A Greedy Approach to Text Steganography using Properties of Sentences

[1]S. Changder, [2]N. C. Debnath, [3]D. Ghosh
[1]Department of Computer Applications, National Institute of Technology, Durgapur, India
[2]Department of Computer Science, Winona State University, Winona, MN, USA
[3]Department of Computer Science & Engineering, National Institute of Technology, Durgapur, India
suvamoy.nitdgp@gmail.com, Ndebnath@winona.edu, profdg@yahoo.com

Abstract – **Steganography is the art and science of covered or hidden writing. The purpose of steganography is covert communication to hide the existence of a message from an intermediary. Digital Steganography algorithms have been developed by using texts, images and audio etc as the cover media. This paper presents a new approach on text steganography through Indian Languages. Considering the properties of a sentence such as number of words, number of characters, number of vowels etc and using the presence of redundant feature code able characters in Indian Languages, this approach hides the message into an innocent cover file containing Indian texts. This approach also presents the extraction of message from the generated cover file by applying the reverse method of hiding. The approach shows satisfactory results on applying to some topic of daily newspaper in Indian Languages like Bengali.**

Keywords – Information Security, Text Steganography, Information Hiding, Text watermarking, Indian Text Steganography.

I. INTRODUCTION

Steganography means covered or hidden writing. The principle of steganography is secret communication to hide a message from an intermediary. This differs from cryptography, the art of secret writing, which is intended to make a message indecipherable by an unintended receiver but does not hide the existence of the secret communication. Although steganography and cryptography are different and distinct, these two can be treated as twin sisters of secret communications. As the app1ication of computer in rea1 life is increasing day by day, the need to secure data is becoming more and more essential and challenging part of message or data transfer and hence the hidden exchange of information has attracted more attention to the researchers.

Steganography is the art and science of hiding information such that its presence cannot be detected [1]**.** In steganography, the secret message is encoded in such a way that information's existence kept hidden from the unintended receivers. The aim of steganography is to establish a secured communication in a absolutely unnoticeable manner [2] and to avoid drawing suspicion to the transmission of a hidden data [3]**.** The steganography is not to keep others from knowing the hidden information, but it is to keep others from drawing suspicion that the information even exists. If a steganography method causes someone to suspect that there is secret information in a carrier medium, then the method has failed [4]**.**The first written evidence about steganography being used to send messages is the Heredotous[5] story about slaves and their shaved heads. Although steganography is an ancient subject, the modern formulation of it is often given in terms of the prisoner's problem [6].

A number of steganography methods have been introduce on different cover media such as images [2,4,7], video files[8,9] and audio files[10]. Due lack of large scale redundancy of information in a text file, in compared to others, text steganography seems to be most difficult kind of steganography [11].

This paper presents a new approach on text steganography through Indian Languages. Considering the properties of a sentence such as number of words, number of characters, number of vowels etc and using the presence of redundant feature code able characters in Indian Languages, this approach hides the message into an innocent cover file containing Indian texts. This approach also presents the extraction of message from the generated cover file by applying the reverse method of hiding. The approach shows satisfactory results on applying to some topic of daily newspaper in Indian Languages like Bengali.

II. RELATED RESEARCH WORKS

Considering the syntactic structures of a text, the syntactical steganography approach build syntactically correct sentences by using the Context Free Grammars (CFG). CFG based Mimicry [12], NICETEXT [13] comes under this category. Though, NICETEXT produces syntactically correct sentences, the output text is almost always set of ungrammatical and semantically anomalous sentences. Using this disadvantage of NICETEXT steganalysis algorithms [14] have been developed to detect the presence of hidden information in the cover file generated by NICETEXT.

In lexical Steganography lexical units of natural language text such as words are used to hide secret bits. . In this approach a word could be replaced by its synonym and the choice of word to be chosen from the list of synonyms would depend upon secret bits.

In Ontological steganography method, to embed information, instead of implicitly leaving semantics intact by replacing only synonymous words an explicit model for the meaning is used to evaluate equivalence between texts. This method is also having the same disadvantage like

978-0-7695-4367-3/11 $26.00 © 2011 IEEE
DOI 10.1109/ITNG.2011.13

NICETEXT that sometimes it may produce semantically incorrect texts.

