Information Processing and Management 000 (2017) 1–14

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Approaches for preserving content integrity of sensitive online Arabic content: A survey and research challenges

Saqib Hakak[a,*], Amirrudin Kamsin[a], Omar Tayan[b], Mohd Yamani Idna Idris[a], Gulshan Amin Gilkar[c]

[a] Faculty of Computer Science & Information Technology, University of Malaya, Malaysia
[b] Department of Computer Engineering, College of Computer Science and Engineering, NOOR Research Center, Taibah University, Saudi Arabia
[c] Department of Computer Engineering, College of Computer Science, Shaqra University, Saudi Arabia

ARTICLE INFO

Article history:
Received 15 October 2016
Revised 27 July 2017
Accepted 17 August 2017
Available online xxx

Keywords:
Online sensitive content
Quran
Integrity
Watermarking
Cryptography
Steganography

ABSTRACT

Trends in Internet usage and accessing online content in different languages and formats are proliferating at a considerable speed. There is a vast amount of digital online content available in different formats that are sensitive in nature with respect to writing styles and arrangement of diacritics. However, research done in the area aimed at identifying the necessary techniques suitable for preserving content integrity of sensitive digital online content is limited. So, it is a challenge to determine the techniques most suitable for different formats such as image or binary. Hence, preserving and verifying sensitive content constitutes an emerging problem and calls for timely solutions. The digital Holy Qur'an in Arabic, constitutes, one case of such sensitive content. Due to the different characteristics of the Arabic letters like diacritics (punctuation symbols), kashidas (extended letters) and other symbols, it is very easy to alter the original meaning of the text by simply changing the arrangement of diacritics. This article surveys the different approaches that are presently employed in the process of preserving and verifying the content integrity of sensitive online content. We present the state-of-the-art in content integrity verification and address the existing challenges in preserving the integrity of sensitive texts using the Digital Qur'an as a case study. The proposed taxonomy provides an effective classification and analysis of existing related schemes and their limitations. The paper discusses the recommendations of the expected efficiency of such approaches when applied for use in digital content integrity. Some of the main findings suggest unified approaches of watermarking and string matching approaches can be used to preserve content integrity of any sensitive digital content.

1. Introduction

The field of analyzing sensitive online content is quite new and growing and requires significant research efforts in terms of identifying different methods that are suitable for checking and analyzing the performance of different methods suitable for different formats like text and image, and so on. Content integrity refers to the process of managing and assuring the

* Corresponding author.
E-mail address: saqibhakak@gmail.com (S. Hakak).

