SMS SPAM DETECTION TECHNIQUES AN ANALYTICAL RESEARCH METHOD

Anchal[#1], Abhilash Sharma[#2]

[#]*CSE Dept, RIMT Mandi Gobindgarh*

[1]anchalgoyal90@gmail.com

[2]abhilash583@yahoo.com

***Abstract*— Use of electronic media for any kind of communication is quite common in today's modern world. SMS (Short Message Service) is a popular and quick service for the communication. The problem occurs when the user does not want to receive a particular text or text from particular type of IDS. The message from the promotional companies makes it annoying for the user. To prevent such kind of message, text classification methods have been proposed. This paper focuses on the text classification methods like tree architecture, ICA algorithm and Neural Network algorithm for the text classification to prevent the user from unwanted spam messages.**

***Keywords*— SMS, SPAM, ICA, NEURAL**

I. INTRODUCTION

SMS is a part of our daily life. People often SMS each other to communicate. SMS can become a problem also if the user does not want to receive a SMS. Promotional companies send bulk SMS to the users which becomes a headache for the user. To identify a sms to be spam certain criteria must be decided. One of the major factors in sms spam detection is the textual analysis of the sms. E messages have become popular means for personal and business communication due to its fast and free availability as well as low or free cost. But several people and companies misuse this facility to distribute unsolicited bulk messages that are commonly called as spam sms. Spam smss may include advertisements of drugs, software, Nigerian scam, adult content, health insurance or other fraudulent advertisements Spam detection problem is becoming more serious now days. It consumes more than half bandwidth of mailboxes. Spam frustrates, confuse and annoy sms users by wasting valuable resources and time. Spam even provides ways for phishing attacks and distributing harmful content such as viruses, Trojan horses, worms and other malicious code. Without a spam filter, one sms user might receive over hundreds of sms daily and find that most of them are of spam category. The spam sms are with no use of sms users. Due to this, serious attention has given to this issue in mailboxes. Several technical solutions like commercial and open-source products have been used to alleviate the effect of this issue.

Spam filtering can be of two types:

- Non-machine learning based
- Machine learning based

Short Message Service (SMS)

SMS is a communication service standardized in the GSM mobile communication systems; it can be sent and received simultaneously with GSM voice, text and image. This is possible because whereas voice, text and image take over a dedicated radio channel for the duration of the call, short messages travel over and above the radio channel using the signaling path [3].

Using communications protocols such as Short Message Peer-to- Peer (SMPP).It allows the interchange of short text messages between mobile telephone devices as shown in figure 1 that describes traveling of sms between parties.

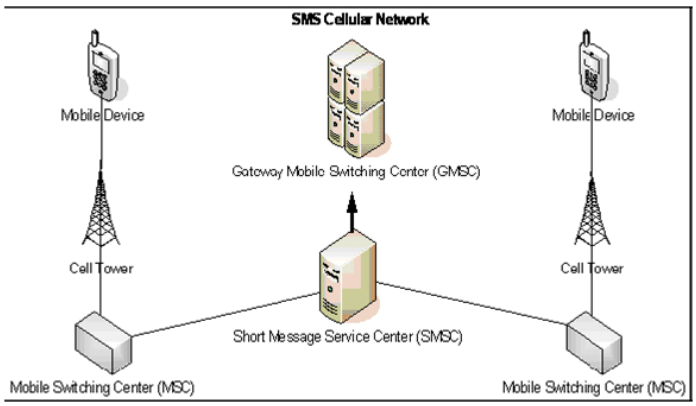

Figure 1: Basic of SMS system

SMS contains some meta-data:

- Information about the senders (service center number, sender number).
- Protocol information (protocol identifier, data coding scheme).
- Timestamp SMS messages do not require the mobile phone to be active and within range, as they will be held for a number of days until the phone is active and within range.

SMS transmitted within the same cell or to anyone with roaming capability. The SMS is a store and forward service, and is not sent directly but delivered via an SMS Center (SMSC). SMSC is a network element in the mobile telephone network, in which SMS is stored until the destination device becomes available. Each mobile telephone network that supports SMS has one or more messaging centers to handle and manage the short messages [3].

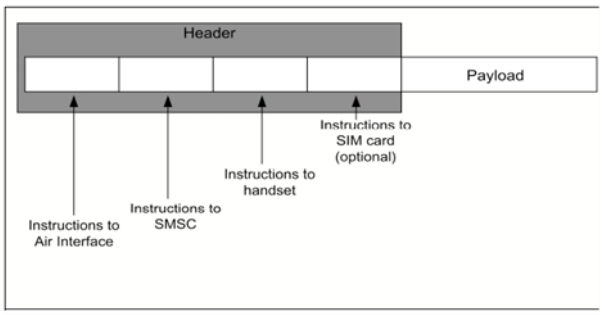

Figure 2: SMS Message Structure

As illustrated in Figure 2, the SMS comprises of the following elements, of which only the user data displayed on the recipient's mobile device:

1) Header - identifies the type of message:
   a) Instruction to Air interface
   b) Instruction to SMSC
   c) Instruction to Phone
   d) Instruction to Subscriber Identity Module (SIM) card.
2) User Data - the message body (payload).

II. TEXTUAL CLASSIFICATION IN SMS

Text classification is a supervised learning process. In this process a task is assign on text data or document for classify this text according to predefined categories or classes according to their contents. For a long time it is very classical problem in information access field, recently this field is attracted due to over loaded amount of text document available in digital form. Some systems are based on text classification like routing, data access, classification, and filtering. There are many numbers of documents available in digital form and day by day availability of documents increasing. These documents represent some information that can be access easily. So to access information from huge amount of documents is very difficult and more time consuming. Organizations that need to access information from huge amount need a technique to solve this difficulty and more work in less time. Data is automatically classified according categories of their contents. There are many algorithm are available to deal with automatic text classification [1].

