Applied Computing and Informatics (2017) 13, 38–46

Saudi Computer Society, King Saud University

Applied Computing and Informatics

(http://computer.org.sa)
www.ksu.edu.sa
www.sciencedirect.com

ORIGINAL ARTICLE

Online integrity and authentication checking for Quran electronic versions

Izzat Alsmadi [a,*], Mohammad Zarour [b]

[a] Computer Information Systems Department, Yarmouk University, Jordan
[b] Prince Sultan University, Saudi Arabia

Received 24 April 2015; revised 18 July 2015; accepted 10 August 2015
Available online 14 August 2015

KEYWORDS

Information retrieval;
Security;
Search engines;
Documents' automatic evaluation;
Hashing algorithms;
Information authentication

Abstract The ability to control data and information through the Internet can be challenging. Preliminary analysis showed that some tampering and forgery may occur to some words of the Quran in the electronic versions that span the Internet. Such small modifications may not be noticed by public audience. The holy book of Quran includes a unique feature in that its worldwide copies are all identical. The 114 chapters (Suras) and all their verses and words are preserved in the exact form. As such, we designed and evaluated a model and a tool to evaluate the integrity of the wording in the e-versions of the Quran through generating a Meta data related to all words in the Quran preserving the counts and locations. Such Meta data can be used in the same way hash algorithms are used in security to check the integrity of a disk and its data files where any small change in the data will result in a different hash value. We conducted several experiments to evaluate the different parameters and challenges that can impact the automatic authentication process of Quranic verses based on information retrieval and hashing algorithms.

1. Introduction

With the huge amount of information uploaded by users all over the world, internet is the central library that all humans in the world are using to exchange information. However, one of the major problems with the current web is that there is no control on the type, nature or details of information uploaded to the web by any user. Accordingly, credibility and accuracy are the foremost concerns when reading information from the web. Credibility can be linked to the website, hosting the information, author of uploaded data and the type of data or information itself. For example, in social networks or even news websites the same story can be told in exactly opposite versions given by two different websites or authors.

Authenticity and integrity of online documents is increasingly getting crucial as many organizations put their documents online. Providing facilities that allow both documents' owners and viewers to be able to ensure that their documents are authentic and not tampered is extremely important. Currently two major techniques or approaches are used to authenticate documents, or users online: Document control

* Corresponding author.
E-mail addresses: ialsmadi@yu.edu.jo (I. Alsmadi), mzarour@psu.edu.sa (M. Zarour).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.aci.2015.08.001

and digital signatures. Document control is related to permissions before and after publishing the document online. Document control then is needed to be dynamically applied in a manner that document can be tracked even after publishing it online. Integrity is about making sure that the document is not altered by any unauthorized person and hence is authenticated. In digital signatures, signed documents should be verified by the people who signed it (i.e. non-repudiation).

Nowadays, many multilingual copies of the holy Quran are available online. The holy Quran is originally written in Arabic language. The exact wording and statements in Quran verses are identical for all Arabic versions of the Quran.[1] However, one problem with translation is that it may change, intentionally or unintentionally the meaning of some verses when translated to another language. This is a general problem when dealing with translation. On the other hand, for the same language, it is possible that the same verse be written in different words due to intentional fraud or due to language translation issues.

Due to the sensitivity and the nature of Quran verses there is a vital need to continuously monitor Quran verses and chapters written through the Internet websites and pages to make sure that they are authenticated and not changed or fraud.

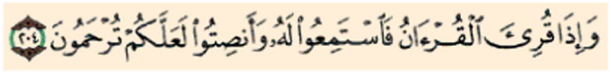
Hashing algorithms or checksums are used to verify the integrity of data in files, disks and databases. The hashing algorithms or tools take file or data as inputs and generate a unique decimal or hexadecimal number in a way where any small change that may occur to the data in the file will cause the hashing algorithm to produce a new random number that is different from the previous number that was generated before the modification. The Holy Quran is reserved and integrated in all its chapters and verses word by word and letter by letter. Hence using hashing algorithms to check automatically and frequently whether there is any possible modification can be an effective tool to authenticate Quran electronic versions either for the whole Quran version or in parts: chapters or verses. Accordingly, Hashing can be generated for the whole Quran as one unit, chapter by chapter or even verse by verse. Fig. 1 below shows an example for hashing codes using the verse:

Removing one letter from the verse may cause all hash values to be changed to totally new values.

As shown in Fig. 1, different hashing algorithms generate different hashing values for the same verse. Moreover, hashing algorithms do not differentiate between small and large modification. Any tiny modification will change the hash codes just like a large modification.

The identity and integrity of the data must be periodically and systematically verified through mechanisms such as Secure Hash Standard (SHS) and Secure Hash Algorithm (SHA) [1]. Quran verses can be considered plagiarized or unauthenticated even for one letter change. Hence, hash algorithms are good option as such algorithms are very sensitive even for a very small change in data or documents.

