Egyptian Informatics Journal (2014) **15**, 79–87

Cairo University

Egyptian Informatics Journal

www.elsevier.com/locate/eij
www.sciencedirect.com

FULL-LENGTH ARTICLE

An improved algorithm for information hiding based on features of Arabic text: A Unicode approach

A.A. Mohamed *

Salman Bin Abdulaziz University, Kingdom of Saudi Arabia

Received 1 August 2013; revised 29 March 2014; accepted 9 April 2014
Available online 19 May 2014

KEYWORDS

Information hiding;
Information security;
Arabic language;
Steganography;
Unicode

Abstract Steganography means how to hide secret information in a cover media, so that other individuals fail to realize their existence. Due to the lack of data redundancy in the text file in comparison with other carrier files, text steganography is a difficult problem to solve. In this paper, we proposed a new promised steganographic algorithm for Arabic text based on features of Arabic text. The focus is on more secure algorithm and high capacity of the carrier. Our extensive experiments using the proposed algorithm resulted in a high capacity of the carrier media. The embedding capacity rate ratio of the proposed algorithm is high. In addition, our algorithm can resist traditional attacking methods since it makes the changes in carrier text as minimum as possible.

1. Introduction

With the rapid growth and wide application of Internet and smart devices in recent years, the network delivery of large bulk of novels, journals and documents in digital way has become more popular. Most of this kind of information can be categorized or directly converted to text formatting so that they are easily copied. Thus, how to protect the transmitted information effectively on the internet has become more important. In addition, the text steganography has become an effective way to solve such a problem very well.

Steganography is a Greek word coming from cover text. “Stegano” means hidden and “Graptos” means writing [1]. In steganography, the secure data will be embedded into another object; so the eavesdropper cannot catch it which literally means “covered writing”.

Obviously the purpose of steganography is to avoid drawing suspicion to the transmission of hidden information. A message is a hidden information in the form of plain text, cipher text, images or anything that can be encoded into a bit stream. This message is embedded in a cover-carrier to create a stego-carrier. A possible formula of the process may be represented as follows:

$$\text{Stego medium} = \text{Cover medium} + \text{Embedded message} + \text{Stego key}$$

* Tel.: +966 201023303373.
E-mail address: dr_ashrafa@yahoo.com

Peer review under responsibility of Faculty of Computers and Information, Cairo University.

Production and hosting by Elsevier

1110-8665 © 2014 Production and hosting by Elsevier B.V. on behalf of Faculty of Computers and Information, Cairo University.
http://dx.doi.org/10.1016/j.eij.2014.04.002

Different types of cover media including image, Sound, video as in [2–4] and text can be used in steganography. Choosing carrier file is very sensitive as it plays a key role to protect the embedded message.

But using text is preferred over other media, because the texts occupy lesser space, communicate more information.

Text steganography represents the most difficult type as there is generally lack of data redundancy in the text file in comparison with other carrier files. The existence of such redundancy can help increase the capacity of hidden data size.

Furthermore, text steganography depends on the language as each language has its own unique characteristics, which are completely different from other languages. For example, the letter shape in English language does not depend on its position in the word. At the same time, Arabic letters have different forms which depend on letter positioning.

After data embedding, the text with secrets which is called stego-text, is sent from sender side to receiver side over the Internet. The security concept is central around the idea that no one can easily discover the secrets embedded into the stego-text by using statistical computation or other methods of detection.

There are three important parameters (criteria) in designing steganography systems [5]:

I. Perceptual transparency.
II. Robustness.
III. Hiding capacity.

I. The security or Perceptual transparency refers to the ability of an eavesdropper to figure, or suspect the hidden information easily. We can achieve high security by minimizing the embedding Impact (distortion). Intuitively, we can try to achieve this by minimizing the distortion between the cover text and stego text and by restricting the distortions to the portions of the cover text that are difficult to model.

II. The robustness refers to the ability of protecting the unseen data from corruption especially when transmitted through the internet.

III. The capacity or the embedded rate refers to the ability of a cover media to store secret data, which can be measured by the amount of secret data (bits) that can be hidden in a kilo byte of a cover media. It is used to measure the hiding efficiency and is defined by Eq. (1) as follows:

$$\text{Embedded rate(Capacity Ratio)} = \text{Size of hiddeng secrete in bits/} \text{the size of carrier in kilobytes.} \tag{1}$$

One reason of why text steganography is difficult is that text contains little redundancy compared to other media. Another is that humans are sensitive to abnormal-looking text. Text steganographic schemes must be specifically designed to exploit the specific characteristics of the target language, because the grammatical and orthographic characteristics of every language are different.

The following parts of the paper are structured as follows. Section 2 reviews the text steganography methods for Arabic text and the related works. The proposed method is detailed in Section 3. The experiments and discussions are presented in Section 4. Some conclusions are made in Section 5. The future work is given in Section 6.

2. Related works

Most of the text steganography methods are applied to English texts. However, there are a few text steganography methods applied to other languages [1,6,7]. A few works have been done on hiding information in Arabic texts [8–10]. The following is a list of different methods of the works for Arabic text carried out and reported thus far.

2.1. Steganography by shifting points

In Arabic languages, dot is very important and 14 of 28 Arabic letters have one or more dots. In this method, data are hidden in Arabic texts by using these letters as explained in [8,11]. Using the approach of the vertical displacement of the points, we can hide information in the texts. However, we need to use a special font created mainly for this purpose.

2.2. Steganography using Arabic diacritics (Harakat)

Arabic language uses different marks or diacritics, Arabic extension character, (Harakat) which are optional to use. The main reason to use these symbols is to distinguish between words that have same letters (i.e., so that every word pronounced differently). The diacritic such as "Fatha" is used to hide 1. However, the rest of the diacritics are used to hide a 0 because it was found out that "Fatha" represents almost half of the diacritics in any Arabic text as explained in [9,12]. The main disadvantage of this method is that it attracts the attention of the reader.

2.3. Kashida based steganography

In this method, we add extension (Kashida in Arabic) to a word to hold secret bit 'one' and we leave the word without extension (Kashida) to hold secret bit 'zero' as explained in [10,11,13]. Note that letter extension does not have any effect on the writing content. The main disadvantage of this method is that it attracts the attention of the reader, increases the file size and changes the apparent of the text.

2.4. Unicode based steganography

Arabic letters have many forms according to Unicode standard. It is divided into two groups of codes for Arabic letters; the representative code and the code of the possible shapes of the letter. In this method, we use the different possible Unicode values of the same letter to hide bits as explained in [14,15,5]. This method is suitable for internet and modern device as smart phone. But it is weak against traditional attacks. Some techniques of Unicode based steganography provide high capacity and low security [9,11] and vice versa.

The disadvantages of Unicode based steganography can be summarized as follows: