International Journal of Computer Science and Engineering (IJCSE)
ISSN 2278-9960
Vol. 2, Issue 3, July 2013, 47-52

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

A SECRET SHARING SCHEME AND NATIONAL SECURITY ENHANCEMENT THROUGH NATURAL LANGUAGE PROCESSING

RAFIDHA REHIMAN K A[1] & LAKSHMI K S[2]

[1]M.Tech Information System Security, RECT IGNOU, Rajagiri School of Engineering and Technology, Rajagiri Valley, Cochi, Kerala, India

[2]Department of Information Technology, Rajagiri School of Engineering and Technology, Rajagiri Valley, Cochi, Kerala, India

ABSTRACT

In Cryptography, it is assumed that language of communication is in English and the major problem of English in communication is frequency analysis. In this paper we propose a new secret sharing scheme through natural language romanization and symmetric key DES cryptography. Many anti social groups and terrorists use cryptography on their natural language to exchange secret SMS messages. In this context we study the importance of language identification for security and develop an identification system for romanized Malayalam and Hindi and plain English with SVM regression. After identification system transliterates the message to corresponding font. Also for testing purpose we create a language corpus including romanized Malayalam, romanized Hindi and English messages. With this system we obtain an accuracy of 94.2381% for identification and 91 % for transliteration.

KEYWORDS: Character n – Gram Approach, Cryptography and Network Security, Frequency Analysis, Language Identification, Natural Language Processing, Romanization, Transliteration

INTRODUCTION

Providing the complete security to the communication system targets a nation's strength and is considered as prime requirement of the government. We can use Cryptography for secret message sharing [1]. The normal message is called the plain text and it is converted into a cipher text by means of cryptographic algorithms and cryptographic keys. Cryptographic algorithms and keys are vulnerable to attacks; an interceptor in the communication channel can perform cryptanalysis and recover the secrets. If key is compromised by brute force attacks then cryptographic algorithms will fail.

Language is primary means for communication and most of the communication between the people can taken place in their native languages. For exchanging it through technologies either one requires a language editor or they can apply romanization on natural language [2]. With the rapid growth of Internet and Information technology many people use English alphabets to express their native language. Romanization is a scheme by which the users represent their mother tongue using English alphabet. In other words it is a transliteration scheme in which natural language expressed using roman characters.

Encrypt the message after romanization provides a new dimension to Cryptography especially it is a big challenge to the cryptanalyst. So to ensure the national level security of a nation we require language identification from Romanization. Language Identification from scripts and fonts is almost settled in both foreign and Indian languages [3] [4] [5] [6]. Written language identification from Romanized text is a strong research thread and has been the topic of significant and varied research. Many researchers are working in this field for foreign languages and not much explored in

Indian languages. It is inflexible and generalization to other languages is virtually impossible. Also manual processing is very expensive and time consuming and we require automatic language identification systems through the use of computers [2][7].

Automatic language identification is a process by which we can identify the language of a written text or speech. Automatic language identification is possible because natural languages are non random and they have regularities. The most widely used technique for automatic language identification is extracting features from a language and based on the existence of these features we can predict the language. The researchers working in this field and published papers using frequency analysis, character n – gram approach, Multi linear regression, Maximum likelihood classifier, Bayesian classifier etc [3][4][5][6][7][8]. Transliteration is a process by which we can convert the languages in one font to another font. Transliteration systems perform one to one mapping of a letter to another letter or a word to another word. After identifying the language of written romanized form we can use a transliteration system to check performance of our identifier [9].The process flow of proposed architecture is shown in figure1