In this paper, an online integrity and authentication system for the Quran is proposed. The goal of such system is to continuously traverse and search information and web pages throughout the Internet for any possible intentional and unintentional Quran verses fraud or change.



Original text	وإذا قرئ القرآن فاستمعوا له وانصتوا لعلكم ترحمون
Original bytes	d988d8a5d8b0d8a720d982d8b1d8a620
Adler32	93763c14
CRC32	42549328
Haval	1b3b4f5c96769ea1b158ce8fa627e2b3
MD2	0583a09b8dd5046ad4ae5add83271f10
MD4	e8e26ce992452933886b17fd7d1b5aaa
MD5	52392ce5e34c1c468241d0cda7b17725

Figure 1 A sample of Quran verses hashing.

It should be mentioned that in this research we are not focusing on encryption but rather integrity checking. The problem is that in our scenario for Quranic documents online, the relation between information providers and readers is not a typical e-commerce one-to-one relation in which typical hash or encryption algorithms can be sent to the known receiver or customer. Accordingly, private keys intended for a particular receiver will not be applicable in our scenario where the tool is expected to check quickly through the Internet that Quranic documents are authentic and do not contain tampering.

The rest of the paper is organized as follows: Section 2 discusses the related work to this research field followed by a discussion of the adopted methodology in Section 3. Section 4 discusses the developed authentication system, while the experiments are presented and discussed in Section 5. Section 6 presents the conclusion and future work.

1.1. Tashkeel, Harakat, or Diacritic marks

A unique aspect in Arabic is the addition of Diacritic marks or Tashkeel symbols (small و, small أ, small ي, Tanween: ٌ ٍ ً, stress or shaddah). Those are added to letters to give different meanings or ways of spelling letters. They may cause problems to indexing and querying of data in cases whether those Diacritics are added or ignored. We will present few of the proposed solutions to handle Tashkeel in Arabic.

Papers that tried to detect Diacritics proposed morphological rules to do that. Known patterns and an initial training set are used to match subject word for one of the morphological rules. Hidden Markov Model (HMM) is used where the hidden states are the possible diacritized selections of the words. In Arabic text, Diacritic marks can be challenging whether they exist or not in the Arabic text. If they exist, the challenge will be to read or parse them correctly especially as they need extra resources in addition to typical letters parsing in all other languages. On the other hand, Diacritic marks are necessary to Arabic words to clarify words and statements from ambiguity.

[1] Although different readings have few literal changes, those can impact the integrity checking process and hence we either use one reading or have different integrity datasets for the different readings. Experiments in this papers use "Hafs" reading.

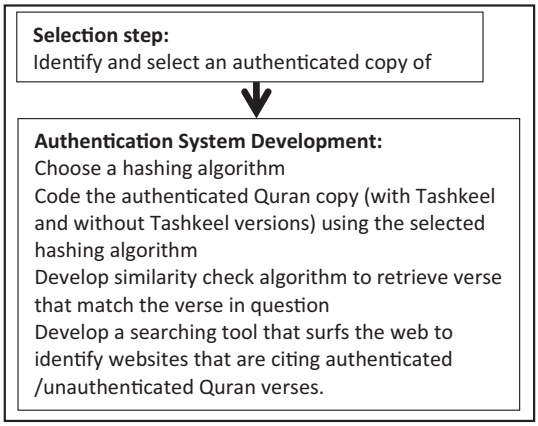


Figure 2 Research methodology.

Letters in Arabic have different forms based on their location, accordingly, indexer needs to understand that: (ا,إ ,أ) are all forms of the same letter. Fig. 2 shows a sample of a Sphinx table of Unicode character set that can be used as part of an indexer to guide the indexer for letters of different forms. The goal here is to acknowledge that many Arabic letters have different forms and hence a Unicode character table needs to have equivalent forms of similar letters. As mentioned earlier such issues are solved technically once the right encoding scheme is used.

2. Related work

While the subject of data or information integrity is not new, few research papers are published in the area of web documents integrity verification. This is particularly true if we want to exclude papers discussing encryption approaches.

The removal of online documents' replication is discussed in [2]. In many cases, the exact document is published in more than one website. This may not be related to plagiarism but to copies replication. This is actually quite common in the Internet world. The work presented in [2] used hashing comparison to discover and remove possible replication in web documents. The aim was to improve indexer optimization where one copy of the document will be stored if many documents have the exact same hash value.

The digital Quran verification and authentication is very rare; an algorithm to verify the fundamental text for Quranic quotes is presented in [3]. In order to generate a standard form of Quran verses, the algorithm is stemming some of the special Arabic characters [3] that may make the comparison process difficult. This includes Harakat or Tashkeel that can change the letters' spelling and the different shapes for some letters as well. While removing such symbols can facilitate the information retrieval processes, they may cause false alarms where some plagiarism may only occur in those parts which will then be missed or not detected. The work in [3] did not specify any limit or range to consider verses as not authenticated. It used Information retrieval and language processing methods with no utilization of security such as hashing methods.

