ISSN:0975-9646

Simultaneous Secure Transmission for Textual Data Using Radomized Bits

Remya R Nair
M.Tech Student, Department of Computer Science and Engineering
Sarabhai Institute of Science and Technology, Vellanad , Trivandrum, India

***Abstract*-Security of data transmitted through internet has put forward a number of challenges. In this thesis work, system is developed in which two techniques are combined, encryption and compression, which provides a strong backbone for its security and reduces extra overhead. The proposed scheme extends security by incorporating pattern recognition, data encryption using SDES and XOR encryption technique, and decreases extra expenses by data compression technique using arithmetic coding. The scenario of present day's information security system includes confidentiality, originality, reliability, non-repudiation. This present work focuses on using the technique for secure and lightweight data transmission approach. Data compression, proposes an attractive technique, to decrease communication costs, by using unused bandwidth effectively. To justify the efficiency of the proposed technique, this proposed technique with different input text sizes has been verified. The security part, associated with the proposed technique is cross checked by methods of preventing capacity against the various security challenges. The compression proportion of the proposed technique has been computed. The result shows that the proposed scheme is proficient and competent when compared to popular existing techniques**.

Keywords— Arithmetic coding; band width; pattern recognition; SDES and XOR encryption mechanism; data compression

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 and the increase in bandwidth transmission speed. Two fields of research have been proposed to enhance the communication security: cryptography and pattern matching. Although they are both applied to the protection of secret message, the major difference is the appearance of the transmitted data. However, reliability issues regarding to data transmission such as confidentiality, data security and data loss are becoming serious concerns [5]. The Client requires that; the transmitted data should not be lost, damaged or manipulated by any unauthorized third party. Data lost can also result from network congestion due to extra overhead [1] [2].

Information Security is the practice of defending information from unauthorized access, use, disclosure, disruption, modification, perusal, inspection, recording or destruction. It is a general term that can be used regardless of the form the data may take. The act of ensuring that data is not lost when critical issues arise

The rapid growth and widespread use of electronic data processing and electronic business conducted through the Internet, along with numerous occurrences of international terrorism, fueled the need for better methods of protecting the computers and the information they store, process and transmit. The academic disciplines of computer security and information assurance emerged along with numerous professional organizations – all sharing the common goals of ensuring the security and reliability of information systems.

Our objective of this paper is to provide an integrated mechanism which can resolve security issues, provide confidentiality, reduce information loss and impose less overhead at the same time.

II. RELATED WORK

The work can be categorized into two parts. The first part is associated with the different data security issues, while the second part depicts the various compression techniques and approaches.

A. Different data security issues

To enhance the safety issues of data transmission, our main focus are ensuring the data security, integrity, and confidentially. A detail discussion about the existing techniques of such issues is drawn in this subsection.

1) Data hiding and Steganography: Steganography is the art of hiding information and an effort to conceal the existence of the embedded information. It serves as a better way of securing message than cryptography which only conceals the content of the message not the existence of the message. Original message is being hidden within a carrier such that the changes so occurred in the carrier are not observable. [2][14]. It can be further categorized into three types such as format based, random and statistical generations, and linguistic method [2]. Water marking [3][4] is one of the example of this approach

2) Cryptographic Mechanism: Cryptography is the process of converting the important data and information into a cipher text form and then convert it again into the decipherable form when it reaches its authorize user [8][9]. The secret key is used to decrypt the message into plain text. It can be classified into two types, including public-key, and private key cryptography [9]. In public key cryptography two keys are used, one for encryption and another for decryption while in the private key cryptography, the single key is used for both encryption and decryption [8].

3) Pattern matching techniques: Pattern matching is a procedure to check a perceived sequence of tokens for the presence of constituents of some predefined pattern.

In our proposed technique, we use cryptographic mechanism, pattern generation and matching technique to provide security confidentiality, and integrity. We have studied the above security scheme to provide better security scheme by finding the weakness of existing techniques.