http://dx.doi.org/10.1016/j.ipm.2017.08.004
0306-4573/

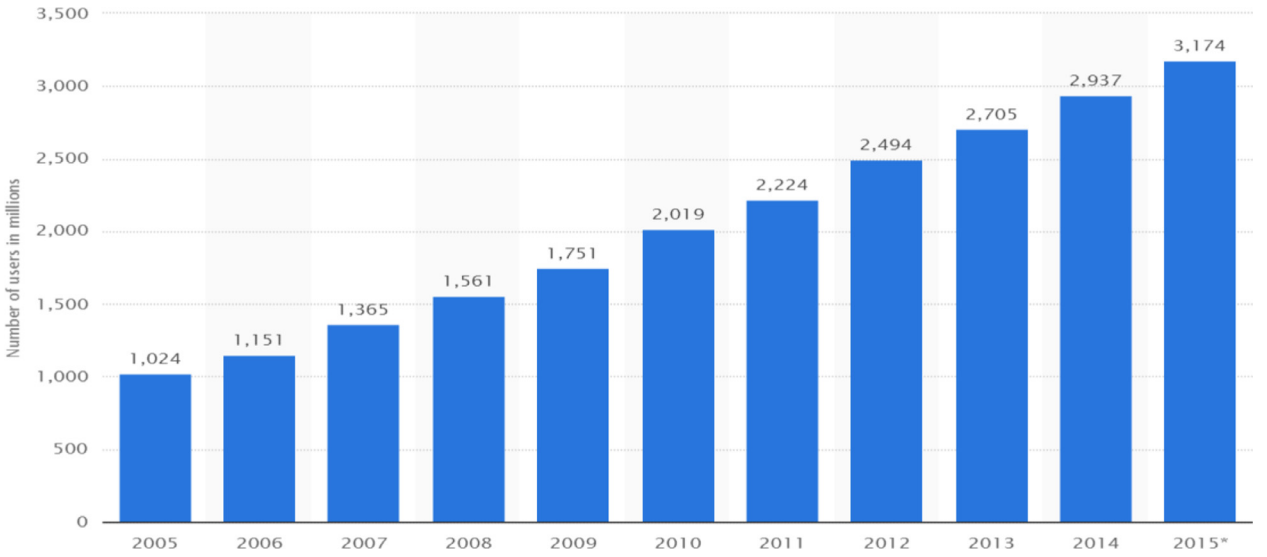

Fig. 1. Trend of increasing number of internet users per year.

accurateness and correctness of the content being transferred or being made available to end users. In order to protect sensitive information, one of the services of confidentiality, integrity and availability has to be deployed. There has been a massive increase in the use of digital content since the past few years, and issues of copyright violation, authenticity and integrity of digital content and data vulnerability then since dramatically increased (Chamlawi & Khan, 2010 ; Pan, Huang, & Jain, 2004). According to internet statistics, the number of internet users has increased fivefold as shown in Fig. 1.

With this trend in the steady increase of internet users, the publication of digital content online has also increased. There is a large amount of content available online in different formats like image format, text format, audio format and video format that requires adequate content integrity checking (Lai & Chang, 2011).

Online Digital Qur'an constitutes one type of such content that is most sensitive to tampering and is also available in different formats. The sensitivity lies in the fact that one single change of a symbol changes the meaning of a complete sentence. Different methods have been used to protect and verify the integrity of the Qur'anic content available online. The two most widely used methods include steganography and watermarking (Chang & Lin, 2006 ; Kee, Johnson, & Farid, 2011 ; Khare, Shivakumara, & Raveendran, 2014). However, there exist other approaches like cryptography, hashing and other sub-methods as well. The lack of proper classification in regard to these different approaches to verifying, authenticating and protecting the digital content has created a certain amount of confusion regarding the applicability of these different approaches.

There are 28 characters in the Arabic alphabet written from right to left. Arabic words can be connected such as مَيِقَتَسْنُم or a single word like نزام. Given the connections between different letters and pointed letters, it is very easy to tamper with words and change the meaning of a whole sentence (Mansour, Haraty, Daher, & Houri, 2008). Based on this fact, a review of different approaches used to protect and verify the integrity of the Digital Qur'an is presented. Future directions and challenges in the area of protecting and verifying content integrity of the Digital Qur'an are also discussed. A better understanding of the protection and verification processes of the integrity of sensitive online Qur'anic content can assist in identifying better and more efficient approaches. Herein lies the significance of this survey. To the best of our knowledge, this work is the first attempt to showcase all possible content integrity approaches together with their classifications based on their formats, methods and attacks.

Thus, briefly summarizing the contributions of this article, the objectives of this article are:

1. To present different approaches used to verify the integrity of sensitive Arabic language content available in text and image formats.
2. To propose new taxonomies based on objective 1, and to help researchers in identifying relevant techniques with respect to verifying the integrity of sensitive content.
3. To help researchers in unifying all converging trends found in the literature of content integrity into one standard approach.
4. To survey the existing methods to verify the integrity of one of the most sensitive Arabic content, namely the Digital Qur'an.
5. To present research challenges and offer recommendations for future research in the field of content integrity verification.

The organization of this paper is as follows: Section 2 describes the different types of online content available and proposes a classification based on the format of sensitive online content. In Section 3, a summary of different techniques that

Please cite this article as: S. Hakak et al., Approaches for preserving content integrity of sensitive online Arabic content: Aa survey and research challenges, Information Processing and Management (2017), http://dx.doi.org/10.1016/j.ipm.2017.08.004