A system to verify the authenticity of Quranic verses based on Unicode comparison is presented in [4]. The algorithm adopted in this system uses the Unicode representation of the characters to verify the fundamental text against the original Unicode text of the Quran in order to authenticate the fundamental text of the Quranic quotation, and then the position of diacritics and Tashkeel are verified. A similar study that used unicode centric string matching approach is presented in [8].

Sbbah and Selamat [9] have developed a framework for Quranic verses authenticity detection in online forum. They used a numerical approach to detect Quran quotes. The numerical approach calculates the diacritical ratio (DR) of each word in the text. Their authentication system depends on defining the weights of Quranic letters and diacritics and calculate the identifiers of distinctive Quranic words.

Yusoff et al. [5] discussed authentication in Hadeeth and used the historical traditional methods used to authenticate Hadeeths in current digital evidence authentication. Writers and collectors of Hadeeths used to follow the sequence or chain of Hadeeths authors and check their authenticity one by one. A deficiency in one member of the chain may weaken the whole chain and the Hadeeth itself.

In order to protect the digital Quran from any distortion or tampering, the work presented in [6] used the meta-codification where an "atomization structure" and Unicode codification is used on the Quran in terms of the number of verses in each chapter, the number of characters in each verse and so on.

Other research works as in [7] discussed the usage of digital signatures to authenticate information or documents through the web. This can work in e-commerce applications and emails where the typical two ways of communication exists between either sender and receiver or client–server. However, in our case, such architecture does exist and hence such methods cannot be used to check whether the information or the document was tampered or not. In Quran verses verification system, an authenticated copy of content is preserved and web content is surveyed for possible – different copies. It is also important to distinguish some changes related to Tashkeel, fonts, etc. that should not be classified as tampering. In particular we mean that some versions of the Quran include Tashkeel or Dialects and others do not. Hence comparison should take this into consideration.

In another but related domain, speech recognition techniques are used to verify Quran recitation as we started seeing many recitation applications which require new techniques to verify the validity of the recitation, see for example [10–12].

In our case, the goal is to use information retrieval techniques to look for documents that include unauthenticated Quranic verses. Authentication algorithms are then required to check whether these millions of documents online have the same – literally authenticated versions of the Quran. Of course such algorithms have two possibly contradictory requirements: performance and accuracy. Algorithms should be able to search through the Internet for possible frauds with a high accuracy and quick speed. False alarms may also arise due to issues such as different readings of the Quran where some specific letters or words are different from one reading to another.

3. Methodology

The aim of this research work is to build an authentication system that can verify any possible "change" or "modification" of

Quran verses. Since Quran verses are literally identical, we developed an authentication system that will index verses one by one in its indexer and will frequently search any possible – mutation in the verses. The authentication system will be able to frequently search the Internet for web pages that may include such fraud verses and index them for monitoring, alert, logging and verification.

Similar to the concept of kids safe search that some search engines used to limit or focus kids search on kids related websites and eliminate websites that may present or contain harmful content to kids, our proposed system should be able to retrieve results for only authenticated versions of Quran verses.

Our research methodology consists of three main steps (see Fig. 2):

1. Selecting a hashing algorithm: MD5 hashing algorithm has been chosen as it is a common and reliable one (see section experiments and discussions).
2. Developing of the authentication system: the authentication system has been developed using C# programming language, ASP.NET. This can be classified as a customized information retrieval or crawling system.

The details of developing the authentication system are discussed in more detail in Section 4. The goal of this research is to design a search engine or authentication system that is able to retrieve results based on a search query. A customized Quran search engine or information retrieval system should focus on parsing and indexing Quran verses from all over the web. For proposed program search user interface and queries will then represent verses or part of verses.

The proposed Quranic search engine website allows the search through the whole internet and retrieves links where the searched verses are located. The focus in this paper is to propose and design a system that can search for and index correct and incorrect versions of Quranic verses that may exist throughout the Internet.

The algorithm to develop the authentication system is shown in Fig. 3. The authentication system will start by checking whether the verse is written with Harakat and Tashkeel or not to decide which version of the Quran to search. Then the algorithm will look for the verses that are similar to the verse in question; some verses or part of them can be replicated in the Quran. The verses that have perfect match (similarity = 100%) or verses that are partially match (similarity $\geqslant$ 70%) are selected to authenticate the verse in question. Then, the hash code for the verse in question is generated and compared to the selected verses. If the hash code of one or more verses is identical to the hash code of the verse in question then the verse is authenticated, otherwise the verse is tampered.

