Global Journals Inc.

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 13 Issue 4 Version 1.0 Year 2013
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

An Intelligent Method of Secure Text Data Transmission through Internet and its Comparison using Complexity of Various Indian Languages in Relation to Data Security

By Devasish Pal, Padiga Raghavendra & Dr. A Vinaya Babu

Sri Sai Jyoth Engineering College

Abstract - Security of data transmitted through internet has posed a number of challenges. Data transmitted can be in the form of text, pictures, audio and video clips. In this paper a study has been carried out to find the relationship between the complexities of various Indian languages and its relation to text data security through an intelligent method of converting the text data before transmission. Complexity has been determined from the percentage retrieval by cryptanalyst using reverse frequency mapping without knowing the key. Percentage retrieval with and without converting the text data into an intelligent intermediate form have been compared and repeated in many languages along with English. The percentage retrieval of encrypted data using a language other than English, after converting the same data into a coded file and a dictionary is almost negligible.

Keywords : complexity, dictionary, compression, frequency, retrieval, occurrence, coded file.

GJCST-C Classification : E.3

An Intelligent Method of Secure Text Data Transmission through Internet and its Comparison using Complexity of Various Indian Languages in Relation to Data Security

Devasish Pal [α] , Padiga Raghavendra [σ] & Dr. A Vinaya Babu [ρ]

Abstract - Security of data transmitted through internet has posed a number of challenges. Data transmitted can be in the form of text, pictures, audio and video clips. In this paper a study has been carried out to find the relationship between the complexities of various Indian languages and its relation to text data security through an intelligent method of converting the text data before transmission. Complexity has been determined from the percentage retrieval by cryptanalyst using reverse frequency mapping without knowing the key. Percentage retrieval with and without converting the text data into an intelligent intermediate form have been compared and repeated in many languages along with English. The percentage retrieval of encrypted data using a language other than English, after converting the same data into a coded file and a dictionary is almost negligible.

Keywords : complexity, dictionary, compression, frequency, retrieval, occurrence, coded file.

I. Introduction

Exponential growth of the internet and free accessibility to all users across the globe, security of data across internet has become a prime concern. This security aspect can be further divided into security of data and information in individual systems and in transit between internet users across the network. Data and information can be further divided into text and non text data eg pictures, graphics, audio and video clips. Earlier the text data used to be only in English language. Introduction of unicode and the process of localization [2] encouraged the information exchange of language based context resulting in text data being transmitted in all languages across the internet. This paper deals with security of text data while being transported across the network. To achieve security of data transmission, cryptography is one of the methods in which the security goals can be achieved by means of encryption and decryption. The key used for cryptography can be symmetric or asymmetric key. The encryption can be of blocks of fixed/variable size bit stream transformed to cipher stream. They use either block cipher or stream cipher techniques for transformation. Parameters in these schemes are mainly algorithm and key. Larger the key size, greater is the security of data and slower is the data rate. One more parameter has been considered i.e the complexity of a language [6] with a case study on Telugu. Greater the complexity of the language, greater is the security of text transmitted in that language keeping other parameters like encryption algorithms and the key constant. A simple logical conclusion is that if the text of a script is complex then the same level of security can be achieved with lesser key size. Subsequently a comparative study has been carried out over English and Telugu with Bengali as a case study [7] and it was observed that percentage retrieval of data in Bengali is less than Telugu and English.

In this paper, a comparative study has been carried out on various other Indian languages and adding a fourth security parameter i.e an intelligent method of text data encryption with security [8].

II. Review

A lot of study has gone into making the job of cryptanalysts simpler. Different languages in the world consist of characters displaying different properties and behavior [3, 4] which help in the process of cryptanalysis. One of the methods of determining the language complexity is by the frequency analysis. In this process frequency of each symbol in the encrypted message is determined. This information is used by cryptanalysts, to determine which cipher text symbol maps to the respective plaintext symbol. In transposition systems, the letter frequencies of a cryptogram are identical to that of the plaintext. In the simplest substitution systems, each plaintext letter has one cipher text equivalent. The cipher text letter frequencies are not identical to the plaintext frequencies, but the same numbers will be present in the frequency count as a whole. A method for fast cryptanalysis of substitution ciphers has been proposed by Thomas Jakobsen [1] which uses the knowledge of diagram distribution of the cipher text. The individual letters of any language occur

