/ **27** (37%)	10/ **27** (37%)	10/ **27** (37%)	9/ **9** (96%)
دين	90	90/ **99** (97%)	90/ **259** (35%)	90/ **259** (35%)	90/ **259** (35%)	90/ **259** (35%)	9/ **9** (69%)
عيسى	25	25/ **25** (100%)	25/ **25** (100%)	25/ **25** (100%)	25/ **25** (100%)	25/ **25** (100%)	16/ **16** (64%)
آية	79	79/ **79** (100%)	79/ **241** (33%)	79/ **82** (99%)	79/ **79** (100%)	79/ **241** (33%)	44/ **44** (85%)
سلام	39	36/ **37** (92%)	39/ **47** (83%)	39/ **47** (83%)	39/ **47** (83%)	39/ **47** (83%)	19/ **19** (57%)
بيت	26	24/ **26** (88%)	26/ **32** (81%)	26/ **32** (81%)	26/ **32** (81%)	26/ **32** (81%)	5/ **5** (34%)
محمد	4	4/ **4** (100%)	4/ **4** (100%)	4/ **4** (100%)	4/ **4** (100%)	4/ **4** (100%)	4/ **4** (100%)
يأن	1	1/ **1** (100%)	1/ **46** (2%)	1/ **46** (2%)	1/ **1** (100%)	1/ **46** (2%)	1/ **1** (100%)
صدقة	5	5/ **5** (100%)	5/ **8** (63%)	5/ **5** (100%)	5/ **5** (100%)	5/ **8** (63%)	4/ **4** (88%)
Overall accuracy		**95.1%**	**59.0%**	**69.3%**	**79.3%**	**59.0%**	**75.8%**

Table 5.4 Words Comparison Assessment Outcomes

The numbers placed in brackets represent the percentage of the environments accuracy. The bare numbers represent the number of total correct retrieved results. The bold bare numbers represent the number of total retrieved results. The accuracy percentage was calculated for each cell by comparing them with the accepted result for each word.

In the prior two tables, the results show that the proposed tool has a higher accuracy percentage based on the correct results that are retrieved in comparison to the traditional Quran search engines. However, the QVT accuracy percentage in table 5.4 did not reach 100% because some of the tested words returned incorrect outcomes. For example, the tested word "**ويل**" retrieved two diverse words with different meanings " **ويلـ**عبون" and "**ويلـ**هيهم". As shown, the retrieved words include the word "**ويل**" in its morphological structure which limits the search patterns' capability.

Although the accuracy percentage for the QVT system in table 5.4 did not reach the maximum percentage, the improvement in the outcomes' results is due to the integration of the regular expression patterns with the phrase-based patterns. This minimizes the suffix, prefix, diacritics, and symbols' obstacles, which are not properly considered by some of the traditional search engines. As seen in prior tables, a considerable increase in the accuracy percentage has been distinguished by using the proposed tool in comparison with the results of the traditional Quran search engines. Thus, a significant improvement in the results will be observed when using the proposed tool.

Chapter 6 : Conclusion and Future Work

6.1 Conclusion

In this thesis, the design, development, and implementation of a new tool called QVT have been discussed. QVT is a web service that facilitates the process of quote validation within Arabic script. Furthermore, it supports integration with several applications' platforms such as web sites, windows applications, browsers and so on. The design's aspects enable the tool to overcome any interoperability obstacles between various system's platforms. On the contrary, other solutions are lacking a mechanism to support the integration of their solutions with different applications.

To enhance the system integration, QVT was developed based on (SOA) organizational pattern. The system was divided into three parts. One part aims to provide information about the service to the developers, the second part aims to apply the search pattern process, and the last part is a separate database. These different parts result in scalable system integration.

To the best of our knowledge, the data revision mechanism determined in this thesis has not been implemented by any Arabic text retrieval system before. The data revision can overcome the obstacles raised from the characteristics of some religious books which distinguish them from other Arabic books. Indeed, integrating the data revision mechanism will enhance the results' accuracy of the text retrieval system and will also lead to a high-precision system. Equally important, the information retrieval mechanism determined in this thesis has contributed a clear enhancement in the retrieved results in comparison to the existing text retrieval systems.

The multi GUI capability for the QVT system depends on the successful integration of the web service. As the QVT is based on XML-based technologies, every time a developer requests to integrate QVT using WSDL, the Software Development

Framework (SDF) is automatically incorporated in the web service and initiates the SOAP message's infrastructure. The integration process requires little interaction from the developer’s side. Accordingly, the QVT systems distribution endeavors will be simplified.