1. Read Quran Verse (QVsi)
2. QVsi has Tashkeel (Yes or No)
3. Similar verses to QVsi exist? (Yes or No)
4. Generate hash code; HiQVsi of QVsi.
5. Compare HiQVsi with original hash code HoQVsi.
6. If HiQVsi equals HiQVsi, authenticate QVsi.
   • Else, the verse QVsi is unauthenticated

Figure 3 Quran verse authentication algorithm.

4. Quran authentication system development

We have conducted an experiment to evaluate the top websites that are used to search for Quranic verses. This is based on the popular search engine Google. We built a customized crawler to search through Google for Quran verses one by one from the first to the last verse. For each verse, we collected information related to the approximate number of index pages in Google for the verse. This represents the approximate number for web or Internet links that include the queried verse. The developed crawler uses Quran verses as search queries. Results are retrieved from embedded information in search web pages. Meta data related to the retrieved web pages are collected and saved in a dataset. Google ranks retrieved results based on its popularity page rank algorithm.

Fig. 4 shows a small sample of the collected data. We collected the first 10 links that are retrieved from Google search which indicate the most visited websites for the specific verse.

The number of indexed pages in Google is approximate, see Fig. 5, where this can be a direct indicator of the verse popularity. Quranic digitally published verses can be indexed as part of the whole Quran in some websites that include the whole Quran chapters and verses. In some other cases, some chapters or verses of the Quran may be indexed or included in some pages as part of non-Quranic text.

In order to deal with the verses having Harakat and Tashkeel, we have considered two versions of the Quran: one that considers Harakat or Tashkeel and one that does not. Quran indexed in the web can be with or without Tashkeel. In many cases, search engines and their queries can be sensitive to these symbols that are unique in Arabic language; hence a search for verse that does not include Tashkeel in the query may not retrieve results for verses that include Tashkeel symbols.

While there are some authentic websites such as www.islamweb.net, www.alro7.net, www.alhawali.com, many of the popular websites in the first search page or the top 10 pages are either from non-Islamic resources, or from some websites that do not represent the main stream of Islam and that may include inaccurate information, content or interpretation of these verses. We consider that the Internet is now the main source for searching for Quran verses and their interpretation not only for Muslims, but also for non-Muslims. Popular social network websites such as Facebook and YouTube are listed frequently as sources for Quranic verses. It is not clear however why Google is presenting them in the top 10 retrieved results. It is possibly that such high rank is not based on the popularity of the specific Facebook personal page or the specific YouTube video, but rather based on the popularity of the mother website (i.e. YouTube or Facebook).

Fig. 5 shows top 20 verses based on Google indexed pages that are popular. This is accomplished after indexing almost the first half of the Quran. Fig. 5 shows that the number of Quran indexed verses may vary from one verse to another and from one chapter to another. The first most popular verse of the Quran is AlBasmalah (بسم الله الرحمن الرحيم). In fact this verse can be used in many contexts not necessary part of citing Quranic verses. People may use it to start a book, a chapter or a process. The verses in the first chapter are also popular for almost the same reasons. Some short verses can get a large number of indexed links as such short verse may be used in non-Quranic contexts (see Fig. 6).

Figure 4 A small sample of Quran search crawler.

Figure 5 A sample of Google retrieved results.

For the Quran authentication system to work properly, we built an indexer that includes all Quran verses (based on the version available on the Islamweb website) where each line or row in the indexer includes one verse. Information about each verse should include its chapter, sequence number, content and number of characters and MD5 hash value. MD5 hash value is saved as the hash table value besides the verse as a key object, see Fig. 7.

Total hash items are 6236 to represent total Quran verses. Fig. 8 shows a small screenshot from Quran indexer including the attributes described earlier.

5. Experiments and discussions

In this section, several experiments are discussed, some of them helped us in making decisions related to the hashing algorithm in developing the authentication system and at the end we have conducted some experiments to test our authentication system. As mentioned earlier, there are several challenges that may face the process of authenticating automatically electronic verses and chapters of the noble Quran. Such challenges are related to different languages, fonts, styles, and Quranic read-