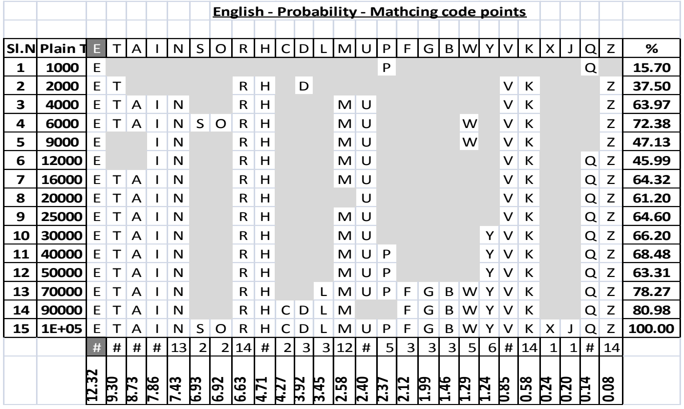
Author α : Associate Professor, IT SSJEC, Hyderabad.
E-mail : dpal55@gmail.com
Author σ : Student, III year IT SSJEC, Hyderabad.
E-mail : raghavendra.padiga@gmail.com
Author ρ : Principal JNTUH. E-mail : dravinaybabu@yahoo.com

with greatly varying frequencies [9]. This factor has been used to solve varying simple ciphers. There are two general approaches to solve simple ciphers. One makes use of the frequency characteristics and the other uses the orderly progression of the alphabet to generate all possible decipherments from which the correct plaintext can be picked up. Statistical analysis of the frequencies of multiple letters when compared to single letters have been found to be more helpful while retrieving part of plain text message. By using the combined techniques of monogram frequencies, keyword rules and dictionary checking the cryptanalytic technique of enhanced frequency analysis has been developed [5].

Plain text is encrypted using the proposed algorithm resulting in cipher text. The frequencies of different characters in the cipher text are extracted. Mapping is carried out between the characters of plain text and cipher text based on these frequencies. Now the characters in cipher text are replaced with the mapped characters of plain text and the percentage of the exact retrieval as compared to plain text is calculated by K.W. Lee et.al [5].

III. Conditional Probability

A vast study has been carried out into the frequency of occurrence of characters of many languages. The characters of different languages have different frequency patterns. This information helps a cryptanalyst to retrieve data from a cipher text by reverse frequency mapping. The percentage of data retrieved increases with the increase of corpus size of a sample text. As an example let us take a case study of English language. First a corpus frequency string is calculated with a very large corpus of English text. Corpus frequency string consists of all the different characters available in the corpus text. Next the percentage of occurrence of each character of corpus frequency string is calculated. To find the percentage retrieval of text after encryption from a new sample text, the new sample text is encrypted. The cryptanalyst using reverse frequency mapping tries to retrieve maximum possible characters. The percentage of occurrence of those characters already calculated earlier using corpus frequency string is added indicating the total percentage retrieval. Eg: retrieved chars: a, r, y and k. If the percentage occurrences of those characters are 8.73, 6.63, 1.24 and 0.58 then the total % retrieval is 8.73 + 6.63 + 1.24 +0.58 = 17.18 as per chart shown in Fig. 1.



English - Probability - Mathcing code points

Sl.N	Plain T	E	T	A	I	N	S	O	R	H	C	D	L	M	U	P	F	G	B	W	Y	V	K	X	J	Q	Z	%
1	1000	E														P										Q		15.70
2	2000	E	T						R	H		D										V	K				Z	37.50
3	4000	E	T	A	I	N			R	H				M	U							V	K				Z	63.97
4	6000	E	T	A	I	N	S	O	R	H				M	U					W		V	K				Z	72.38
5	9000	E			I	N			R	H				M	U					W		V	K				Z	47.13
6	12000	E			I	N			R	H				M	U							V	K			Q	Z	45.99
7	16000	E	T	A	I	N			R	H				M	U							V	K			Q	Z	64.32
8	20000	E	T	A	I	N			R	H					U							V	K			Q	Z	61.20
9	25000	E	T	A	I	N			R	H				M	U							V	K			Q	Z	64.60
10	30000	E	T	A	I	N			R	H				M	U						Y	V	K			Q	Z	66.20
11	40000	E	T	A	I	N			R	H				M	U	P					Y	V	K			Q	Z	68.48
12	50000	E	T	A	I	N			R	H				M	U	P					Y	V	K			Q	Z	63.31
13	70000	E	T	A	I	N			R	H			L	M	U	P	F	G	B	W	Y	V	K			Q	Z	78.27
14	90000	E	T	A	I	N			R	H	C	D	L	M			F	G	B	W	Y	V	K			Q	Z	80.98
15	1E+05	E	T	A	I	N	S	O	R	H	C	D	L	M	U	P	F	G	B	W	Y	V	K	X	J	Q	Z	100.00
		#	#	#	#	13	2	2	14	#	2	3	3	12	#	5	3	3	3	5	6	#	14	1	1	#	14	
		12.32	9.30	8.73	7.86	7.43	6.93	6.92	6.63	4.71	4.27	3.92	3.45	2.58	2.40	2.37	2.12	1.99	1.46	1.29	1.24	0.85	0.58	0.24	0.20	0.14	0.08	