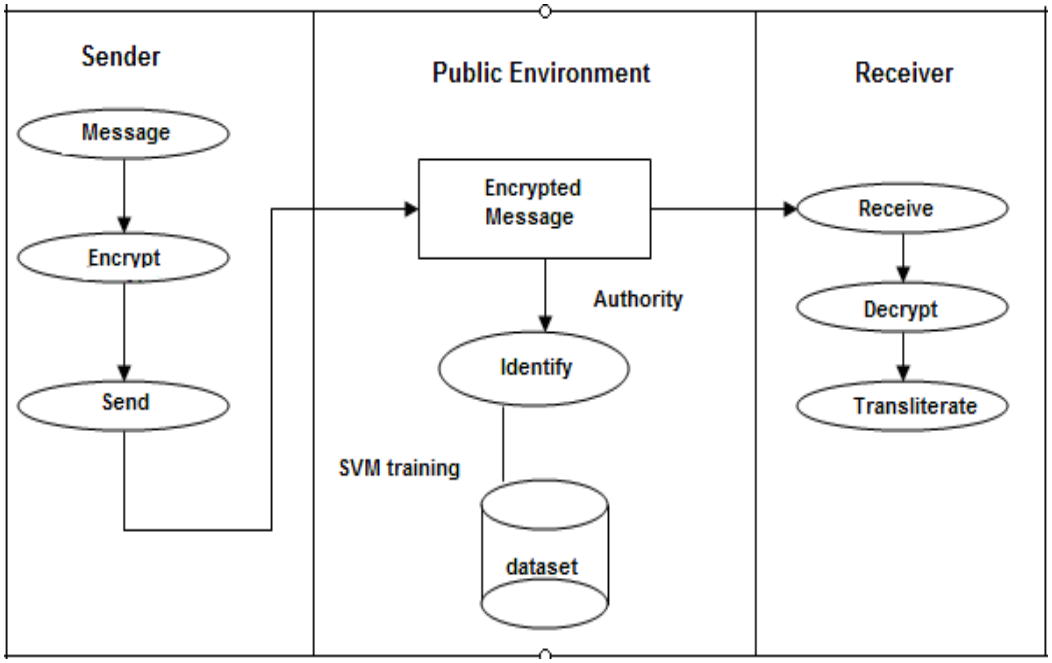

Figure 1: Process Flow of Proposed Architecture

LANGUAGE IDENTIFICATION

Data Collection and Pre – Processing

Many factors will affect the success of Data Mining algorithms. The quality of data is a major concern. If dataset contains irrelevant and redundant information or data is noisy or unreliable then it is a more difficult process to select the attributes. Attribute selection is a process of identifying and removing irrelevant and redundant information from the dataset so that training algorithms operate faster and effectively [10]. If user has background knowledge about data then he can select his own attributes and is better than using attribute selection methods.

For implementing our system we collect 100 romanized SMS of Malayalam and Hindi and Plain English excluding copyright protected and offensive messages from Internet and save it in an excel file. Since there is no standard corpus exist for romanized Indian languages. In a pre processing step we eliminate special characters, commas and numerals in the file. Also avoid the messages which combine romanized form and pure English for this we applied manual evaluation. Then convert all uppercase characters into lowercase and keep the blank space as a valid character.

Feature Extraction

The process of identifying the keywords from sample message is known as feature extraction. These information rich features are used to build the model under the supervision of the class, the language. We extract the features include letter frequency by counting the occurrences of a particular alphabet from the SMS and divide it by total number of alphabets in the message, word frequency by examining the number of distinct words in a message or existence of a pattern word divided by total number of words in the message, frequencies of permissible tri - grams in the message. For this we designed three separate applications. Then create the data set including letter frequency, word frequency and n- gram frequencies of the highest occurrences.

Model Selection

Meta learning has been developed in the field of Data Mining to aid experts in selecting the best algorithms to be used with certain datasets. Algorithm selection is a time consuming task which involves experimentation with different classifiers and analyzing the performance of those classifiers [11]. Studies on the subject try to develop a reliable method to accelerate classifier selection process by using learning algorithms. Here, for selecting the learning algorithm suitable for our dataset we use a previously designed Meta learning framework using weka [12]. The frame work selects SVM as the suitable algorithm for our dataset. Also SVM can take a large number of features and learn to weigh them appropriately. So we can avoid further attribute selection procedure for regression analysis [12].

Experiments

Then we designed an application based on SMO (Sequential Minimal Optimization) algorithm to identify the language. Then test our application based on 43 messages from SMS collections in internet and obtains 94.2381% of accuracy and the confusion matrix is given below in the table1. From these results we can prove that there exist clear distinction between English and natural languages.

Table 1: Confusion Matrix from Data Mining Experiment

a	b	c	← Classified as
15	1	0	a = Malayalam
2	14	0	b= Hindi
0	0	11	c=English

Romanization, Encryption and Decryption

The main focus of our study in the area of cryptography is to propose a new scheme for secret sharing and implement a method similar to Data Encryption Standard (DES) symmetric key cryptosystem. Almost all cryptosystems are crackable and vulnerable to attacks hence we propose a new method combined romanization on natural language and symmetric key cryptography. Our proposed method helps to keep our secret message concealed from unauthorized parties in the communication channel through double encipherment.

Here we romanize the message with natural languages Malayalam and Hindi and Malayalam is one of the toughest languages in the world and only a few people use this and can use to prepare the message. The secret messages in romanized Malayalam are then encrypted using symmetric key algorithm DES so the massage is visible to only the intended person with secret key. Also we know that DES is crackable but at least 256 key searches are required. When key is obtained the analyst can decipher it but the message again in scrambled format and need identifiers or some times they try with other crackers and we can conceal our message from unauthorized parties. The example is given in the table 2.

Table 2: Sample SMS Romanized then Encrypted

Message in English	Romanized Malayalam Message	Encrypted Message
See you tomorrow evening	nale vykeettu kanam	0x4b79a3503a33d46709f187b72c265ae80b7eb7a7538ff9ef
Get ready for a plain hijack	vimaanam ranjan thayyarayirikkuka	0x8752d5533bbb942924e582025decd8436ec781ce7389199831e7b4d0402a43706ac583c5b63a820c
Send lawyers with money and weapons	panavum ayudangalumayi doothane ayakkuka	0x7fadcc3a24c804cc6b040bf4a194a407b4e5a2ee0b9477bd252d63f5292572f9198cbb9b16af3d8f

Transliteration

Systematic transliteration is a mapping from one system of writing into another, word by word, or ideally letter by letter. Here we design transliteration system for Malayalam and Hindi, and maps the sounds in romanization to matching script.

