Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 7, July 2015, pg.103 – 108

RESEARCH ARTICLE

Expert System to Detect Suspicious Words in Online Messages for Intelligence Agency Using FP-growth Algorithm

Placida Tellis, N. Deepika

M.Tech (CSE), New Horizon College of Engineering Bangalore

Senior Assistant Professor (CSE), New Horizon College of Engineering Bangalore

***Abstract:** As increase in technology where users are surrounded by social networks sites, it is very effortful job to track the suspicious words which are communicating between user that might lead to cybercrime and other terrorist activity. These Instant Message(IM) are difficult to intercept even by Intelligence Agency, as there are Millions of Messages sent over network. The proposed system predict the suspicious words used in SNS and delivers the detailed report of type of action happening. Ontology based Information extraction (OBIE) to retrieve specific domain and ARM (Association Rule Mining) and FP-growth Algorithm we can list out the frequently used suspicious word.*

***Keywords:** Instant Message(IM), Intelligence Agency(IA), Association rule mining(ARM), FP-Growth Tree, Social networking sites(SNS).*

I. Introduction

The Indulgence of SNS in every aspects of communication makes e-crime a major abstract entity to resolve. Given a easier way of communication users misuse the technology to disregard the agreement of SNS. Cybercrime actions are increased leaving no stone for Intelligence Agency. The CIA, FBI and other federal agencies are constantly integrating domestic and foreign intelligence information to rescue future cyber crime. Internet Crime Complaint Center (IC3) released the report 2012 of cybercrimes, with the current data and trends of online crime activity [1].

The architectures of Smart Phones, Instant messengers and Social Networking sites [2,3]. WordNet, is a lexical database [4], contains a large quantity of words consisting of (155287 words organized in over 117000 Synsets for a total of 207000 wordsense pairs).Ontology based information Extraction technique (OBIE)[5] , which is ruled by with pre-defined Knowledge based rules, then detecting the frequent suspicious words using the ARM[6] and FP-Growth Algorithm[7]which will guide to capture the frequent words sent through the chatters of online messaging. The FP-Growth Algorithm proposed by Han is an efficient and scalable method for mining the complete set of frequent patterns by pattern fragment growth, using an extended prefix-tree structure for storing compressed and crucial information about frequent patterns named frequent-pattern tree (FP-tree).

The big challenge is in automated message surveillance is the recognition of messages containing suspicious words or some other web content. A classic approach to this problem is constructing a set of keywords. Here two difficulties arise First is, it is reasonable to assume that such particular relatively static keywords will not always be present in messages that would otherwise declare as suspicion. Second, there is little guarantee that a sufficiently intelligent individual will not recognize such surveillance is in place and instead use substitute words in place of known keywords.

This section gives the need of the surveillance of online messages and also provides the technique used in this proposed framework. Section II provides the related survey of the system and detailed diagram including the system design. Section III shows the proposed design flow and the algorithm of the framework.

II. Related work and system design

There have been many works done to detect malicious messages, emails, chat and also detect the phishing attacks in SNS. But understanding the logic behind the relationships between offenders can help to identify suspects and understand offender action.

Michael Robertson, Yin Pan and Bo Yuan [8] explained about the social approach to detect malicious content in web technologies for Facebook with security heuristics is limited to identify malicious URL links. Detection of suspicious emails from static messages using decision tree induction proposed which is purely dependent on highest information entropy that identifies the messages are deceptive or non-deceptive [9].

Mohd Mahmood Ali and Lakshmi Rajamani [10] proposed framework with an idea of instant message secure system that identifies suspicious messages by intruders. But does not concentrate on encrypted and code words and short form messages. This paper used various algorithm like tree alignment, stemming algorithm and Apriori algorithm.

John Resig Ankur Teredesai [11] in their paper explored Framework for IM and various data mining issues and how they relate to Instant Messaging and current Counter-Terrorism efforts. And this paper detects does not tend to fully detect suspicious messages, not even detection of topics and social network analysis.

The above papers describe may use different technique to capture suspicious work but nowadays users are uses tricks like sending short words or code words which is not so easily understood by admin. Our proposed algorithm will help to detect these words and try to find the culprits behavior.

System Design

The design of SWDS (Suspicious word detection system) Monitoring system consists of the following sub-systems:

• Instant Messaging system with clients and web browser. Directory Server helps in authenticating a client and contains the identities of the clients. Messaging server where the on line messages are forwarded whenever the chatter starts the next chatting simultaneously using SMTP. Suspicious words from instant messages that are exchanged between the chatters through pre-defined rules used in database.
• SWDS monitoring system applying OBIE and classification rules. Data pre-processing is to extract domain and context using ontology (OBIE). Applying CBA and generate useful association rules.
• Database (Message DB, code word DB, Ontology DB, short word)
The Figure 1 shows design of the SWDS monitoring system in Instant Messaging System (IMS), with the interrelationship among various subsystems in order to detect the suspicious words based on context of the instant message.