In Text Steganography by Hiding Information in Specific Character of Words [15] approach, specific characters from some particular words are selected to hide the information. For example, the first character of every alternative word hides the secret message.

Text Steganography by Line Shifting method [16, 17] is another useful approach where lines are shifted vertically to some degree. For example, lines are shifted vertically to degree say α or – α. For α, the information is 1 and for – α, the information is 0. This method is appropriate for printed text.

Information hiding in Random Character and Word Sequences method [18] generates a random sequence of characters or words to hide the information.

In Text Steganography by Word Shifting method [16, 19], information is hidden in the text by shifting words horizontally and by changing the distance between the words.

Text Steganography by Feature Coding method [20, 21] changes the feature or structure of the text to hide data. For example, elonging or shortening end portion of some characters, or by vertical displacement of points of characters like 'i', 'j' etc. In this method a large volume of data can be hidden in the text.

In Text Steganography by adding Open Spaces method [22], the information can be hidden by adding extra white spaces in the text.

Information can be hidden by Creating Spam Texts [18] in a HTML file. This approach uses the flexibility of HTML regarding case-sensitiveness. In HTML the tags
,
,
 are equally valid. So by changing case, one can hide information in a HTML documentation text. But in WML this method will not work since all the tags are in lower case letters.

Besides all these, some algorithms for text steganography through Indian Languages have been proposed by using feature coding method [23, 24, 25, 26] and dynamic programming method [27] etc.

III. PROPOSED ALGORITHM

Let P be a property of a sentence, e.g. number of characters, number of words etc. Now we can assign a value to a sentence S as:

Value (S) = 0, if P is even, and
= 1, if P is odd.

Therefore, we can represent a binary digit by a sentence following the desired property. For example, we can represent 0(zero) by the sentence "if P is even", as the number of characters present in the sentence is even (including spaces). Similarly we can represent 1(one) by the sentence "if P is odd", since the number of characters present in the sentence is odd. By following the same, we can represent m bits by a sentence which is following m number of desired properties. To explain this let us consider three properties of a sentence say P_C, P_W, P_A as number of characters, number of words and sum of ASCII values of the characters in the sentence respectively, we call this as a property array, P. P can be represented by fixed length code i.e. P_C, P_W and P_A can be represented by 00, 01, 10 respectively, and let us assume this coding is shared publicly with the sender. Therefore to represent a binary stream of 3-digits say '010' we will select a sentence where the number of characters, number of words and sum of ASCII values of the characters are even, odd and even respectively.

So, to hide a secret message i.e. a stream of binary digits of length n (considering n = b X m, if not, the message can be padded with some false bits), we can choose b number of sentences, each following m number of desired properties, say $S_1, S_2, \ldots, S_b$ from a text called cover media . This cover media is sent to the receiver.

In the receiver side, during extraction of the message from the cover media, the job of the receiver is divided into following tasks:

- to locate the sentences $S_1, S_2, \ldots, S_b$, containing the blocks of hidden bits.
- to place the block of bits in its proper position.

Since the receiver is nothing except the cover file, therefore, to accomplish the above mentioned tasks the sender have to follow some locating and indexing mechanism during the time of encoding or hiding the data. To solve the problem of locating, we have used the feature coding method, that is, if a sentence S is selected by the encoding algorithm then the feature coding method will be applied to a selected character (decided by the indexing algorithm as discussed later) of the sentence, so that during extraction of the secret message the receiver can locate the sentence S by the presence of feature coded character and to solve the indexing problem we have used the ordering of the feature code able characters as they appears in the particular Indian Language. For indexing the sentences let us define a ordered set $C_{ORGINAL}$ as the collection of feature code able characters (individual and with medial vowels) of the particular Indian Language and a set $C_{FECODED}$ as the collection of modified characters of $C_{ORIGINAL}$, maintaining the same order as they appear in $C_{ORIGINAL}$. Now let us define the Indexing scheme as follows:

$Index(c_i) < Index(c_j)$,
$Index(c_i) < Index(c_ic_i) < Index(c_ic_j) < Index(c_ic_ic_i) < \ldots..$ so on,
where $c_i, c_j \in C_{ORIGINAL}$ and $i<j$.

