Linguistic Grammar Approach to Textual Steganography

Conference Paper · March 2017

CITATIONS
0

READS
229

2 authors, including:

Smriti Priya Medhi
Assam Don Bosco University
9 PUBLICATIONS **11** CITATIONS

SEE PROFILE

Proceedings of the 11th INDIACom; INDIACom-2017; IEEE Conference ID: 40353
2017 4th International Conference on "Computing for Sustainable Global Development", 01st - 03rd March, 2017
Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM), New Delhi (INDIA)

Linguistic Grammar Approach to Textual Steganography

Urjashee Shaw
Student, Department of Computer Science and Engineering & IT, Assam Don Bosco University
Guwahati, India
Email Id: urjashee09@gmail.com

Smriti Priya Medhi
Assistant Professor, Department of Computer Science and Engineering & IT, Assam Don Bosco University
Guwahati, India
Email Id: smriti.medhi@dbuniversity.ac.in

Abstract - Text Steganography is the technique of concealing secret or sensitive data within some text. Sending encrypted data risks drawing attention of hackers and crackers, where they might attempt to crack and reveal the original message. Steganography has gained prominence in the last decade due to the constant need for data concealment. The communicated data or messages can be viewed by an attacker in the network, so work of steganography is to hide these communicated data without giving away the fact that sensitive data is hidden behind them. Various steganography techniques has been proposed in the past, but the problem still remains in hiding text behind some text. The model proposed in this paper reads binary data and uses a dictionary and a style source to generate an innocuous text file that can later be converted back, with the help of the dictionary, to the original data. The dictionary contains valid words, classified according to type.

***Keywords – Cryptography, Plain Text, Stego Text, Steganography* .**

NOMENCLATURE

ASCII – American Standard Code for Information Interchange.

I. INTRODUCTION

In this digital age it is imperative to secure our information. Many encryption techniques has been developed to solve the problem of securing data. Now a day it is almost impossible to break encryption schemes. However the use of encryption implies that the transmitted data is important. To any hacker transmitting encrypted data is enough to incite suspicion. Encrypted data can be tampered and disrupt the communication process between two parties. The idea is to accept other solutions to information security [3]. This need has led us to look for Steganography techniques, which is another solution to information security but with the exception that any hacker or third party is impervious to any form of secret data exchange.

Steganography is the technique or rather the art of hiding secret messages instead of encrypting it. The work of steganography is such that the existence of secret message is known to the sender and the intended receiver only. The Greek words "stegos" and "grafia" meaning "covered writing" has coined the term Steganography.[1] People might confuse steganography for cryptography since they are related terms but the work of cryptography is to scramble original data so that it becomes illegible to any third party, whereas steganography tries to hide the message behind a text, image, sound or video. Most of the steganography works has been centered around video and images due to the large amount of redundant data in their file format. However text Steganography is the most difficult kind of Steganography; because of the fact that there are no redundant data in text files, as opposed to video and sound files which have a lot of redundancy which Steganography can take advantage of. A text document when send by a sender can be recovered in the exact same format by the receiver, which is not the case for image, sound or video files. Text documents are the same as they appear to the naked eye but other file formats have the ability to hide data. [7]

In text Steganography the hidden message is called the secret message. The text file that will hide the data is called cover file or cover document. The file containing the hidden message is called the stego-document. The entire process of hiding and extracting hidden message is called stego-system. [2]

II. TEXT STEGANOGRAPHY

Below mentioned are some of the techniques proposed by various researchers and authors to implement text steganography.

A. Text Steganography Types

1) Line-Shift Coding

 ISSN 0973-7529; ISBN 978-93-80544-24-3

In Line-Shift Coding method bits are added to the document to hide the secret message by vertically moving in a paragraph. The hidden message can be later extracted from the cover file. [6]

2) Word-Shift Coding

In word-shift coding words are shifted horizontally to hide a secret message in a cover document. The space between the words must be different in order to apply this particular method. Disadvantage to this method is that the original document would be required to crack the hidden message. [6]

3) Format-based methods

In Format-based method the format of the cover document is changed to hide data. The contents of the cover documents remain exactly the same. Here white spaces are used to hide secret texts. Bit "0" is represented by a single white space and bit "1" is represented by two white spaces. The advantage of this method is that it can be applied to all types of text format. A disadvantage is that very little data can be hidden. [7][8]

4) Random and statistical generation methods

This method uses statistics to generate a stego document containing the secret message. This method exploits the properties of a particular language and statistically predicts the next word. A Context Free Grammar (CFG) language model is used with probability associated with it. The first word start at the root and random rules are applied for the next words. [7][8] Disadvantage of this method is that it works on probability, there is no guarantee that you will retrieve the correct message.

