Linguistic Approach for Text Steganography through Indian Text

S.Changder, D. Ghosh
Dept. of Computer Applications, Dept. of CSE
National Institute of Technology, Durgapur
West Bengal, India
suvamoy.nitdgp@gmail.com, profdg@yahoo.com

N. C. Debnath
Dept. of Computer Science
Winona State University
Winona, USA
ndebnath@winona.edu

***Abstract*—Steganography is the art and science of hiding a message inside another message without drawing any suspicion to the others so that the message can only be detected by its intended recipient. With the other Steganography methods such as Image, Audio, Video, a number of text Steganography algorithms have been introduced. This paper presents a new linguistic approach for Steganography through Indian Languages by considering the flexible grammar structure of Indian Languages. To add more security to the system, instead of hiding the original message it is converted to an irrelevant binary stream by comparing the message bits with the pixel values of an Image. Thereafter, the bits of this binary stream are encoded to some part-of-speech and by creating meaningful sentences starting with a suitable word belonging to the mapped part-of-speech, the proposed method hides the message inside a cover file containing some innocuous sentences. Similarly in receiving side, the algorithm finds the corresponding part-of-speech of the starting word of each sentence and place the bit stream of the mapped part-of-speech to recover the converted message. After comparing these bits with the Image pixels, the algorithm extracted the original message from the cover file. The proposed method exhibits satisfactory result on some Indian Languages like Bengali.**

Keywords- Information Security; Text Steganography; Information Hiding; Text watermarking; Linguistic Text Steganography

I. INTRODUCTION

As the number of data being exchanged on the Internet increases, the network security is becoming more and more important. Therefore, the confidentiality and data integrity are required to protect against unauthorized access. This has resulted in an explosive growth of the field of information hiding. Various methods like Cryptography, Steganography, coding and so on have been used for this purpose. However, during recent years, Steganography perhaps has attracted more attention than others.

Steganography is the method of hiding information such that its presence cannot be detected [1]. A secret message is encoded in such a manner that the existence of the information is hidden. Combining with existing communication methods, Steganography can be used to carry out hidden exchanges.

The objective of Steganography is to establish a secured communication in a completely undetectable manner [2] and to avoid drawing suspicion to the transmission of a hidden data [3]. It is not to keep others from knowing the hidden information, but it is to keep others from suspecting that the information even exists. A Steganography method fails if it causes a suspicion that there is secret information in a carrier medium [4].

Trithemius, the author of Polygraphia and Steganographia, invented the word Steganography and the word is derived from two Greek words steganos, meaning "covered", and graphein, meaning "to write". The first written evidence about Steganography being used to send messages is the Heredotous[5] story about slaves and their shaved heads.

The modern representation of Steganography can be given in terms of the prisoner's problem [6]. Specifically, in the general model for Steganography, illustrated in "Fig. 1", let Alice wishing to send a secret message S to Bob. In order to do so, Alice embeds S into a cover-object C to obtain the stego-object Ć. The stego-object Ć is then sent through the public channel. In a pure Steganography framework, the technique for embedding the message is unknown to Wendy and shared as a secret between Alice and Bob. However, it is generally not considered as good practice to rely on the secrecy of the algorithm itself. In private key Steganography Alice and Bob share a secret key which is used to embed the message.

In compared to other Steganography such as on images [2, 4, 7], video files [8, 9], audio files [10] etc., text Steganography seems to be most difficult kind of Steganography due to the lack of large scale redundancy of information in a text file[11].

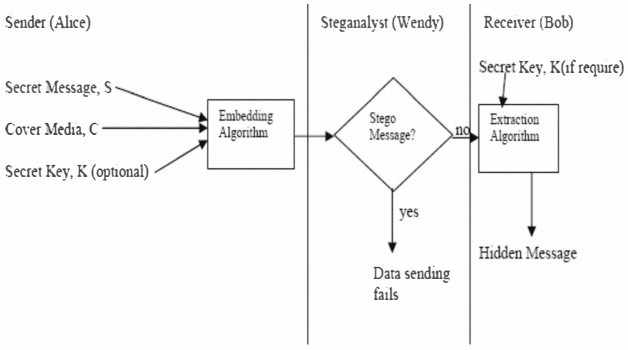

Figure 1. General model for Steganography.

978-1-4244-8845-2/10/ $ 26.00 © 2010 IEEE

There are many steganographic applications for digital image, including copyright protection, feature tagging and secret communication. Copyright notice or watermark can be embedded inside an image to identify it as intellectual property.