B. Data Compression:
Compression is a well-defined approach for reducing the number of bits needed to store or transmit data over the network. It can be categorized into two categories such as lossless or lossy. In the lossless compression technique compressed data can be decompressed to its original value .While in the lossy compression technique, actual data cannot be retrieved completely as it discards "unimportant" data, during compression. Here, the focus is lossless compression technique as we have taken this approach as a part of our proposed algorithm and lack of space [13].
1) Huffman Coding: Given a set of data symbols or alphabet and their frequencies of occurrences, build a set of variable-length code words with the smallest average length and assign them in the place of these symbols. Here each time two symbols with the smallest probabilities are selected, and added to the top of the partial tree, deleted from the list, and replaced with an auxiliary symbol representing the two original symbols. When the list is reduced to just one auxiliary symbol, the tree is complete. The tree is then traversed to determine the code words of symbols and replaced with their corresponding code words[6][7].
2) Run Length Encoding: In this approach, any sequence of identical symbols will be replaced by the number of repetitions of this particular symbol followed by this particular symbol. For instance, the text 'aaaa' is coded as '4a'. RLE is widely used in early graphics file format for Compressing black and white images [13].
3) Lempel Ziv: It is the dictionary-based encoding technique. Some predefined codes represent the sequence of characters from matching previously stored database. In this mechanism the search is done within the search buffer and the longest matching string is taken to replace the character or symbols. [13].
4) DEFLATE algorithm: It combines both LZ77 and Huffman compression technique to compress data. LZ77 is a dictionary-based compression technique, so it uses a 32K sliding window to record the repetitive characters. In present scenario, many software implementations such as PKZIP, zlib/gzip, 7-Zip/Advance COMP use Deflate algorithm. It searches duplicated strings in input data. Second occurrence of a string is replaced by a pointer to previous string, in form of a pair [10].

A data loss [5] was not covered adequately in the previous systems which is an accidental or deliberate exposure of classified, sensitive or official information into an uncontrolled or unauthorized environment or to persons without a need-to-know can be called Data spill. A data spill is sometimes referred to as unintentional information disclosure or a data leak. This risk has been mitigated to a large extent in the proposed system since we have minimal data loss in this system. Many of the existing systems are not capable of preventing the deadly sniffer and chosen sniffer text attack. Moreover existing systems, is dependent of the bit stream. Majority of the time there is an Extra overload [1][2] due to network congestion which can result in Data Loss.
In the case of Huffman coding, 1. Relatively slow. 2. Depends upon statistical model of data. 3. Decoding is difficult due to different code lengths. 4. Overhead due to Huffman tree. In the case of LZW 1.Management of string table is difficult. 2. Amount of storage needed is indeterminate. 3. Royalties have to be paid to use LZW compression. In Run length encoding ,Compression ratio is low as compared to other algorithm