Many studies are discussing the text retrieval systems that support keyword search methodology. However, their approaches mostly depend on a bare word search. QVT search techniques facilitate the search efforts to overcome obstacles resulting from the Arabic’s diacritics, unique symbols and multi character formats.

Phrase matching is a study area that is acquiring grand attention among researchers and software developers. This thesis had exposed the architecture of a phrase based search system. Implemented in the architecture is a matching service that can be developed in order to increase the outputs accuracy with high ratio precisions for different Arabic books recourses.

Many studies and developers are trying to provide a predefined pattern search service within text retrieval systems. The search service that is offered in this thesis is based on regular expression prototype and demonstrates an improvement in the accuracy of the retrieved results in comparison with those which are designed with different prototypes. Regular expression affords valuable meta-characters that can be utilized in a predefined pattern search.

The service stability and consistency are the core objectives of researchers and developers to provide a distinct service to the end user. The DBaaS service that is utilized in this thesis is based on a cloud computing infrastructure. Adoption of such a service has supported the system to introduce a distinguished service to the end user. Besides, it demonstrates improvement in the services sustainability compared with local database services. As well, it minimizes the efforts to call, maintain and move the database.

Enhancing existing text retrieval systems is always the main objective of the researchers who study this area. The features developed within QVT, which are the XML-based web service, a data revision mechanism, a phrase matching service, a predefined pattern search service, and DBaaS cloud computing will provide complete and precise support to the Arab users as their desires are fully considered.

6.2 Limitations

Due to the distinctive characteristics of the Arabic language, some words used in Arabic texts differ in the verbatim translation from the true meaning behind the use of the word in a given sentence.

There is some performance limitation related to the use of cloud computing such as the response time. The response time between the QVT and DBaaS was not considered in this thesis. This is due to the fact that DBaaS cloud service was recently launched.

6.3 Future Work

There are some essential outstanding works that need to be accomplished. First, the translations of Arabic books need to be validated. Most of the Arabic books were translated into different languages. Some of these translations contain multiple translation productions depending on the translator and the language glossary. Various translations may lead to different intended meanings of the translation. Some translations may depend on the verbatim translation. So, the translation's validation process in different languages is a promising area of study for researchers.

Secondly, implementing the semantic search methodology within the validation tool will improve the search accuracy in the system. As well, it will facilitate the implementation of the translation's validation process in the future.

Third, Improving the QVT performance is one of the future aims by reducing the time required to retrieve the information.

Fourth, the future plan is to implement the validation concept in general. In this thesis, the validation concept was applied for quotations that are extracted from the given Arabic books. The long-term goal is to implement the validation concept in general Arabic texts by applying the general rules of the Arabic language which is what distinguishes it from other languages. Some of these rules which can be applied are the grammatical and morphological rules. In addition, improving the tool by including the process of diacritic's validity in the general sentences such as Thumm'ah for the Doer, Fett'ha for the Object and Kas'rah for the Genitive and so forth.

References

[1] Miniwatts Marketing Group., "INTERNET WORLD USERS BY LANGUAGE," *Internet World Statistics*, 2010. [Online]. Available: http://www.internetworldstats.com/stats7.htm. [Accessed: 01-May-2011].

[2] Communications and Information Technology Commission Saudi Arabia, "The use of computers in the Kingdom of Saudi Arabia-all sectors from 2007 to 2009," 2009.

[3] TRA, "Annual report 2009," Abu Dhabi, 2009.

[4] Ministry of Communications and Information Technology and Egypts, "Summary report on indicators of Communications and Information Technology," 2011.

[5] Ministry of Communications and Information Technology and Egypts, "Yearbook 2010," 2010.

[6] Ministry Of Communications and Information Technology Saudi Arabia, "The National Communications and Information Technology Plan The Vision Towards the Information Society."

[7] King Abdulaziz City for Science and Technology, "Strategic Priorities for Information Technology." p. 46, 2006.

[8] King Abdulaziz City for Science and Technology, "King Abdullah Initiative for Arabic Content," 2010. [Online]. Available: http://www.econtent.org.sa/Pages/Default.aspx. [Accessed: 01-May-2011].

[9] 19th Session of the Council of the League of Arab States, "The Riyadh Declaration," *the Council of the League of Arab*, 2001. [Online]. Available: http://portal.mofa.gov.sa/Detail.asp?InSectionID=5445&InNewsItemID=62706. [Accessed: 01-May-2011].

[10] A. Blogs and A. Embodiment, "Arabic Blogs : An Embodiment of Freedom of Expression," no. 1, pp. 1-6, 2006.