For implementing this create two libraries one for Hindi and one for Malayalam including all character set and corresponding UTF – 8. Then equate English alphabet or combination of alphabets to alphabets in Malayalam and Hindi. When user enters input, the input word is split into consonants and vowels of the language through a split function and map these consonants and vowels to UTF-8 notations corresponding to Malayalam or Hindi in our library. For measuring the performance of transliteration we tested with 50 Malayalam and Hindi words and are transliterated with some spelling mistakes and obtain an overall accuracy of 91%.

CONCLUSIONS

With this work we propose a new cryptographic scheme by which we can exchange information securely through unsecured channel. Create our message through romanization then encrypt it by using a symmetric key algorithm similar to well known Data Encryption Standard (DES). If sender use Malayalam then it is very difficult to hack because Malayalam is one of the toughest languages in the world and only a few people use it. Also well known frequency analysis is not possible in it [13].

Applying Symmetric key Cryptography scheme alone on document does not lead to a complete solution for privacy as well as protection. Moreover, even if the secret keys have been obtained the original message should not be revealed. The suggested method is efficient and enhances the security of a nation against the attacks.

Always all SMS text contain romanized text and very difficult to understand the meaning without a language identifier. If attackers use such messages then scientists can use the identifier to break the secrets. Our system identifies romanized Malayalam, romanized Hindi and Plain English with 94.2381% of accuracy. After identifying the languages as Malayalam and Hindi the system transliterate it to corresponding language font with 91 % of accuracy.

Limitations

Identification limited to two Indian languages Malayalam and Hindi and global language English so the messages in other language gives wrong predictions. Some times system is not able to exactly transliterate into the natural languages handled.

Future Enhancements

Now the system is implemented with symmetric key cryptosystem and need a secure channel for key exchange. In the future we can implement it with public key cryptosystem to avoid the problems in key exchange. The Mathematicians and Computer Scientists used Machine translation techniques to break the secrets of attacker. Future of this work is to

generalize the language identification system for other Indian languages. There is no standard corpus for Indian languages so need to develop a training corpus including more languages.

ACKNOWLEDGEMENTS

A special thanks to Dr A Sreekumar, Department of computer applications, Cochin University of Science and Technology for suggesting the idea.

REFERENCES

1. William Stallings, Cryptography and Network security, (4th Ed: Prentice-Hall, 1999).
2. P K Saxena ,Nveen Gaba, Identification of encryption schemes for Romanized Indian language.
3. Gary Adams ,Philip Resnik, A Language Identification Application built on JAVA Client Server platform,J Burstein and C Leacock (Eds), From research to Commercial applications:making {NLP} work in practice(pp 43-47.Somerset,New Jersey:Association for computational linguistics).
4. Vinosh Babu J , Baskaran S, Automatic Language Identification Using Multivariate Analysis, (Gelbukh Ed, CICLing 2005, LNCS 3406, pp.789 -792, 2005 (c) Springer – Verlag Berlin Heidelberg 2005).
5. Kavi Narayana Murthy, G Bharadwaja Kumar : Language Identification from small text samples,(Journal of Quantitative Linguistics 2006, vol 13,Number 1, pp 57 – 80).
6. Abdel Malek Amine, Zakaria Elberrichi, Michel Simonet, Automatic Language Identification : An alternative unsupervised approach using a new hybrid algorithm,(International journal of computer science and applications, vol.7,No.1 pp 94 – 107,2010,(c) Technomathematics Research Foundation).
7. Shri Kant, Veena Sharma , Neelam Verma ,Identification scheme for Indian languages from their plain and ciphered bit stream (Pre International convention on Mathematical science – Pre ICM 2008).
8. Rafidha Rehiman K A, A Sreekumar , Asurvey of machine translation approach to enhance national security , proccedings of 2nd national conference , NCILC – 2012.
9. M.Hanumathappa , Mallamma V Reddy, Natural language identification and translation tool for natural language processing,(IJSAIT,107-112, 2012).
10. Mark A. Hall and Geoffrey Holmes, Benchmarking attribute selection technique for discrete class data mining.
11. Silviu Cacoveanu, Camelia Vidrighin, Rodica Potolea, Evolutional meta learning framework for Automatic Classifier selection.
12. Phil Antony Mingo , Rafidha Rehiman K A , Kannan Balakrishnan , An autonomous framework for classifier selection in weka (IJECS Volume 2 Issue 3 March 2013 Page No. 696-703).
13. Rajesh Ramachandran, Use of non English language to enhance Network Security,Georgian Electronic Scientific Journal: Computer Science and Telecommunications 2009/ No.5(22).