III. PROPOSED SCHEME

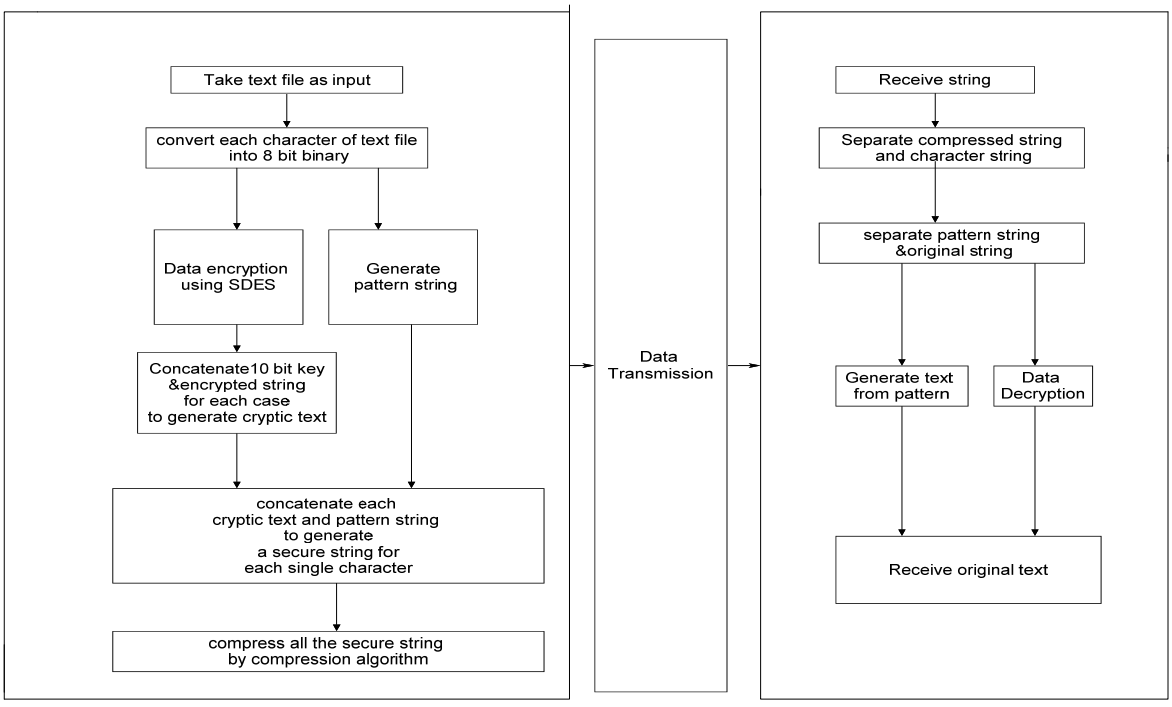

Figure 3.1: Architectural Diagram

To demonstrate our proposed model, we have described all the operations, related to our proposed model, in the following sub sections

1. Data Pre-processing First, data over large data sets are often bottlenecked by the I/O system, such as reading data from disk or streaming through memory. As a result, encryption[8] schemes that significantly increase the size of the data can slow down data processing. Thus, one challenge lies in partitioning the data into parts that can be executed using the available encryption schemes on an untrusted area, and parts that must be executed on the trusted client. Some of these schemes achieve efficiency by revealing additional information to the file, such as the order of items for sorting. Any sequence of identical symbols will be replaced by the number of repetitions of this particular symbol followed by this particular symbol. It is the dictionary-based encoding technique. Some predefined codes represent the sequence of characters from matching previously stored database. In this mechanism the search is done within the search buffer and the longest matching string is taken to replace the character or symbols

2. Pattern String Generation Each character of input plan text (pt) is converted represented into 15 bits binary string by the following way:

1. If size of pt is Text Size , total 8 bits input characters

$$\overline{InputLen} = \left(\frac{TextSize}{8}\right) \quad (1)$$

2.If the character sequence of pt is represented by (pt_m) and if bit sequence of each pt_m is represented by (pt_{mi}), then convert each of the pt_m to the decimal value DEC_m, where DEC_m, $pt_m \in \sum$ and from equation (1), we have,

$$DEC_m = \sum_{i=0}^{7}(pt_{mi} \times 2^i), for\ 0 \le m < \overline{InputLen}$$

3.Separate all the digits of DEC_m , in the form of D_{mk} where $0 \le DEC_m \le 127\ and\ initialy\ D = \langle\emptyset\rangle$.

$$D_{mk} = \left(\left\lfloor\frac{DEC_m}{10^j}\right\rfloor\right) modulo\ 10, for\ 2 \le j \le 0\ ; 0 \le k \le 2$$

4. Construct pattern table by plotting actual value of each D_{mk} according to X axis, and position value according to Y axis. For example, the decimal value of A is 65 and actual value of D_{mk} are 0, 6, 5. The position values of D_{mk}, i.e. the value of k are 0, 1, 2. Place them in pattern table.

Position	Actual Value									
Value	(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(0)	D_{m0}	0	0	0	0	0	0	0	0	0
(1)	0	0	0	0	0	0	D_{m1}	0	0	0
(2)	0	0	0	0	0	D_{m2}	0	0	0	0

Table 3.2.1 Pattern Table

5. Calculate each position value, p_k, of each D_k from the pattern table by

$$p_{mk} = (k \times 10) + D_{mk}\ , for\ 0 \le k \le 2$$

6. If x , y are integers and Z_m is 5 bits length string, convert each p_{mk} into 5 bits strings, where $0 \le k \le 2$

$$\left.\begin{array}{l} y = P_{mk \div 2^i} \\ x = y\ modulo\ 2^i \\ Z_{mki} = x \end{array}\right\} for\ 0 \le i \le 4\ and$$