SurahNO	VerseNO	Verse	countlinks	First Link	Second Link
1	1	بسم الله الرحمن الرحيم	1290000000	www.youtube.com/watch?v=2Y2-9QK2Q	www.youtube.com/watch?v
1	2	الحمد لله رب العالمين	168000000	www.youtube.com/watch?v=2Y2-9QK2Q	www.youtube.com/watch?v
1	3	الرحمن الرحيم	97200000	www.youtube.com/watch?v=2Y2-9QK2Q	www.youtube.com/watch?v
3	2	الله لا إله إلا هو الحي القيوم	10200000	www.islamweb.net › ... › سورة البقرة ›	www.islamweb.net › ... › بقرة ›
1	4	مالك يوم الدين	7540000	www.youtube.com/watch?v=2Y2-9QK2Q	www.youtube.com/watch?v
20	26	ويسر لي أمري	6290000	https://ar-ar.facebook.com/allah.akbar11	https://ar-ar.facebook.com/
14	20	وما ذلك على الله بعزيز	2120000	www.islamweb.net › ... › سورة فاطر ›	www.islamweb.net › ... › فاطر ›
15	93	عما كانوا يعملون	1050000	www.islamweb.net › ... › سورة الحجر ›	www.islamweb.net › ... › حجر ›
20	28	يفقهوا قولي	1020000	www.flickr.com/photos/mahyawa/68647	www.flickr.com/photos/mal
3	8	ربنا لا تزغ قلوبنا بعد إذ هديتنا و	947000	www.islamweb.net › ... › سورة البقرة ›	www.islamweb.net › ... › بقرة ›
15	38	إلى يوم الوقت المعلوم	811000	www.islamweb.net › ... › سورة الحجر ›	www.islamweb.net › ... › حجر ›
2	163	وإلهكم إله واحد لا إله إلا هو الر	747000	www.youtube.com/watch?v=2Y2-9QK2Q	www.youtube.com/watch?v
20	4	تنزيلا ممن خلق الأرض والسما	740000	www.islamweb.net › أضواء البيان › سورة طه ›	www.islamweb.net › سورة طه ›
4	136	يا أيها الذين آمنوا آمنوا بالله ورس	739000	www.islamweb.net › ... › تفسير سورة النساء ›	www.islamweb.net › ... › نساء ›
20	51	قال فما بال القرون الأولى	707000	www.islamweb.net › ... › البغوي › سورة طه ›	www.islamweb.net › ... › طه ›
1	6	اهدنا الصراط المستقيم	631000	www.youtube.com/watch?v=2Y2-9QK2Q	www.youtube.com/watch?v
19	92	وما ينبغي للرحمن أن يتخذ ولدا	605000	www.islamweb.net › ... › تفسير سورة مريم ›	www.islamweb.net › ... › مريم ›
20	8	الله لا إله إلا هو له الأسماء الحسنى	583000	www.holyquran.net/cgi-bin/prepare.pl?q	www.holyquran.net/cgi-bin/
19	88	وقالوا اتخذ الرحمن ولدا	581000	www.islamweb.net › ... › تفسير سورة مريم ›	www.islamweb.net › ... › مريم ›

Figure 6 Top 19 verses based on rank.

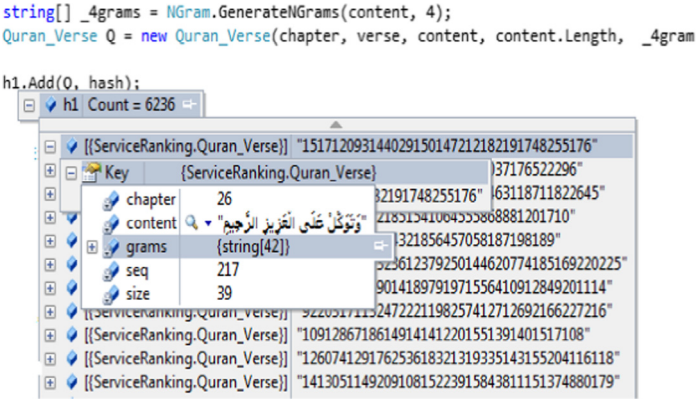


Figure 7 Quran indexer content snapshot.

Figure 8 A sample from Quran indexer.

Original text	إذا وقعت الواقعة
Original bytes	d8a5d990d8b0d98ed8a720d988d98ed982d98ed8b9d98ed8aa... (length=53)
Adler32	e00d2455
CRC32	c1c4f7d0
Haval	49dc4d21b959e0ca711ba0f83e721adb
MD2	2d1536072011971f14b88fdbfa4d6a73e
MD4	b341d2b6c312d18ff72c5e4c022803f1
MD5	10910a94d5f67c532125fdd9b181ce70
RipeMD128	2e639d7f66c36b136d62f960610e2954
RipeMD160	3c4bd593d7342eaa8c43d1d2d6cf86a5697e2440
SHA-1	83cf881fdda0b5e1f18defa6abe1fa965901fd90
SHA-256	94c9ceefd2fe268070e6521df2381b4e5c320017dc245f564a7f1027bb167b85

(a)

Original text	اذا وقعت الواقعة
Original bytes	d8a7d8b0d8a720d988d982d8b9d8aa20d8a7d984d988d8a7d9... (length=30)
Adler32	3f1714bf
CRC32	eb20db0d
Haval	eb77b8b126eb79f49eb4d613865a65b2
MD2	e0cf7b49a35fea6edd774c1be78ae66a
MD4	a7b1cbe7647ad7478e53a4d7a7838792
MD5	594a239d0af8fa6113e183bce300dab2
RipeMD128	01323d2e7b02c3406a4094ab5f867c3b
RipeMD160	419c524751068fdf617d626ef874dafad253e747
SHA-1	964c0decc70268f2a4d2c72b74e00aa96dba1ba4
SHA-256	c96aafa51f116570bb25bd580578c5c0c0a00639eaa17432a4ae825023b9152a

(b)

Figure 9 (a) Evaluating hashing values using Tashkeel. (b) Hashing values without Tashkeel.