5) Linguistic method

Linguistic method uses the linguistic structure of a language to hide secret text. There are two types of linguistic method- Syntactic method and Semantic method. Syntactic method uses special characters (, . ? / ; : " ") to hide secret data in the cover document. Semantic method uses synonym of words for some pre-selected words. [7][8]

6) White Steg

White Steg as the name suggests uses white spaces for hiding messages. Inter Sentence Spacing places white spaces after each character. Single white space for bit 0 and two white spaces for bit 1. Placing white spaces after each line is End of Line Spacing. Inter Word Spacing is similar to Inter Sentence Spacing, where instead of putting white space after characters we put white spaces after each word. Single white space for bit 0 and two white spaces for bit 1. [8]

7) Spam Text

HTML tags of HTML files are also used to hide data. 0 is inferred when the tags are different, meaning starting and ending tags are not similar, 1 is inferred when tags are similar. Lack of space in a tag is inferred as 0 and presence of a space is inferred as 1. [8]

8) SMS-Texting

In SMS Texting method binary data is represented by its shortened form or full form to hide data. Whenever full form of a word is used it is inferred to as bit 0 and whenever shortened form is used it is inferred as 1. For convenience sake a dictionary is maintained to map shortened words to corresponding full form. [8]

B. Previous Works

- S. Roy and M. Manasmita in their paper [1] proposed a hybrid model combining Line Shifting, Word Shifting and use of Special Characters techniques. In Line-Shift Coding method bits are added to the document to hide the secret message by vertically moving in a paragraph. In word-shift coding words are shifted horizontally to hide a secret message in a cover document. The space between the words must be different in order to apply this particular method. Information is hidden in binary form. [1]

- M. Garg in the paper [2] proposed a unique method of using both the primary and secondary attribute of a HTML tag to hide a message. Ordering of attributes of any HTML tag does not change the appearance in the webpage, so this property can be exploited to hide secret messages. HTML tags have primary and secondary attributes, for e.g. *src* is the primary attribute and *alt* is the secondary attribute for tag *img.* So if a tag has both the primary as well as the secondary attribute, it is inferred as bit 1 whereas if a primary attribute is not followed by any secondary attribute, it is inferred as bit 0. [2]

- M. Morran and G. Weir in their paper [4] proposed a unique method that uses the grammar of a language for the basis of substitution. The proposed method creates a cover file that tags parts of speech to form a grammatically correct sentence. From this cover files select some particular words to perform substitution. Substitution should not hamper the correctness of the sentence. These substituted words are mapped to particular letters that form our secret message. The parts of speech targeted for substitution are nouns, verbs, prepositions and determiners are avoided. [4]

- I. Banerjee *et al* in their paper [5] proposed a method where the secret message is first encrypted using a SSCE (Secret Steganography Code for Embedding) method that uses a SSCE table to encrypt the secret text. This secret message is

embedded into the cover file using the AMT (article Mapping Technique) method. Starting with the first two bits of the cover file use the AMT mapping technique to create the stego-document. At the receiver side perform message extraction and text decryption to get the secret message. [5]

III. PROPOSED WORK

This proposed algorithm uses parts of speech to form the cover document and words are mapped to some predefined dictionary to form the stego document. This method reads binary data and uses a dictionary to generate an innocuous text file that can later be converted back, with the help of the dictionary, to the original data. The dictionary contains valid words, classified according to type. Mapping words from the dictionary is followed by the conversion of a secret text to meaningful sentences of the format "Article-Noun-Verb-Article-Object".

For example we take the word "Hi". ASCII value of H is 0100 1000 and ASCII value of i is 0110 1001. We get a concatenated16 bit as 01001000 01101001.