III. ALGORITHM FOR TEXT ANALYSIS

There are several algorithms for the identification of the data text analysis and the percentage in which they have been copied. Here is a review of some of the finest algorithms.

A. ICA(Increment Component Analysis):

They employ a generalized suffix-tree that can be updated efficiently when the source changes [4]. The amount of effort required for the update only depends on the size of the change, not the size of the code base. Unfortunately, generalized suffix-trees require substantially more memory than read-only suffix-trees, since they require additional links that are traversed during the update operations. Since generalized suffix-trees are not easily distributed across different machines and the memory requirements represent the bottleneck with respect to scalability. Consequently the improvement in incremental detection comes at the cost of substantially reduced scalabilty.

B. AST Based Incremental Method

Nguyen et al. presented [5] an AST-based incremental approach that computes characteristic vectors for all sub trees of the AST for a file. Text analysis is detected by searching for similar vectors. If the analyzed data changes, vectors for modified files are simply recomputed. As the algorithm is not

distributed, its scalability is limited by the amount of memory available on a single machine. A related approach that also employs AST sub tree hashing is proposed by Chilowicz et al. [5]. However, such systems often contain substantial amounts of cloning [3] making text analysis management for them especially relevant. Instead, his approach does not require a parser.

C. Neural Logistics for text classification

The neural networks are non-linear statistical data modeling tools that are inspired by the functionality of the human brain using a set of interconnected nodes [6] networks are widely applied in classification and clustering, and its advantages are as follows. First, it is adaptive; second, it can generate robust models; and third, the classification process can be modified if new training weights are set. Neural networks are chiefly applied to credit card spread sheet data, automobile insurance spread sheet data and corporate fraud. Literature describes that neural networks can be used as a financial fraud detection tool. The neural network fraud classification model employing endogenous financial data created from the learned behavior pattern can be applied to a test sample. The neural networks can be used to predict the occurrence of corporate fraud at the management level [7]**.**

A neural network (NN) is a feed-forward, artificial neural network that has more than one layer of hidden units [7] between its inputs and its outputs. Each hidden unit, j, typically uses the logistic function1 to map its total input from the layer below, xj , to the scalar state, yj that it sends to the layer above.

$$y_i = logistic(x_j) = 1/1 + e^{-x_j},\ x_j = b_j + \sum y_i w_{ij} \quad (1)$$

where b_j is the bias of unit. j, i is an index over units in the layer below, and w_{ij} is a the weight on a connection to unit j from unit i in the layer below. For multiclass classification, output unit j converts its total input, x_j, into a class probability, p_j.

D. Text Based Techniques

Text based techniques perform little or no transformation to the raw source data of spread sheet before attempting to detect identical or similar (sequences of) data. [8].

E. Token Based Technique

Token-based techniques apply a lexical analysis (tokenization) to the source code and, subsequently, use the tokens as a basis for text analysis detection. [9]

IV. CONCLUSION

The paper concludes that there are several methods of text classification to prevent the user from annoying and unauthorized text .This paper has also focused on classification methods like ICA , NEURAL and TREE ARCHITECTURE system . A combination of such algorithms can be helpful in increasing the performance of the classification of the SMS spam.

ACKNOWLEDGEMENT

I would like to express my gratitude to all the people who have given their heart welling support in making this completion a magnificent experience.

REFRENCES

[1] H. A. Basit, D. C. Rajapakse, and S. Jarzabek. “A study of clones in the STL and some general implications. In Proc. of the Int'l Conf. on Software Engineering, “pages 451{459, 2005.

[2] I. D. Baxter, A. Yahin, L. M. de Moura, M. Sant'Anna, and L. Bier. “Clone detection using abstract syntax trees. In Proc. of the Int'l Conf. on Software Maintenance ” pages 368{377, 1998.

[3] K. Beck. “Extreme Programming explained, embrace change” Addison-Wesley, 2000.

[4] Hang Dai and Jingshi He Dongguan “China Research Journal of Applied Sciences, Engineering and Technology”6(5): 895-899, 2013 ISSN: 2040-7459; e- ISSN: 2040-7467 2013.

[5] T. T. Nguyen, H. A. Nguyen, J. M. Al-Kofahi, N. H. Pham, and T. N. Nguyen, “Scalable and incremental clone detection for evolving software,” ICSM’09, 2009.

[6] Ghosh, S., & Reilly, D. L. (1994). “Credit card fraud detection with a neural- network”, 27th Annual Hawaii International, Conference on System Science 3 (1994) 621–630.

[7] Beasley, M. (1996).” An empirical analysis of the relation between board of director composition and financial statement fraud. The Accounting Review”,71(4), 443–466.

[8] J. H. Johnson, “Identifying redundancy in source code using fingerprints,”in Proc. of CASCON ’93, 1993, pp. 171–183.

[9] M. Fisher and G. Rothermel, “The EUSES spreadsheet corpus: a shared resource for supporting experimentation with spreadsheet dependability mechanisms,” ACM SIGSOFT Software Engineering Notes, vol. 30, no. 4, pp. 1–5,2005.

[10] I. D. Baxter, A. Yahin, L. M. de Moura, M. Sant'Anna, and L. Bier, "Clone detection using abstract syntax trees," in Proc. of ICSM '98, 1998, pp. 368–377.

[11] R. Komondoor and S. Horwitz, "Using slicing to identify duplication in source code," in Proc. of SAS '01, 2001, pp. 40–56.

[12] G.D.K.Kishore1, Maddali Sravanthi Automated Anomaly and Root Cause Detection in Distributed Systems.International journal of engineering trends and technology-Volume3Issue1-2012.s