Table 1 A sample hash of verses (including Tashkeel).

Verse	Hash1	Hash2	Hash3
56:1	937dbc692b153fec-d3bcf890121a54ea	578D50BD96B68D5-A315491A333419D11	578d50bd96b68d5a3-15491a333419d11
3:3	3cc96e9210508e78b-3ecae50f5295f63	C64AC514F615D3E4-3DA500959AFACFED	c64ac514f615d3e4-3da500959afacfed
9:4	8709dce3dcf2d41c47-52f76faa1c542a	17126C6DFA2CDE5344-5F017D079A6378	17126c6dfa2cde5344-5f017d079a6378

Table 2 Hash of verse (without Tashkeel).

Verse	Hash1	Hash2	Hash3
56:1	a1612dca935376349712c736c71b94b3	3EBC3778DB0FE1E6E27BA6C638D2CB55	3ebc3778db0fe1e6e27-ba6c638d2cb55
3:3	1cc9e35f76da3988f50c50c13bfec3fc	D2BFE01BFFD05BFFC3AF18139DDD4805	d2bfe01bffd05bffc3af18139ddd4805
9:4	16f463927f497032dd9b284ac9f20f0c	6BF0D7336E4203AE294E34CCE0DC5C14	6bf0d7336e4203ae294-e34cce0dc5c14

ings. Accordingly, we want first to evaluate whether such differences can be handled by hashing algorithms or not.

In the first study, we have selected an experimental case study of several different verses, different fonts, colors, and occurrences of small editing or changes and see their hashing values using different hashing algorithms.

We want to see first whether Tashkeels may impact hashing algorithms. Fig. 9(a and b) shows different hash algorithm values for the same small part of Tashkeel

Bytes conversion gave the same value for the part of the verse with and without Tashkeel as Tashkeel values are not recognized by the conversion. This byte value can be then used for verification if Tashkeel is to be ignored. All the evaluated hashing algorithms including the following: Alder32, CRC32, Haval, MD2, MD4, MD5, RipeMD128, RipeMD160, SHA-1, SHA-256, SHA-384, SHA-512, Tiger, and Whirlpool gave different values for the verses with and without Tashkeels. We repeated the process several times and noticed that all hashing algorithms for all evaluated verses show different hash values comparing with and without Harakat or Tashkeel. We noticed however, that different implementations for the same hashing algorithm may give different hash values for the same text. As such, the system should use the same algorithm implementation for initial and consecutive authentications or verifications. This is what we have done in developing the authentication system where we used the MD5 hashing algorithm for hashing purposes.

We also tried some small changes related to using for example (ه) instead of (ة) and all hashing algorithms showed different values.

This simple experiment showed that if we want to check correctness or authenticity of verses with Tashkeel and exact form, hashing algorithms can be used. However, if we want to stem those parts, hashing algorithms will not be then effective unless those additions are stemmed from all verses. In the second experiment and in order to check whether same hash functions can produce same hash codes, we tried different implementations for the same hashing algorithm and check hash values for several verses. We believe that verse should be the standard unit to measure with. Sizes of smaller or larger than a verse can have several issues and challenges specially related to false alarms.

In this experiment we have fixed the source of Quran text to one website, in this case, (http://quran.com). First, we evaluated using MD5 hashing from three websites: http://www.file-format.info referred to (Hash1), http://www.hashemall.com referred to (Hash2) and http://www.md5hashgenerator.com referred to (Hash3). Table 1 shows results of three samples using verses with Tashkeel.

Results showed inconsistent findings where two algorithms gave exact hash values (ignoring letter cases) unlike the third one.

Table 2 shows same verses and same implementations for the verses but without Tashkeel. All values are different from

Table 3 Hash of verse with different fonts and font sizes: verse 56:8: MD5 (without Tashkeel).

Font-size	Hash1	Hash2	Hash3
1	c3bc805e0a302bdf11591cf180a5d083	F1A223ED0DEC063DB88A916EF20405AE	f1a223ed0dec063db88a916ef20405ae
2	c3bc805e0a302bdf11591cf180a5d083	F1A223ED0DEC063DB88A916EF20405AE	f1a223ed0dec063db88a916ef20405ae
3	c3bc805e0a302bdf11591cf180a5d083	F1A223ED0DEC063DB88A916EF20405AE	f1a223ed0dec063db88a916ef20405ae

Table 4 Comparison of the hash code for some of the Quran verses from Qaloon version using our authentication system.