Figure 1 : English probability matching code points

IV. Security Model

a) Normal Method

i. Sample text → encrypted → reverse frequency mapping → decrypted → compared with sample text → calculate % retrieval using corpus frequency string.

b) Proposed Method

ii. Sample text --→ dictionary file + coded file → dictionary file encrypted → reverse frequency mapping → decrypted → compared with sample text → calculate % retrieval using corpus frequency string.

Difference in % retrieval between (a) and (b) is the advantage as per the proposed plan.

Here the text file is converted into a dictionary file which consists of frequency of occurrence of words arranged in descending order and an extended ASCII value is given to each word referred to as the code for that particular word. The extended ASCII values selected are from 33 to 250 i.e. a total of 218 different words are covered. If the number of different words exceed more than 218, then the numbers are repeated appending to its previous values eg 219th word will be 33 33, next will be 33 34 and so on. Based on the coded values of various words in the dictionary, the text file is converted into a coded file which can be decoded only by the dictionary. The dictionary created for different text files is different, hence dynamic in nature. Each text file will have a unique dictionary. The paper [8] speaks about the concept of dictionary and the coded file but left open how the dictionary is to be transmitted in a secure way. It also does not speak of any other language other than English. ie it has not considered language complexity as a factor for security of text data. The coded file created in this paper is different with only the coded values as it serves the purpose. The dictionary is encrypted and transmitted. The coded file is transmitted without encryption. The attacker cannot decode the coded file without getting the information of the dictionary file. The actual percentage of data retrieved from coded file (which represents the plain text data file) will finally be much less than the percentage retrieved from the dictionary file. The percentage data retrieved from dictionary file in various languages for a fixed corpus sizes have been found out using programs in Python 2.7 and displayed in figures 4 and 5 below.

c) Sample Text

Dance, little baby, dance up high,
Never mind baby, mother is by;
Crow and caper, caper and crow,
There little baby, there you go:
Up to the ceiling, down to the ground,
Backwards and forwards, round and round.
Then dance, little baby, and mother shall sing,
With the merry gay coral, ding, ding, a-ding, ding.

d) Dictionary File

```
and -> !
baby, -> "
little -> #
the -> $
ding, -> %
to ->&
mother -> '
Never -> (
caper, -> )
dance -> *
is -> +
There -> ,
mind -> -
forwards, -> .
down -> /
high, -> 0
a-ding, -> 1
caper -> 2
Then -> 3
coral, -> 4
there -> 5
Dance, -> 6
ceiling, -> 7
you -> 8
With -> 9
round. -> :
ground, -> ;
Backwards -><
gay -> =
shall ->>
Baby's -> ?
go: -> @
Up -> A
crow, -> B
dance, -> C
merry -> D
The -> E
by; -> F
Crow -> H
ding. -> I
up -> J
round -> K
sing, -> L
```

e) Coded File

```
6 # " * J 0( - " ' + F
H ! ) 2 ! B, # " 5 8 @
A & $ 7 / & $ ;< ! . K ! :
3 C # " ! ' > L 9 $ D = 4 % % 1 I
```

% retrieval normal method for sample text is 84.75% and proposed method is 0.0%.