Therefore, during encoding, if an unused sentence follows the properties represented by the block of binary digits, the minimum indexed, unused character present in that sentence will be replaced by its corresponding feature coded character so that at the time of extraction the receiver has to find the feature coded characters to locate the sentence and to place the blocks, extracted from the properties of that sentence, in sorted order based on the Index of the feature coded characters.

Followings are the algorithms for Hiding and Extraction of the secret message:

Algorithm HIDE:

1) Convert the secret message M to its equivalent binary say M_B.

2) At the beginning of M_B concatenate binary equivalent of $|M_B|$ to get a new binary stream M_{BN}.
3) Determine the size of the property array |P| and concatenate binary conversion of |P| at the beginning of the property array to get P_N.
4) Select a text file T_{COVER}, containing Indian scripts and sufficient enough to hide all the blocks, as the cover media.
5) for i =1 to |P|+8 // first 8 bits to hide the binary conversion of |P| //
6) Scan P_N and apply feature coding method to T_{COVER}, i.e. if $P_N[i] = 0$ the concerned character of T_{COVER} remains unchanged and if $P_N[i] = 1$ then replace the character with the respective feature coded character.
7) end of for loop
8) Divide M_{BN} into blocks, $b_1, b_2, \ldots, b_i$, each of size |P| (pad with required number of 0 and/or 1, if require, for the last block to maintain equal block boundary).
9) For each block in step 8 scan T_{COVER} from the next sentence where step 6 terminates to find a sentence S,
   a) *That does not contain any element belongs to the set* $C_{FECODED}$.
   b) *That follows the properties represented by the bit stream of the block.*
   c) *That contains the desired character(s) of the element* $e_i \in C^*_{ORIGINAL}$.
10) If Found(S), replace character(s) of $e_i \in C^*_{ORIGINAL}$ of S by the corresponding characters of $f_i \in C^*_{FECODED}$, and mark e_i as not reusable. Else search for the next character(s) of the element $e_{i+1} \in C^*_{ORIGINAL}$ in S and so on until $C^*_{ORIGINAL}$ is exhausted (in worst case, when no element from $C_{ORIGINAL}$ is present in the sentence).
11) Once $C^*_{ORIGINAL}$ is exhausted, repeat step 9 and step 10 to search for another sentence with the same block.
12) End.

Algorithm EXTRACT:

1) Scan $T_{GENCOVER}$ and find first eight feature code able characters and place the code bits to get the size of the property array |P|.
2) for i = 1 to |P|
3) Scan $T_{GENCOVER}$, starting from the next character where step 1 terminates, to find the corresponding code bits of the feature code able characters to get the property array.
4) end loop
5) Scan $T_{GENCOVER}$, starting from the next sentence where step 3 terminates, to find the sentences containing feature coded character(s). Place the code bits, found by comparing the properties of sentence with the property array, as per the index value (sorted in ascending order) of the feature coded character(s).
6) From the first seven bits find the length, $|M_B|$, of the message and discard the all the bits after $|M_B|$.
7) Convert the code bits to its character equivalent and get the original message.
8) End.

IV. Advantages and Disadvantages

The advantage of the method is that a large volume of data can be hidden. Since we are hiding the data with minimal change of the structure of the cover file therefore it will draw less attention to the unintended recipients.

Since for each block we are searching the whole file therefore time complexity will be much higher and that may be considered as a disadvantage to the system.

V. Experimental Results

We have discussed our experiment by dividing into Hiding and Extraction of the message. Each division shows Inputs, Outputs and the intermediate steps or results of the corresponding algorithms.

A. Algorithm HIDE

1) Input

The secret message M = 'DO'.

The cover text T_{COVER} = Text file in Bengali(an Indian Language), shown in Figure 1, containing a portion of a topic of Sports(Football) taken from a daily Bengali Newspaper.

The set of Feature code able characters, $C_{ORIGINAL}$ as,

{আ, র, ড়, য়, খা, গা, ঘা, ঝা, ণা, থা, ধা, না, পা, বা, মা, যা, রা, লা, শা, সা, য়া, র্ক, র্খ,. . . . .}

The property array as {Number of words, Number of Characters, Number of characters with point, Number of vowels}. We have represented these properties by fixed length coding 00, 01, 10 and 11 respectively.