[11] T. C. Legal, W. H. Dutton, A. Dopatka, M. Hills, G. Law, and V. Nash, "Freedom of Connection – Freedom of Expression," no. November, 2010.

[12] Y. Tian, "Re-thinking Intellectual Property," *Human Rights*, p. 359, 2009.

[13] A. Al-Zoghby, N. a. Ismail, and T. Hamza, “Mining arabic text using soft-matching association rules,” *2007 International Conference on Computer Engineering & Systems*, pp. 421-426, 2007.

[14] F. a. Allah, a. El qadi, S. Boulaknadel, and D. Aboutajdine, “Arabic Information Retrieval System Based on Noun Phrases,” *2006 2nd International Conference on Information & Communication Technologies*, vol. 1, pp. 1720-1725, 2006.

[15] S. M. Alzahrani, N. Salim, and M. M. Alsofyani, “Work in Progress: Developing Arabic Plagiarism Detection Tool for E-Learning Systems,” *2009 International Association of Computer Science and Information Technology - Spring Conference*, pp. 105-109, 2009.

[16] R. a. Haraty and R. Varjabedian, “ADD: Arabic duplicate detector - a duplicate detection data cleansing tool,” *ACS/IEEE International Conference on Computer Systems and Applications, 2003. Book of Abstracts.*, p. 137.

[17] R. Lopes, “On the Credibility of Wikipedia : an Accessibility Perspective,” *Challenge*, pp. 27-34, 2008.

[18] J. Fong, F. Pang, and C. Bloor, “Converting relational database into XML document,” in *Database and Expert Systems Applications, 2001. Proceedings. 12th International Workshop on*, 2001, pp. 61–65.

[19] S. A. Almulla and C. Y. Yeun, “Cloud computing security management,” in *Engineering Systems Management and Its Applications (ICESMA), 2010 Second International Conference on*, 2010, pp. 1–7.

[20] Y. Kidawara, “Information credibility analysis of web content,” *Proceeding of the 2nd ACM workshop on Information credibility on the web - WICOW ’08*, p. 3, 2008.

[21] A. Amin, J. Zhang, H. Cramer, L. Hardman, and V. Evers, “The Effects of Source Credibility Ratings in a Cultural Heritage Information Aggregator Categories and Subject Descriptors,” *Source*, pp. 35-42, 2009.

[22] L. C. M. Tang, Y. Zhao, S. Austin, M. Darlington, and S. Culley, “A characteristic based information evaluation model,” *Proceeding of the 2nd ACM workshop on Information credibility on the web - WICOW ’08*, p. 89, 2008.

[23] T.-yuan Liu, S. Member, W.-hsiang Tsai, and S. Member, “Quotation Authentication : A New Approach and Efficient Solutions by Cascaded Hashing Techniques,” vol. 5, no. 4, pp. 945-954, 2010.

[24] R. M. B. Al-Eidan, H. S. Al-Khalifa, and A. S. Al-Salman, “Towards the measurement of Arabic Weblogs credibility automatically,” *Proceedings of the*

11th International Conference on Information Integration and Web-based Applications & Services - iiWAS '09, p. 618, 2009.

[25] M. F. Noordin and R. Othman, "An Information Retrieval System for Quranic Texts: A Proposed System Design," *2006 2nd International Conference on Information & Communication Technologies*, pp. 1704-1709, 2006.

[26] J. M. O'Toole, *Authenticity in a Digital Environment (review)*, vol. 1, no. 4. 2001, pp. 537-539.

[27] H. Moukdad, "Lost in cyberspace: How do search engines handle Arabic queries," in *the 32nd Annual Conference of the Canadian Association for Information Science*, 2004, pp. 1-7.

[28] B. Pinkerton, "Webcrawler: Finding what people want," Citeseer, 2000.

[29] C. Zhang and S. Zhan, "RESEARCH AND IMPLEMENTATION OF FULL-TEXT RETRIEVAL SYSTEM USING COMPASS BASED ON LUCENE," *Engineering and Technology*, vol. 11, no. 9, pp. 1120-1126, 2011.

[30] K. Patterson, C. Watters, and M. Shepherd, "Document Retrieval using Proximity-based Phrase Searching," in *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual*, 2008, pp. 137–137.

[31] G. Rasool and N. Asif, "Software Artifacts Recovery using Abstract Regular Expressions," *2007 IEEE International Multitopic Conference*, pp. 1-6, Dec. 2007.

[32] Y. Zhenjun and J. Xiangyu, "A simplified application of regular expressions: With the extraction of Chinese cultural terms as an example," in *Computing, Communication, Control, and Management, 2009. CCCM 2009. ISECS International Colloquium on*, 2009, vol. 1, pp. 439–442.