Q-verse	Hash MD5	Correct version	Hash MD5
إِنْ لَكُمْ فِيهِ لَمَا يَتَخَيْرُونَ	711522207422 381951060 1321301341 1313059119	إِنْ لَكُمْ فِيهِ لَمَا تَخَيْرُونَ	692913921 2491711 6387200246 23413536 15774149
وَلَوْلَا إِذْ دَخَلْتَ جَنَّتَكَ قُلْتَ مَا شَاءَ اللَّهُ لَا قُوَّةَ إِلَّا بِاللَّهِ إِن تَرَنِ أَنَا أَقَلَّ مِنكَ مَالًا وَوَلَدًا	137228453917919 3691054153 36120220 162142166	وَلَوْلَا إِذْ دَخَلْتَ جَنَّتَكَ قُلْتَ مَا شَاءَ اللَّهُ لَا قُوَّةَ إِلَّا بِاللَّهِ إِنْ تَرَنِ أَنَا أَقَلَّ مِنْكَ مَالًا وَوَلَدًا	109215698 193251 58821615221 42081591148673
وَقَاتِلُوا فِي سَبِيلِ اللَّهِ الَّذِينَ يُقَاتِلُونَكُمْ وَلَا تَعْتَدُوا إِنَّ اللَّهَ لَا يُحِبُّ الْمُعْتَدِينَ	38128941281 77153692381111 4025172191 70205176	وَقَاتِلُوا فِي سَبِيلِ اللَّهِ الَّذِينَ يُقَاتِلُونَكُمْ وَلَا تَعْتَدُوا إِنَّ اللَّهَ لَا يُحِبُّ الْمُعْتَدِينَ	42427067 351552 3024590161 391672 082015110

Table 5 Example of verse tampering discovered by the authentication system from another source.

Original	وَقَاتِلُوا فِي سَبِيلِ اللَّهِ الَّذِينَ يُقَاتِلُونَكُمْ وَلَا تَعْتَدُوا إِنَّ اللَّهَ لَا يُحِبُّ الْمُعْتَدِينَ
MD5	b73a3bde38abdda5f073482de2044787
الباحث	وَقَاتِلُوا فِي سَبِيلِ اللَّهِ الَّذِينَ يُقَاتِلُونَكُمْ وَلَا تَعْتَدُوا إِنَّ اللَّهَ لَا يُحِبُّ الْمُعْتَدِينَ
MD5	25dcb1ccfabd68ffcd42556e2a9dd991

those of Table 1 as Tashkeel changed hash values. On the other hand, Hash 2 and 3 gave same values (ignoring letter cases).

This experiment shows that not all hashing functions are identical or should always produce same hash values for same text. Nonetheless, this is not due to reliability issue with hashing algorithms. Rather, it has to do with the detail implementation of the hashing algorithm. Hashing values for the same algorithm implementation for the same text should always produce same hash values.

Table 3 shows results of changing text size and font. All three MD5 implementations gave same results and were not impacted with changing verse font type or size.

As a summary of hashing experiments, we found that the authentication systems including the indexer should be developed using the same hash function. This is since values of different hashing algorithms for the same exact verse may result or display different hashing values.

We decided to consider MD5 hashing algorithm due to its popularity and reliability. As mentioned earlier, it was important to select only one hashing algorithm as each algorithm produces different hashes in comparison with the others for the same verses (see Tables 4 and 5).

We have used our authentication system to check one of the online copies of the Holy Quran available in one of the Arabic websites that index Arabic books and documents (The comprehensive library: http://shamela.ws/), and we noticed that there are records of some Quranic distortion in some verses. The risk is that such versions may be exposed to non-Muslims or Muslims who have little experience of the exact wording of Quran verses.

One popular example of this is a popular program called (Qaloon). At least one version of this program is discovered with examples of incorrect verses' wording. Following are examples of the incorrect wording of some verses and their correct ones. We also included hashing values for both cases to simply show that an automatic hash based algorithm can easily detect such distortion.

Things to notice in the results:

1. Hash can detect distortion even in dialects or Tashkeels.
2. Hash values are randomly generated. Hence one difference may result in a completely different hash. Hence the idea is not to see how much the text is far from the actual text. The idea is to come up with a Boolean decision: Version is identical relative to the original authenticated copy or not.
3. One of the serious issues through the Internet is that if we search using the incorrect versions that Qaloon program used, you will see thousands of websites indexing these incorrect versions.

Qaloon is not the only program that had such incident. Another program called "الباحث" included some incorrect verses. Results showed that both showed completely different hashes.