2) Output

The generated Cover file $T_{GENCOVER}$, shown in Figure 2. There are no such noticeable difference between T_{COVER} and $T_{GENCOVER}$.

3) Discussion of Intermediate steps

$C_{FECODED}$ represents the set of the corresponding changed characters of $C_{ORIGINAL}$.

The secret message M = 'DO'

a) Step 1: Converted message in binary, M_B = 0100010001001111

b) Step 2: Length of M_B, $|M_B|$ = 16 and hence the binary conversion of $|M_B|$ = 00010000. So, the new message M_{BN} = 0001000001000100010001001111

এই মাঠ আসলে ব্যারেটোর। ঘাসের নীচে লুকিয়ে থাকে বৃষ্টির ছিটে। পা টেনে ধরে। গতি মন্থর করে দেয়। ডজ করলেই চলকে ওঠে জল।
এই মাঠ আসলে ব্যারেটোর। কলকাতায় কতবার এমন ভিজে মাঠে ব্যারেটো-মায়ায় সম্মোহিত হয়েছে বিপক্ষ। এই ব্যারেটোর জন্য সমর্থকরা পাগল। দীর্ঘদিন ধরে ব্যারেটো মাঠের রংকে চিরসবুজ করে রেখেছে। নাকে, গালে, কপালে ঘাসের টুকরো লেগে ব্যারেটোর মুখে মৃদু হাসি। এই মাঠ আসলে ব্যারেটোর সমর্থকদের জন্যও। তাই চোখের মধ্যে হাসি লেগে সমর্থকদেরও মধ্যে।

Figure 1. The Original Cover File, TCOVER

এই মাঠ আসলে ব্যারেটোর। ঘাসের নীচে লুকিয়ে থাকে বৃষ্টির ছিটে। পা টেনে ধরে। গতি মন্থর করে দেয়। ডজ করলেই চলকে ওঠে জল।
এই মাঠ আসলে ব্যারেটোর। কলকাতায় কতবার এমন ভিজে মাঠে ব্যারেটো মায়ায় সম্মোহিত হয়েছে বিপক্ষ। এ ব্যারেটোর জন্য সমর্থকরা পাগল। দীর্ঘদিন ধরে ব্যারেটো মাঠের রংকে চিরসবুজ করে রেখেছে। নাকে, গালে, কপালে ঘাসের টুকরো লেগে ব্যারেটোর মুখে মৃদু হাসি। এই মাঠ আসলে ব্যারেটোর সমর্থকদের জন্যও। তাই চোখের মধ্যে হাসি লেগে সমর্থকদেরও।

Figure 2. The Program Generated Cover File, $T_{GENCOVER}$

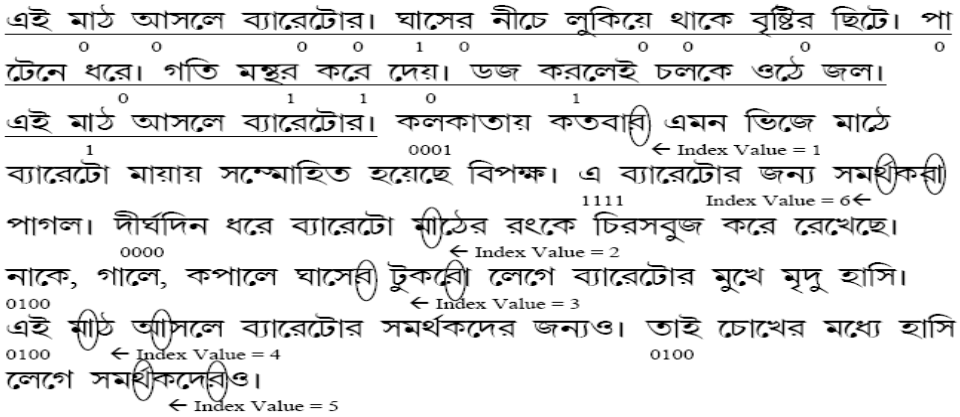

Figure 3. Bits ,corresponding characters, sentences in the Generated Cover File

c) Step 3: P_N = 0000100000011011, as the property array P = 00011011 and the binary conversion of |P| = 00001000.