7. Concatenate three, 5 bits z_{mk} to generate each 15 bits pattern string, Pat_m by,

$$\left.\begin{array}{l} Pat_{mi} = Z_{m0i} \\ Pat_{m(i+5)} = Z_{m1i} \\ Pat_{m(i+10)} = Z_{m2i} \end{array}\right\} for\ 0 \le i \le 4$$

3. Encrypted string generation

SDES and XOR algorithm is using here. After generation of each 8bits cipher text, each 10 bits .Secret key is concatenated to each other to generate an 18 bits encrypted string

4. Data compression

Compression is a well-defined approach for reducing the number of bits needed to store or transmit data over the network. To accomplish the compression, we will take each pattern string and encrypted string to generate a secure string for each single character of the input text. Compressed all secure string by arithmetic coding

- Arithmetic coding is a form of encoding used in lossless data compression
- Arithmetic coding encodes the entire message into a single number, a fraction *n* where ($0.0 \le n < 1.0$).
- It has high compression ratio than any other compression technique

One advantage of arithmetic coding over other similar methods of data compression is the convenience of adaptation. Adaptation is the changing of the frequency (or probability) tables while processing the

Encoding is the process of putting a sequence of characters (letters, numbers, punctuation, and certain symbols) into a specialized digital format for efficient transmission or transfer. Decoding is the opposite process -- the conversion of a digital signal into a sequence of characters. The symbols are encoded into variable length output codes. The length of the output codes varies based on the probability of frequency of symbol. Low probability symbols are encoded using many bits, and high probability symbols are encoded using fewer bits. In encoding codeword is a floating point number between 0 and 1. Bigger the input size, the number of digits in the output becomes more. An input symbol after compression is represented by an interval of real numbers between 0 and 1. The range of interval is initially defined by two values, high and low, which are equal to 1 and 0 respectively. The interval is successfully subdivided as when is each new source symbol is encoded. A randomly generated probability is multiplied with the probability of occurrence to generate the range. The secret key is the seed value of the random number In Encoding a source ensemble is represented by an interval between 0 and 1 on the real number line. The coding assumes an explicit probabilistic model of the source. It is a defined-word scheme which uses the probabilities of the source messages to successively narrow the interval used to represent the ensemble. A high probability message narrows the interval less than a low probability message, so that high probability messages contribute fewer bits to the coded ensemble. The method begins with an unordered list of source messages

and their probabilities. The number line is partitioned into subintervals based on cumulative probabilities.

5.Retrieval of text at the receiver end

To retrieve the text at receiver end ,

a) First separate compressed string and character string using separator.

b) Then we separate pattern string and encrypted string and retrieve characters from one of them by the reverse way.

Please note, If one of them is corrupted or modified during transmission then, we can also retrieve characters from other one. Thus we can reduce the error and provide robustness.

IV. CONCLUSION

The biggest challenge though is likely to be systems integration and assurance. The proposed scheme is two folds, initially the encrypted string and pattern string are generated to provide the security to the input text, and then compression is done to reduce to the extra overhead. It is also independent of the bit stream. Hence, if some of the bits are modified or lost during the travelling time, it does not significantly impact the original data. Also, For the smooth functioning of client and server it is designed as separate threads in the application. The proposed algorithm is also time efficient for both data incorporation and retrieval. The Proposed system is more robust and prepared to prevent security attack. At the same time data Information loss will be certainly to the minimal Its more capable of compressing different size inputs text. To enhance the safety issues of data transmission, our main focus are ensuring Data security, integrity and confidentiality by Cryptographic Mechanism[8] and Pattern Matching techniques[11][12] & Data Compression [13].