6. Conclusion and future work

With the current continuous expansion of Internet information and services, authentication and credibility of users and information through the web is vital. For the holy Quran book in particular, authentication is very critical. The authenticity and originality of verses in the digital media can be challenging especially as it is required to be accomplished in high agile and quick manners. In this paper, we proposed an approach to implement Quranic authentication system based on information retrieval techniques and hashing algorithms. We conducted several experiments to evaluate the different characteristics that may impact the accuracy of the authentication process. Results showed that hashing verification can be a good candidate for the automatic authentication process with high confidence. Some exceptional cases such as the different Quranic readings are yet to be investigated and evaluated. Combining hashing methods with typical text plagiarism methods can be also another alternative approach for the automatic authentication process. We believe however, that such process is very important as well as applicable. A typical Quran indexer can include one table of verses in rows. Content of the verse, its number, chapter and character size can be the major attributes to include in the table or the Quran indexer.

A customized Quranic crawler is developed to search for Quranic verses through the Internet and index pages and information about websites that include those verses. We developed a case study to analyze Google top 10 search results for Quranic verses where we retrieved results for the verses one by one for the whole Quran. This was part of an initial assessment to evaluate the credibility of websites that include Quranic verses and that are most popular according to Google ranking algorithms.

Results showed that while some of the most frequently visited websites for verses search are authentic and credible, however, in many other cases, some verses are indexed by less credible websites according to some subjective metrics that we defined or according to the fact that some of those pages are related to personal or social networks that are managed by individuals without any official entity or establishment.

While we only conducted the current experiments on Quran verses, we believe that same process and tasks can be applied to evaluate Hadeeths as well. The case of Quranic verses can be considered easier as Quranic verses are literally identical even in the different Quran readings (with some few and minor exceptions). However, for the case of Hadeeths, some Hadeeths are mentioned in different wordings where this needs to be considered in the authentication evaluation process.

As a consequence of this study another research work will be conducted in future to rank the websites that cite or document the Quran verses based on their authenticity and provide the readers with the top most websites that document the authenticated Quran verses.

References

[1] Alfred J. Menezes, Paul Van Oorschot, S.A. Vanstone, Chapter hash functions and data integrity, in: Handbook of Applied Cryptography: The CRC Press Series on Discrete Mathematics and Its Applications, CRC Press, 1997, pp. 321–383.

[2] P.S. Tomar, M. Shreevastava, The study of detecting replicate documents using MD5 hash function, Int. J. Adv. Comput. Res. 1 (2011) 2277–7970.

[3] A. Alshareef, A.E. Saddik, A Quranic quote verification algorithm for verses authentication, in: International Conference on Innovations in Information Technology (IIT) Abu Dhabi, United Arab Emirates, IEEE, 2012.

[4] Y. Alginahi, O. Tayan, M. Kabir, Verification of qur'anic quotations embedded in online arabic and islamic websites, Int. J. Islamic Appl. Comput. Sci. Technol. 1 (2013) 10–16.

[5] Y. Yusoff, R. Ismail, Z. Hassan, Adopting hadith verification techniques in to digital evidence authentication, J. Comput. Sci. 6 (2010) 613–618.

[6] A.F. Shamsudin, A. Farooq, AI natural language in metasynthetics of Al-Qur'an, in: 2000 TENCON Proceedings: Intelligent Systems and Technologies for the New Millennium, Kuala Lumpur, Malaysia, 2000, pp. 464–467.

[7] M. Warasart, P. Kuacharoen, Paper-based document authentication using digital signature and QR code, in: 4th International Conference on Computer Engineering and Technology (ICCET 2012) Bangkok, Thailand, 2012.

[8] Kamsin, Amirrudin, Abdullah Gani, Ishak Suliaman, Salinah Jaafar, Rohana Mahmud, Md Sabri, Aznul Qalid et al., Developing the novel Quran and Hadith authentication system, in: Information and Communication Technology for The Muslim World (ICT4M), 2014 The 5th International Conference on, IEEE, 2014, pp. 1–5.

[9] Sabbah, Thabit, Ali Selamat, A framework for Quranic verses authenticity detection in online forum, in: Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences, 2013.

[10] Noor Jamaliah Ibrahim, Zaidi Razak, Zulkifli Mohd Yusoff, Mohd Yamani Idna Idris, Emran Mohd Tamil, Noorzaily Mohamed Noor, Abdul Rahman, Noor Naemah, Quranic Verse recitation recognition module for support in J-QAF learning: a review, Int. J. Comput. Sci. Netw. Secur. (IJCSNS) 8 (8) (2008) 207–216.

[11] Razak, Zaidi, Noor Jamaliah Ibrahim, Zulkifli Mohd Yusoff, Mohd Yamani Idna Idris, Emran Mohd Tamil, Quranic verse recitation feature extraction using mel-frequency cepstral coefficient (MFCC), in: 4th International Colloquium on Signal Processing and its Applications (CSPA 2008), 2008.

[12] Mohammed Ammar, Mohd Shahrizal Sunar, Md Sah Hj Salam, Quranic verses verification using speech recognition techniques, Jurnal Teknologi 73 (2) (2015).