d) Step 5 – Step 7: Hides the bit stream P_N = 0000100000011011(concatenating P and binary conversion of |P|) by using feature coding method to first 16 feature code able characters available in the selected cover media T_{COVER}. Hidden bits and the feature code able characters are shown in the underlined sentences of the text in Figure 3.

e) Step 8 – Step 10: M_{BN} is divided into 6 blocks 0001, 0000, 0100, 0100, 0100 and 1111, each of size equal to number of considered properties. By scanning T_{COVER}, starting from the next sentence of the underlined sentences, we found that the first sentence is following the desired properties represented by the first block 0001 i.e. Number of words, Number of characters, Number of Characters with point and Number of vowels are even, even, even and odd respectively. Since this is the first block therefore we have indexed the sentence with the index value 1(one) and to maintain this we have replaced the corresponding character

by its respective feature coded character, which is shown within the circle. Similarly we have encoded the second, third and so on to hide M_{BN}. Blocks, replaced characters with index value of the sentence is given at the below of each sentences of $T_{GENCOVER}$, shown in Figure 3.

B. *Algorithm EXTRACT*

1) Input

Cover Media = $T_{GENCOVER}$.

2) Output

Message = 'DO'

3) Discussion of Intermediate steps

a) Step 1: From the first eight feature code able characters we got |P| = 00001000 i.e. 8(eight).

b) Step 2 – Step 4: From the next eight feature code able characters of $T_{GENCOVER}$ we got the property array P = 00011011, i.e. 00, 01, 10 and 11 respectively.

c) Step 5 – Step 8: We found the value of the sentences by finding properties of each sentence. And by finding the Index value of the feature code able characters and placing the bits accordingly we got the message 'DO'.

VI. Conclusion and Future Research Scope

In this paper we have introduced a new approach on text steganography through Indian Languages. Considering the properties of a sentence such as number of words, number of characters, number of vowels etc and using the presence of redundant feature code able characters in Indian Languages, this approach hides the message into an innocent cover file containing Indian texts. To do this first we have divided the message into a number of blocks each of size equal to the number of properties. Then by comparing the property represented by the bits and the properties that followed by the sentence we have hidden the secret message. This approach also presents the extraction of message from the generated cover file by applying the reverse method of hiding. On applying the proposed method to the topic of daily newspaper in Indian Languages like Bengali we got satisfactory results with unnoticeable differences between original text file and the program generated cover file.

Future research can be made to decrease the time complexity of the algorithm.

VII. References

[1] C . Cachin, "An Information-Theoretic Modelfor Steganography", in proceeding 2nd Information Hiding Workshop, vol. 1525, pp. 306-318, 1998

[2] R Chandramouli, N. Memon, "Analysis of LSB Based Image Steganography Techniques", IEEE pp. 1019-1022,2001.

[3] N.F. Johnson, S. Jajodia, "Staganalysis: The Investigation of Hiding Information", IEEE, pp. 113-116, 1998.

[4] D. Artz, "Digital Steganography: Hiding Data within Data", IEEE Internet Compufing, pp. 75-80, May-Jun 2001.

[5] Herodotus. The Histories. Penguin Books, London, 1996. Translated by Aubrey de Sélincourt.

[6] G. Simmons, "The prisoners problem and the subliminal channel," CRYPTO, pp.51–67, 1983.

[7] J. Chen, T. S. Chen, M. W. Cheng, "A New Data Hiding Scheme in Binary Image," in Proc. Fifth Int. Symp. on Multimedia Software Engineering. Proceedings, pp. 88-93 (2003).

[8] G. Doërr and J.L. Dugelay, "A Guide Tour of Video Watermarking", Signal Processing: Image Communication, vol. 18, Issue 4, 2003, pp. 263-282.

[9] G. Doërr and J.L. Dugelay, "Security Pitfalls of Frameby- Frame Approaches to Video Watermarking", IEEE Transactions on Signal Processing, Supplement on Secure Media, vol. 52, Issue 10, 2004, pp. 2955-2964.