ACKNOWLEDGEMENT

This work was supported in part by the Department of Computer Science & Engineering, SIST, Trivandrum. I would like to show my gratitude to Dr. C G Sukumaran Nair (HOD), Associate Professor, Ms. Sudha SK and Assistant Professor, Mrs. SoyaChandra C S, for their valuable guidance

REFERENCES

[1] Z. Jalil; A. M. Mirza; H. Jabeen, "Word length based zerowatermarking algorithm for tamper detection in text documents,"Computer Engineering and Technology (ICCET), 2010 2nd International Conference on , vol.6, no., pp.V6-378-V6-382, 16-18 April 2010

[2] Dauneria; K. Indu, "Encryption Based Data Hiding Architecture with Text Pattern Authentication and Verification," Computer and Information Technology Workshops, 2008. CIT Workshops 2008. IEEE 8th International Conference on, vol., no., pp.236-241, 8-11 July 2008

[3] W. Jinwei; L. Guangjie; L. Shiguo, "Security Analysis of Content-Based Watermarking Authentication Framework," Multimedia Information Networking and Security, 2009. MINES '09. International Conference on, vol.1, no., pp.483-487, 18-20 Nov. 2009

[4] C. Ning; Z. Jie, "A multipurpose audio watermarking scheme for copyright protection and content authentication," Multimedia and Expo, 2008 IEEE International Conference on, vol., no., pp.221-224, June 23 2008-April 26 2008

[5] Z. Zhang; S. Qibin; W. Wai-Choong; J. Apostolopoulos; S. Wee, "Rate-Distortion-Authentication Optimized Streaming of Authenticated Video," Circuits and Systems for Video Technology, IEEE Transactions on , vol.17, no.5, pp.544-557, May 2007

[6] L. Hemme; L. Hoffmann, "Differential Fault Analysis on the SHA1 Compression Function," Fault Diagnosis and Tolerance in Cryptography (FDTC), 2011 Workshop on , vol., no., pp.54-62, 28-28 Sept. 2011

[7] L. Zhiqiang; M. W. Hoffman; W.D. Leon; N. Schemm; S. Balkir, "A CMOS Front-End for a Lossy Image Compression Sensor," Circuits and Systems, 2007. ISCAS 2007. IEEE International Symposium on ,vol., no., pp.2838-2841, 27-30 May 2007

[8] J. K. Pal; J. K. Mandal, "A Random Block Length Based Cryptosystem through Multiple Cascaded Permutation Combinations and Chaining of Blocks," Fourth International Conference on Industrial and Information Systems, ICIIS 2009, 28–31December2009, SriLanka.

[9] Y. Ji; K. Hongbo, "FPGA implementation of dynamic key management for DES encryption algorithm," Electronic and Mechanical Engineering and Information Technology (EMEIT), 2011 International Conference on , vol.9, no., pp.4795-4798, 12-14 Aug. 2011

[10] A. Yazdanpanah; M. R. Hashemi, "A simple lossless preprocessing algorithm for hardware implementation of DEFLATE data compression," Electrical Engineering (ICEE), 2011 19th Iranian Conference on , vol., no., pp.1, 17-19 May 2011M. M. Alani; , "DES96 - improved DES security," Systems Signals and Devices (SSD), 2010 7th International Multi-Conference on , vol., no., pp.1-4, 27-30 June 2010

[11] J. Wang; F. Kang; X. Xu; J. Chen, "A Fast Single Pattern Matching Algorithm Based on the Bit-Parallel," Frontier of Computer Science and Technology (FCST), 2010 Fifth International Conference on , vol., no., pp.17-21, 18-22 Aug. 2010

[12] M. Paul; M. Murshed, "An Optimal Content-Based Pattern Generation Algorithm," Signal Processing Letters, IEEE , vol.14, no.12, pp.904-907, Dec. 2007

[13] U. Mehta; K.S. Dasgupta; N.M. Devashraye, "Survey of Test Data Compression Technique Emphasizing Code Based Schemes," Digital System Design, Architectures, Methods and Tools, 2009. DSD '09. 12th Euromicro Conference on, vol., no., pp.617-620, 27-29 Aug. 2009.

[14] M. Zamani; A. Manaf; R. B. Ahmad; F. Jaryani; H. Taherdoost; A.M. Zeki, "A secure audio steganography approach," Internet Technology and Secured Transactions, 2009. ICITST 2009. InternationalConference for, vol., no., pp.1-6, 9-12 Nov. 2009