[33] D. Booth, H. Haas, F. McCabe, E. Newcomer, C. Ferris, and D. Orchard, "Web Services Architecture," *W3C*, 2004. [Online]. Available: http://www.w3.org/TR/ws-arch/. [Accessed: 24-Nov-2011].

[34] C. F. Goldfarb, "Information processing - Text and office systems Standard Generalized Markup Language SGML," *International Organization for Standardization, Geneva, Switzerland*, 1986.

[35] K. Sahin and M. Gumusay, "Service oriented architecture (SOA) based web services for geographic information systems," in *XXIst ISPRS Congress. Beijing*, 2008, pp. 625–630.

[36] M. Mertz and M. Gryning, "Cloud Computing and SOA." p. 65, Jan-2009.

[37] C. Doukas, T. Pliakas, and I. Maglogiannis, “Mobile healthcare information management utilizing cloud computing and android OS,” in *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, 2010, vol. 2010, pp. 1037–1040.

[38] P. Teregowda, B. Urgaonkar, and C. L. Giles, “Cloud computing: A digital libraries perspective,” in *Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on*, 2010, pp. 115–122.

[39] O. Vechtomova and M. Karamuftuoglu, “Approaches to high accuracy retrieval: Phrase-based search experiments in the HARD track,” in *Proceedings of TREC*, 2004, no. 1.

[40] D. Cutting, J. Pedersen, and P. K. Halvorsen, “An object-oriented architecture for text retrieval,” in *Conference Proceedings of RIAO*, 1991, vol. 91, pp. 285–298.

[41] H. K. Al Ameed et al., “Arabic Search Engines Improvement: A New Approach using Search Key Expansion Derived from Arabic Synonyms Structure,” *IEEE International Conference on Computer Systems and Applications, 2006.*, pp. 944-951, 2006.

[42] H. Moukdad and A. Large, “Information retrieval from full-text Arabic databases: Can search engines designed for English do the job?,” *Libri*, vol. 51, no. 2, pp. 63–74, 2001.

[43] E. Lagerspetz and S. Tarkoma, “Mobile search and the cloud: The benefits of offloading,” *1st IEEE PerCom Workshop on Pervasive Communities and Service Clouds*, pp. 117-122, 2011.

[44] D. Box et al., “Simple Object Access Protocol (SOAP) 1.1,” *W3C*, 2000. [Online]. Available: http://www.w3.org/TR/2000/NOTE-SOAP-20000508/#_Toc478383532. [Accessed: 01-Dec-2011].

[45] E. Christensen, F. Curbera, G. Meredith, and S. Weerawarana, “Web Services Description Language (WSDL) 1.1,” *W3C*, 2001. [Online]. Available: http://www.w3.org/TR/wsdl#_how-s. [Accessed: 09-Dec-2011].

[46] M. Sabbouh, S. Jolly, D. Allen, P. Silvey, and P. Denning, “World Wide Web Consortium,” *W3C*. San Jose, p. 4, 2001.

[47] M. A. Aabed, S. M. Awaideh, A. R. M. Elshafei, and A. A. Gutub, “Arabic diacritics based steganography,” in *Signal Processing and Communications, 2007. ICSPC 2007. IEEE International Conference on*, 2007, no. November, pp. 756–759.

[48] M. Widenius and D. Axmark, *MySQL reference manual: documentation from the source*. O'Reilly Media, Inc., 2002, p. 172.

[49] Xeround Inc., "Xeround Cloud DataBase," 2011. [Online]. Available: http://xeround.com/. [Accessed: 25-Nov-2011].

[50] Ketaballah.net, "The Holy Quran search engine." [Online]. Available: http://www.ketaballah.net/searchquran.html. [Accessed: 08-Dec-2011].

[51] Muslim-web.com, "The Holy Quran." [Online]. Available: http://quran.muslim-web.com/?lang=en. [Accessed: 08-Dec-2011].

[52] Holyquran.net, "The Quran's prospector." [Online]. Available: http://www.holyquran.net/search/sindex.php. [Accessed: 08-Dec-2011].

[53] K. F. C. for the P. of the H. Qur'an, "The Holy Quran-Textual Search." [Online]. Available: http://www.qurancomplex.org/quran/Search/Search.asp?Adv=0&l=arb&TabID=1&SubItemID=10. [Accessed: 08-Dec-2011].

[54] alawfa.com, "ALAWFA-Search engine in Quran." [Online]. Available: http://www.alawfa.com/#. [Accessed: 08-Dec-2011].