This paper presents a new approach for Linguistic text Steganography through Indian Languages. Considering the flexible grammar structure of Indian Languages, to hide the secret message, first the algorithm map the bits of the binary stream of the secret message to the corresponding part-of-speech and then creates the cover file by creating meaningful innocuous sentences starting with a suitable word belonging to that selected part-of-speech. Also to add more security, instead of hiding the original message the proposed method converts the original message to an irrelevant binary stream by comparing the message bits with the pixel values of an Image, which can be considered as a key. Similarly in receiving side, first, the algorithm finds the corresponding part-of-speech of the starting words of each sentence and map to the corresponding code stream of that part-of-speech. After comparing the bits of this binary stream with the Image pixels, the algorithm extract the original message from the cover file.

II. Related Works on Text Steganography

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. NICETEXT uses the cover text as a source of syntactic patterns and by running the cover text through a part-of-speech tagger, this algorithm obtains a set of "sentence frames. Then by using the dictionary, a large list of (type, word) pairs where the type may be based on the part-of-speech tagger or its synonyms, the system randomly generate sequence of words to form a sentence. 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 technique, 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 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 $-\alpha$. For α, the information is 1 and for $-\alpha$, 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, by using feature coding method we have proposed some algorithms for text Steganography through Indian Languages such as Bengali [23, 24] and Hindi [25, 26].

III. Proposed Algorithm

In this proposed work we have divided the encoding system into two subsystems to ensure more security. As the first level of security, in subsystem 1 we have converted the secret message to some meaningless (with respect to the original message) binary stream by applying the method as discussed in subsection A. Then in subsystem 2, discussed in subsection B, we have considered the converted binary stream as the input and encoded this stream by forming some meaningful, innocuous sentences using Indian Languages. Here we have proposed our system on Indian Languages because most of the Indian Languages shows the following properties which are not usually followed by English Language:

- Flexible use of grammar for the formation of sentences or phrase structures. So it may draw less attention to the unintended readers.
- In Indian Languages, personal pronouns shows three and two points of formalities in the second person and third person respectively, rather than one as used in English Language. By using this property we can create large number of different sentences without drawing any suspicion to the others.

A. Functions and operations of Subsystem 1

Let us consider an image as a collection of pixels $P_{i, i=1 \text{ to } n}$ and a secret message (compressed, if necessary) $S_{j, j=1 \text{ to } m (<= n)}$ in binary form.

Create the converted secret message CS as follows:

$CS_{k,(k=1 \text{ to } 8)}$ = bits of binary conversion of the length(S).

$CS_{k,(k=9)} = 0$,

$CS_k = 1$, if $(P_k \bmod 2) = S_k$,

$= 0$, otherwise for k = 10 to (m+9)

$CS_k = 0$, for k = (m+10) to until ((k mod 3) = 0).

So here in this process, we have converted the actual message to an irrelevant binary stream so that it will not lead to any meaningful message to any unintended recipient.

B. Functions and operations of Subsystems 2

To do the work in this tier let us define the followings:

1. The input set S where S = {s | s is a binary k-string where, k=3}.
2. A set POS of all Part of speech used in an Indian Language.
3. A subset SPOS of POS defined as: SPOS = {Proper Noun, Common Noun, Proper Noun Genitive, Common Noun Genitive, Proper Noun Accusative, Common Noun Accusative, Pronoun 1st person, Pronoun Accusative 1st Person, Verb, Pronoun Possessive 1st person, Pronoun 2nd person, Pronoun Accusative 2nd person, Pronoun possessive 2nd person, Pronoun 3rd person, Pronoun Accusative 3rd person, Pronoun possessive 3rd person}.
4. Let O be a relation defined over SPOS X SPOS as: O = {{Proper Noun, Verb}, {Common Noun, Pronoun possessive 1st person}, {Proper Noun Genitive, Pronoun 2nd person}, {Common Noun Genitive, Pronoun Accusative 2nd person}, {Proper Noun Accusative, Pronoun possessive 2nd person}, {Common Noun Accusative, Pronoun 3rd person}, {Pronoun 1st person, Pronoun Accusative 3rd person}, {Pronoun Accusative 1st person, Pronoun possessive 3rd person}}.
5. F be an one-to-one function defined as: F(S)→ O.
6. Let E be a system which consider a word as an input and creates a meaningful sentence in Indian Language, starting with the selected word, as the output.

To encode the message we have considered each three consecutive bits of CSk as the input to the function F(S)→ O. For example let us consider the binary stream '010' as the input to the function F. As per the above definitions, for this particular input, F will produce the element {Proper Noun Genitive, Pronoun 2nd person} from O. Without loss of generality, by selecting any part-of-speech from this element randomly, e.g., 'Proper Noun Genitive' for this case, the system considered a word corresponding to the Proper Noun Genitive and created meaningful sentence by using the system E. In this way we have encoded all the bits of the converted binary stream to get an innocent text file for the Steganography system.