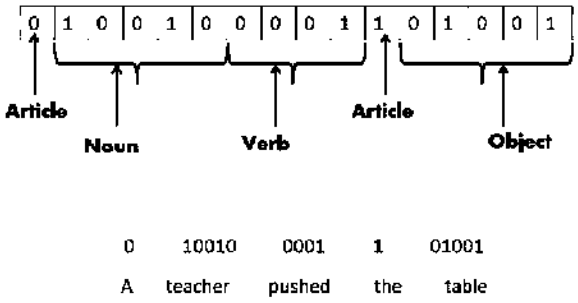

Fig. 1. An example showing bit pattern replacement

Here "Hi" becomes "A teacher pushed the table". We replaced the values using a predefined dictionary of words. The number of words included in each dictionary type must be of power of 2. If there are 5 bits then the number of words for that dictionary will be 2^5, which is equal to 32 words. There are two steps needed: (1) convert the secret message into some cover document called the stego-document. This step is done by the sender and (2) convert the stego-document to the secret message by the intended receiver.

A. *Converting Secret message to Stego-document*

Secret_msg_to_stego()

Step 1: Write the secret text.
Step 2: Extract ASCII values of each character of the secret text.
Step 3: Combine the ASCII values in groups of two characters to get a 16 bit concatenated value. White spaces are considered as a character.
Step 4: Assign an article to the 1st bit, a noun to the next 5 bits, i.e., from 2nd to 6th bit, assign a verb to the next 4 bits,i.e., from 7th to 10th bit, an article to the 11th bit and an object to the next 5 bits,i.e., from 12th to 16th bit.
Step 5: Repeat steps 2 to 4 till the end of string or end of secret message.
Step 6: In case the total number of characters are not even add an additional character to the end of the string.
Step 7: End algorithm.

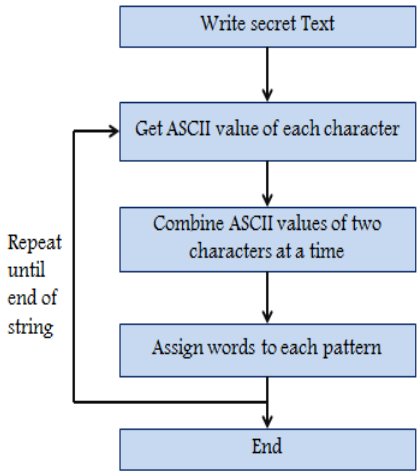

Fig. 2. Hiding Process

B. *Converting Stego-document to Secret message*

Stego_to_secret_msg()

Step 1: Write the stego-document.
Step 2: Repeat steps 3 to 6 for each sentence.
Step 3: For each word find the corresponding bit patterns.
Step 4: Combine these bit patterns to form binary number of 16 bit.
Step 5: Divide this 16 bit binary number into 2 groups of 8 bit.
Step 6: For each of this 8 bit find its corresponding ASCII value. Replace this ASCII value with its corresponding character.
Step 7: End algorithm.

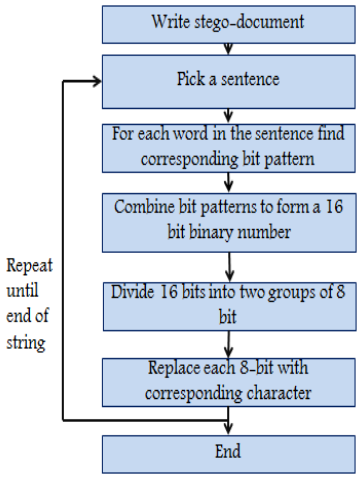

Fig. 3. Extracting Process

C. Data dictionary for NOUN

Following is the data dictionary that we use to replace the bits that correspond to the noun part.

TABLE I. NOUN DICTIONARY

Noun Dictionary	
Bits	*Words*
10000	Friend
10001	Woman
10010	Father
10011	Mother
10100	Man
10101	Dog
10110	Bird
11000	Teacher
11001	Neighbors
11010	Lady
11011	Boy
11100	Girl
11101	Professor
11110	Worker
00001	Dancer
00010	Singer
00011	Swimmer
00100	Cat
00101	Driver
00110	Assistant
01011	Manager
01000	Person

D. Data dictionary for Verb

Following is the data dictionary that we use to replace the bits that correspond to the verb part.

TABLE II. VERB DICTIONARY

Verbs Dictionary	
Bits	*Words*
0001	Pushed
0101	Scratched
1001	Painted
1101	Washed
0000	Climbed
1000	Poked

E. Data dictionary for Object

Following is the data dictionary that we use to replace the bits that correspond to the object part.