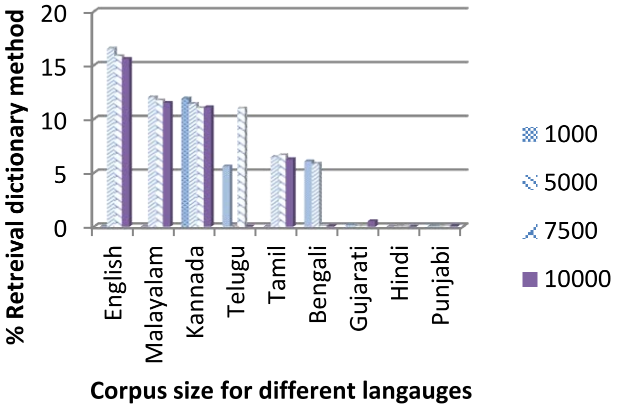


Figure 2 : % Retrieval by dictionary method graph for different languages

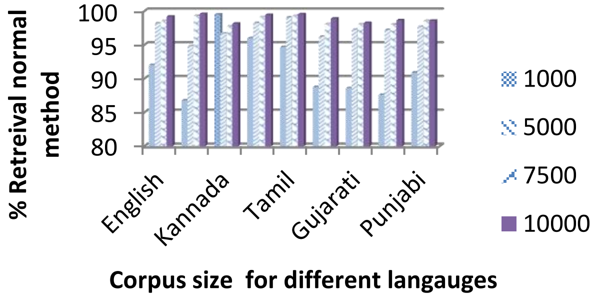


Figure 3 : % Retrieval by normal method graph for different languages

Language	% retrieval dictionary method			
	Corpus size			
	1000	5000	7500	10000
English	0	16.55	15.87	15.6
Malayalam	0	12.03	11.76	11.51
Kannada	11.91	11.43	11.05	11.1
Telugu	5.6	10.78	11.0	11.45
Tamil	0	6.48	6.67	6.27
Bengali	6.07	5.86	0.07	0.05
Gujarati	0.15	0.1	0.05	0.52
Hindi	0	0.02	0.01	0
Punjabi	0.04	0.024	0.045	0.078

Figure 4 : Table displaying the retrieval percentage by dictionary method

Language	% retrieval normal method			
	Corpus size			
	1000	5000	7500	10000
English	92	98.18	98.53	99.18
Malayalam	86.78	94.83	99.31	99.57
Kannada	99.49	96.68	97.74	98.14
Telugu	95.99	98.23	99.11	99.43
Tamil	94.69	99.05	99.27	99.53
Bengali	88.75	96.19	98.09	98.89
Gujarati	88.58	97.25	98.07	98.24
Hindi	87.58	97.19	98.02	98.64
Punjabi	90.89	97.68	98.55	98.57

Figure 5 : Table displaying the retrieval percentage by normal method

V. Findings

a) The percentage data retrieved using conditional probability following a normal method is as explained in IV(a) and after intelligently converting the same sample text into a dictionary form and carrying out the same process as in IV(a) displays a vast difference in the percentage retrieval of data,(Figs 4 & 5) thereby making the proposed system as explained in IV(b) strongly secured.

b) Carrying out the same procedure for similar corpus sizes, the percentage retrieval of data in various Indian languages is far less than English proving that text data transmitted in regional languages is more secured than English language.

c) Amongst the various Indian languages, Gujarati, Hindi and Punjabi display a very low percentage of retrieval of data, making it more secure as far as transmission of data is concerned compared to other languages considered.

d) The three Indian languages Gujarati, Hindi and Punjabi prove to be the most secure amongst the languages considered as case study, are stroke based unlike the languages of the southern part of India which are curvature based.

VI. Conclusions

Security of transmitted data over the internet is most secure when transmitted in any of the Indian languages compared to English language after converting the data into an intermediate form (dictionary and the coded file).

By creating the dictionary, the percentage retrieval compared with plain text file is far less than without creating the dictionary file.

By mapping the retrieved data from dictionary file to coded file the actual data to be retrieved is likely to be far lesser compared to what has been projected in Figs. 4 and 5.

Of the languages considered for case study, Guajarati, Punjabi and Hindi provide better security and they happen to be stroke based than curvature based (south Indian languages).

References Références Referencias

1. Jakobsen, T: A fast Method for Cryptanalysis of Subsittution Ciphers. J. Cryptologia, Volume 19, Issue 3 1995, pp. 265-274.
2. Adam Stone: Internationalizing the Internet. J. Internet Computing. 3, 2003, pp. 11-12.
3. Bauer F L: Decrypted secrets-Methods and Maxims of Cryptology, Springer, 2007.
4. Menezes A. J. P: Handbook of Applied Cryptography. CRC Press, 2001.
5. Lee K.W., C.E. Teh, Y.L. Ta: Decrypting English Text Using Enhanced Frequency Analysis: National Seminar on Science, Technology and Social Sciences 2006 (Ui TM-STSS 2006). pp. 1-7.
6. Bhadri Raju MSVS, Vishnu Vardhan B, Naidu G A, Pratap Reddy L, Vinaya Babu A : Effect of Language Complexity on Deciphering Substitution Ciphers - A Case Study on Telugu.
7. Devasish Pal, Raju Ejjagiri, Dr. A Vinaya Babu: Complexity of Bengali Language and its relation to data security volume1 Issue 4 - 2012 (IJACIT) ISSN 2277-9140.
8. Dr. V.K. Govindan, B.S. Shajee mohan:An Intelligent text data encryption and compression for high speed and secure data transmission over internet.
9. Bao-Chyuan Guan, Ray-I Chang, Yung Chung Wei, ChiaLing Hu, Yu-Lin Chiu: An encryption scheme for largeChinese texts: IEEE 37th Annual.