C. Algorithms for Steganography.

Algorithm Hide ( )

1. Input an Image I (m x n), the secret message S.
2. for i=1 to m
3. for j= 1 to n
4. IM[(i-1)*n + j] = I[i, j] mod 2.
5. store the binary conversion of S into an array SM.
6. store the binary conversion of the length of SM into an array L[8].
7. for i=1 to 8
8. CS[i] = L[i]
9. CS[9] = 0
10. for i= 10 to (length(CS)+9)
11. if IM[i-9] = SM[i-9]
12. CS[i]=1
13. else CS[i] = 0.
14. while ((i mod 3)<>0)
15. CS[i]=0, i=i+1
16. for i = 0 to ((length(CS)/3) -1)
17. m = 4*CS[3*i+1] + 2*CS[3*i+2] + CS[3*i+3] + 1
18. n = Random[1,2]
19. select a word 'W' from the list under the part-of-speech found in POS[m, n]
20. create/select a meaningful sentence starting with the word 'W' to create the cover file C.
21. end.

For extracting the message we have to consider the same Image used for encoding and the cover file C generated by the Algorithm Hide ( ) as the input to the decoding system. From the starting words of first three sentences we found the length of the message. And from the rest sentences, mapping the starting words to its corresponding part-of speech and their respective code, we found the converted message. By comparing the converted message with the Image we found the binary conversion of the secret message and so its to character equivalent to get the actual secret message. The algorithm for decoding is as follows:

Algorithm Extract ( )

1. Input the Image I(m x n) and the cover file C.
2. for i=1 to m
3. for j=1 to n
4. EIM[(i-1)*n + j] = I[i, j] mod 2.
5. for each sentences of C
6. find the starting word W.
7. find the corresponding part-of-speech of W and place the equivalent code bits respectively into an array TEMP[ ].
8. from the first eight bits of TEMP[ ], find the length L of the converted message.
9. for i = 10 to (L+9)
10. ECS[i - 9] = TEMP[i].
11. for i = 1 to L
12. if ECS[i] =1
13. ESM[i] = EIM[i]
14. else
15. ESM[i] = (1 - EIM[i]).

16. Convert ESM[ ] to its equivalent string of characters to find the original message S.
17. end.

IV. ADVANTAGES AND DISADVANTAGES

The advantage of the method is that a large volume of data can be hidden. Since we are not changing the structure of the cover file therefore it will draw less attention to the unintended recipients. Furthermore due to use of the two-tier security system the data will be more secure and protected.

Since it is not so easy to create the meaningful sentences it may be considered as a disadvantage to the system.

V. EXPERIMENTAL RESULTS

For our experiment we have considered the input I as the Image shown in "Fig. 2" and the message to be hidden S as the string "meet me at 11am". The array IM is created by applying 'modulo 2' arithmetic to each pixel of the Image in row major form. To discuss the experiment, here we have shown only 18 bits of the array. For our image we got the first 18 bits as IM[ ] = 000000000000000000. The program converted the secret message to its equivalent binary stream and for our experiment we got (first 18 bits) SM[ ] = '011011010110010101'. The system found the length of SM as 120 for this example, and stored the equivalent binary stream '01111000' to the array L[ ]. As discussed in Subsystem 1, lines 7-14 of the algorithm created the converted secret message, by copying L[ ] to the array CS[ ], padding the ninth position by 0 (zero) and the rests by comparing the entries of IM[ ] and SM[ ]. After execution of the loop we got the converted secret message CS[ ] as '011110000100100101001101010'. For the first iteration of line 17, i.e. for i=0, line 18 produced m = 4 (m=4*CS[1]+2*CS[2]+CS[3]+1). With the randomly generated value of n = 2 in line 19, from the array POS[ ] as given in Table 1, we got the corresponding part-of-speech as POS[4,2] = 'Pronoun Accusative 2nd person'. Then the program selected a corresponding Bengali (an Indian Language) word belongs to POS[4,2], and created a meaningful sentence starting with that word. By iterating the loops in line 17 we got the corresponding sentences for all the bits of CS. "Fig. 3" shows intermediate steps and explanation by considering each three consecutive data bits from CS[ ], corresponding Part-of-Speech, the Bengali sentence(starting word in bold) and the corresponding English sentence(English equivalent of the corresponding starting Bengali words are given in bold letter). The cover media C, shown in "Fig. 4", is generated by accumulating all these sentences,.