TABLE III. OBJECT DICTIONARY

Object Dictionary	
Bits	*Words*
10000	Ball
10001	Gate
10010	Bat
10011	Hat
10100	Mat
10101	Cat
10110	Dog
11000	Horse
11001	House
11010	Rat
00001	Table
00010	Chair
00011	Balloon
00100	Basket
00101	Car
00110	Book
00111	Box
01000	Bag
01001	Scarf
01010	Glass
01011	Cap
01100	Shoe
01101	Tray
01110	Painting
01111	Desk
00000	Cupboard

F. Data dictionary for Articles

Following is the data dictionary that we use to replace the bits that correspond to the articles part.

TABLE IV. ARTICLE DICTIONARY

Articles Dictionary	
Bits	*Words*
0	A
1	The

II. RESULTS AND ANALYSIS

The platform used for developing this algorithm is NetBeans IDE 7.1.2. For any given secret text the program successfully generate a string of text called the stego-document. Some of the input output strings are.

INPUT STRING	OUTPUT STRING
Hello	A father pushed the car. A boy pushed the shoe. A boy washed the horse.
Attack	A friend scratched the mat. A professor pushed the table.
GOOD BYE	A teacher washed the cap. A woman washed the desk. A mother washed the basket. A person pushed the chair. A bird scratched the car.
HELLO there	A father pushed the car. A mother pushed the shoe. A mother kicked the cupboard. A professor pushed the bag. A neighbor scratched the bat. A neighbor scratched the horse.

V. CONCLUSION

Text Steganography has a bright prospect due to the fact that it can hide data without the attacker knowing about the existence of it. In this paper we have presented a method where we use the ASCII values of each character to hide information. The basic idea behind this method is to hide groups of two characters from the hidden text into a sentence. The sentence is constructed by matching binary patterns of the corresponding characters. The binary patterns follow a dictionary of words which are replaced when a binary pattern matches. The sentences which hide the information are a part of the cover document known as stego-document. This part is known as data hiding, which is done by the sender. The second part is message extraction, which will try to extract the hidden message from the stego-document. Extraction can be done by the receiver only if he has the necessary dictionary of words.

REFERENCES

[1] Sangita Roy and Manini Manasmita, "A Novel Approach to Format Based Text Steganography," *Research Gate Publication No. 220846511, Conference Paper January 2011.*

[2] Mohit Garg, "A Novel Text Steganography Technique Based on Html Documents," *International Journal of Advanced Science and Technology Vol. 35, October, 2011.*

[3] Debnath Bhattacharyya, Poulami Das, Samir Kumar Bandyopadhyay, and Tai-hoon Kim, "Text Steganography: A Novel Approach," *International Journal of Advanced Science and TechnologyVol. 3, February, 2009.*

[4] Michael Morran and George R.S. Weir, *"An Approach to Textual Steganography"*, Department of Computer and Information Sciences, University of Strathclyde, Glasgow.

[5] Indradip Banerjee, Souvik Bhattacharyya and Gautam Sanyal, "Text Steganography using Article Mapping Technique(AMT) and SSCE," *Journal of Global Research in Computer Science, Volume 2, No. 4, April 2011.*

[6] Masoud Nosrati Ronak Karimi and Mehdi Hariri, "An introduction to steganography methods," *World Applied Programming, Vol (1), No (3), August 2011, pp. 191-195.*

[7] Shraddha Dulera, Devesh Jinwala and Aroop Dasgupta, "Experimenting with the novel approaches in text steganography," *International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.6, November 2011.*

[8] Monika Agarwal, "Text steganographic approaches: a Comparison," *International Journal of Network Security & Its Applications (IJNSA), Vol.5, No.1, January 2013.*

[9] Wesam Bhaya, Abdul Monem Rahma and Dhamyaa AL-Nasrawi, "Text steganography based on font type in MS-Word documents," *Journal of Computer Science 9 (7): 898-904, 2013.*

[10] [10] Chand, V., Orgun, C.O. "Exploiting linguistic features in lexical steganography: design and proof-of-concept implementation," *Proceedings of the 39th Annual Hawaii International Conference on System Sciences, vol. 6, p. 126b. IEEE, Los Alamitos (2006).*

[11] William Stallings, *Cryptography and Network Security: Principles and Practice 5/e.*, India, Prentice Hall, 2011.

[12] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding", *IBM Systems Journal*, vol. 35, Issues 3&4, 1996, pp. 313-336.

[13] M.Chapman, G. Davida, and M. Rennhard, "A Practical and Effective Approach to Large-Scale Automated Linguistic Steganography", *Proceedings of the Information Security Conference,* October 2001, pp. 156-165.