[10] K. Gopalan, "Audio steganography using bit modification", Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing,(ICASSP '03), vol. 2, 6-10 April 2003, pp. 421-424.

[11] J.T. Brassil, S. Low, N.F. Maxemchuk, and L.O'Gorman, "Electronic Marking and Identification Techniques to Discourage Document Copying", IEEE Journal on Selected Areas in Communications, vol. 13, Issue. 8, October 1995, pp. 1495-1504.

[12] P. Wayner, "Mimic functions", Cryptologia XVI, pp. 193–214, July 1992.

[13] M. T. Chapman, "Hiding the hidden: A software system for concealing ciphertext as innocuous text", Master's thesis, University of Wisconsin-Milwaukee, May 1997.

[14] Peng Meng, Liusheng Huang, Zhili Chen, Wei Yang, Dong Li," Linguistic Steganography Detection Based on Perplexity", International Conference on MultiMedia and Information Technology, pp 217-220, 2008.

[15] T.Moerland,"Steganography and Steganalysis", May 15, 2003, www.liacs.nl/home/tmoerlan/privtech.pdf.

[16] S.H. Low, N.F. Maxemchuk, J.T. Brassil, and L.O'Gorman, "Document marking and identification using both line and word shifting", Proceedings of the Fourteenth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '95), 2-6 April 1995,vol.2, pp. 853 - 860.

[17] A.M. Alattar and O.M. Alattar, "Watermarking electronic text documents containing justified paragraphs and irregular line spacing ", Proceedings of SPIE – Volume 5306, Security, Steganography, and Watermarking of Multimedia Contents VI, June 2004, pp. 685-695.

[18] K. Bennett, "Linguistic Steganography: Survey, Analysis, and Robustness Concerns for Hiding Information in Text", Purdue University, CERIAS Tech. Report 2004-13.

[19] Y. Kim, K. Moon, and I. Oh, "A Text Watermarking Algorithm based on Word Classification and Inter-word Space Statistics", Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR'03), 2003, pp. 775–779

[20] K. Rabah, "Steganography-The Art of Hiding Data", Information Technology Journal, vol. 3, Issue 3, pp. 245-269, 2004.

[21] Shirali-Shahreza, M.H.; Shirali-Shahreza, M., "A New Approach to Persian/Arabic Text Steganography ",Computer and Information Science, 2006. ICIS-COMSAR 2006. 5th IEEE/ACIS International Conference on 10-12 July 2006 Page(s):310 – 315

[22] D. Huang, and H. Yan, "Interword Distance Changes Represented by Sine Waves for Watermarking Text Images", IEEE Transactions on Circuits and Systems for Video Technology, vol. 11, no. 12, December 2001, pp. 1237-1245.

[23] S. Changder, N.C. Debnath , "An Approach to Bengali Text Steganography", Proceedings of the International Conference on Software Engineering and Data Engineering (SEDE-08), ISBN: 978-1-880843-67-3, pp. 74-78, July, 2008, Los Angeles, California, USA.

[24] S. Changder, N.C. Debnath , "A new approach for steganography in Bengali text", Journal of Computational Methods in Science and Engineering(JCMSE), IOS Press, ISSN1472-7978 , Pages111-122,2009.

[25] S. Changder, Narayan C. Debnath, “New Techniques and Algorithms for Text Steganography through Hindi Text.” Proceedings of the International Conference on Software Engineering and Data Engineering (SEDE-09), ISBN: 978-1-880843-71-0, pp 200-204 June 2009, Las Vegas, USA.

[26] S. Changder, N.C. Debnath, D. Ghosh , “A New Approach to Hindi Text Steganography by shifting Matra”, Proceedings of the IEEE International Conference on Advances in Recent Technologies in Communication and Computing ARTCOM2009, ISBN: 978-0-7695-3845-7, pp. 199-202, October, 2009, Kerala, India.

[27] S. Changder, N.C. Debnath, D. Ghosh, “LCS based Text Steganography through Indian Languages” Proceedings of 2010 3rd IEEE International Conference on Computer Science and Information Technology(ICCSIT 2010), ISBN: 978-1-4244-5539-3,pp. 53-58(vol 8) July, 2010, Chengdu, China