Similarly for extracting the message we have followed the reverse method, that is, first we have considered the Image I, shown in "Fig.2" and the cover file C, shown in "Fig. 4", as the input to the system. After storing the 'modulo 2' values of each pixel of I into EIM[ ] (in row major form), the algorithm selects the starting words of all the sentences of the cover file and place the code bits of part-of-speech of the corresponding words into an temporary array TEMP[ ]. From the first eight bits of TEMP[ ] we found the length of the message L and creates the array ECS[ ] by copying all the entries from TEMP[10] to TEMP[L+9]. Then by comparing the entries of ECS[ ] with the entries of EIM[ ] we got the binary equivalent ESM[ ] of the secret message which is converted to its character equivalent to get the original message S.

VI. CONCLUSION AND FUTURE RESEARCH SCOPE

In this paper we have introduced a new approach for Text Steganography through Indian Language. This system is taking the advantage of the properties of Indian Languages such as sentence formation property and different level of formalities of personal pronouns. Also for implementing the higher level of security we have used an encoding of the original message to some irrelevant binary string. Then we have mapped the bits to the part-of-speech commonly used in Indian Languages by using the defined mapping function. After that, by considering the selected part-of-speech we have chosen a word corresponds to that part-of-speech and encoded the bits by a sentence starting with the selected word. Similarly for extracting the message first we have applied the inverse function to the starting word of the sentences and then by applying the method reverse to the encoding method we got the original message. This method can be used to exchange secret information and text watermarking. Beside electronics Steganography these methods can be applied to hard copy also.

Future research can be made on by considering more part-of-speech and changing the mapping order to increase the message size and more security to system respectively.

TABLE I. BITS AND CORRESPONDING PART-OF-SPEECH TABLE.

POS[ ] [ ]	1	2
1	Proper Noun	Verb
2	Common Noun	Pronoun Possessive 1st person
3	Proper Noun Genitive	Pronoun 2nd person
4	Common Noun Genitive	Pronoun Accusative 2nd person
5	Proper Noun Accusative	Pronoun Possessive 2nd person
6	Common Noun Accusative	Pronoun 3rd person
7	Pronoun 1st person	Pronoun Accusative 3rd person
8	Pronoun Accusative 1st person	Pronoun Possessive 3rd person

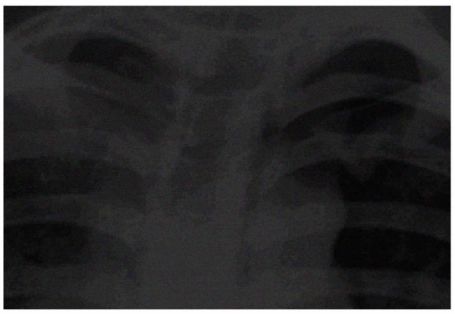

Figure 2. Image File

Data	m	n	POS[m,n]	Bengali Sentence	English equivalent
011	4	2	Pronoun Accusative 2nd Person	তোমাকে টাকা পাঠাচ্ছি।	Sending money to you
110	7	1	Pronoun 1st person	আমি যেতে পারছি না।	I will not be able to go
000	1	2	Verb	কিনে নিও কিছু।	Please buy something
100	5	1	Proper Noun Accusative	রামকে বই কিনে দিও।	Buy and give the book to Ram
100	5	1	Proper Noun Accusative	রাজাকে জামা কিনে দিও।	Buy and give the shirt to Raja
101	6	2	Pronoun 3rd person	সে বলেছিল জামা লাগবে।	He asked for the shirt
001	2	1	Common Noun	টাকা আরো লাগলে জানিও।	Please tell me for more money
101	6	1	Common Noun Accusative	ভাইকে বলবে পড়াশোনা করতে।	Tell to brother to study properly
010	3	2	Pronoun 2nd person	তুমি ঠিক করে থাকবে।	You take care about yourself

Figure 3. Intermediate results of each iteration

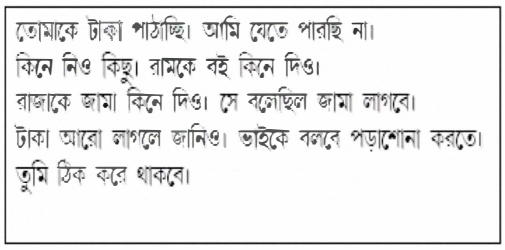

Figure 4. The cover File

REFERENCES

[1] C .Cachin, "An Information-Theoretic Model for 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", ARTCOM2009, ISBN: 978-0-7695-3845-7, pp. 199-202, October, 2009, Kerala, India.