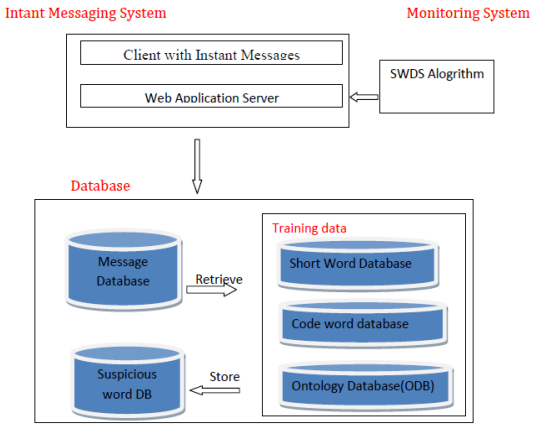

Figure 1. Architecture of the SWDS monitoring system in IM

III. Proposed Framework

Data pre-processing to extract domain using ontology (OBIE):

The pre-processing includes removal of stop words and Root word Extraction.

Stop Words Removal: From each instant message, stop words i.e. words that are not significant are removed, i.e. prepositions, conjunctions, articles, adjectives, adverbs, etc. [12]. Stop word examples are: from, into, in, for, while, a, an, the, that, these, those, under, over, about, although how, what, when, who, whom, etc, stored in database.

Short and code words: After removing unnecessary words system checks for words like 'kl' which means 'kill' short words and code words are stored and detected in chats like if words are 'us' means 'Obama' or 'marriage' means 'meeting'**.** These words are stored by initially in database using brain storming method else the SNS can also update or add any new words occurred in any session**.**

Root word Identification is performed to reduce the derived or inflected words to their stem i.e. to its basic or root form. It is also referred to as conflation. The root words are stored in tree format ie parent-child form restricted to level 2, the child are any suspicious word which are mapped into root word.

Suspicious Identification: Getting the root for each chat messages later finding the frequent pattern or frequent messages sent by user using Association Rule Mining(ARM) and FP-Growth Algorithm, the support and confidence value can be considered as per the requirement to find real suspicious words which are occurred in online messages.

This Phase is important part of this system, where FP-Growth Algorithm provides the frequent words in system, using this we assure that words transmitted are suspicious and can be detected.

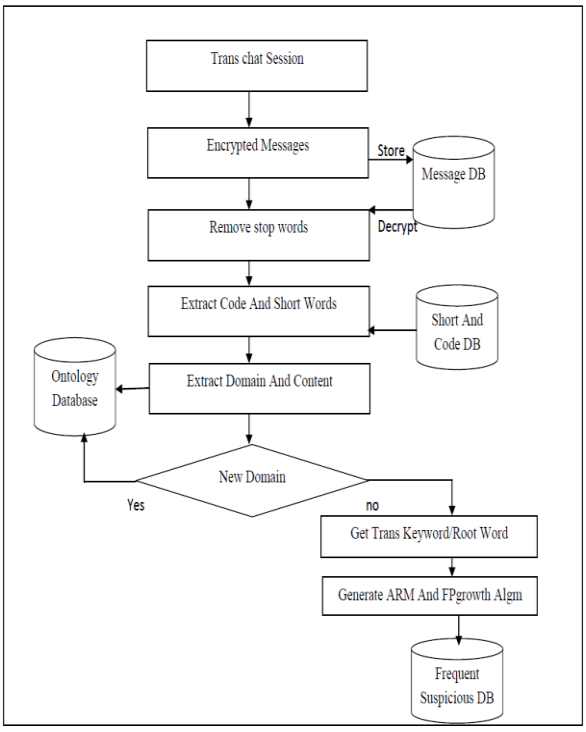

Figure 2 Activity diagram – System workflow of SWDS

Algorithm: Suspicious word detection system

Input: Text messages of chatters
Output: Suspicious words
Steps:
Start
1. Store encrypted message in database.
2. Decrypt the messages //use any encryption/decryption algorithm
3. Remove stop words
4. Search for code and short words
5. Get domain ontology of words from ODB
6. Apply Classification ARM and FP-Growth Algorithm

7. If found new frequent domain go step 8 else step 9
8. Append Semantic Lexicon to ODB.
9. Store and display the filtered suspicious words.
10. Stop

In step1, the messages sent over the network are encrypted form these messages stored physically in the database. And in step2, decryption [13] is applied to messages so that the suspicious words can be extracted.

Step3 and step4 removes the stop words (unnecessary words) and search for the code and short words which are compared with are trained data. Step5 is mapping of stem words to the root words using OBIE and storing in ODB. In step6 applying ARM and FP-Growth Algorithm [14] to retrieve the frequent suspicious words. If found any new suspicious words append to ODB or else store in suspicious words database and display the results.

This technique of illegal activities is analyzed with the help of ontology. Even new code words that are not available in predefined database are also extracted with the help of data mining techniques and added into ontology database. The dataset used here are from brainstorming session which uses GTD(Global terrorist Database) [14] which has recorded information on terrorists.

Conclusions and Future work:

Instant Messaging Systems (IMS) generically cannot detect many suspicious words online hence they are vulnerable for cyber frauds. To overcome this difficulty, a framework for detection of online messages using ontology domain from text messages using Classification Based Association rules and FP-growth Algorithm is proposed. If the proposed framework is integrated at Server side, there will be significant reduction in cyber crime.

The future work consists following:

- Steganography techniques are not detected
- Support for Multilingual languages to be included[
- Integration with HADOOP to solve Big Data problems .

References:

[1](2012).[Online].Available:http://www.fbi.gov/sandiego/press-releases/2012/ic3-2011-internet-crime-report released.
[2] 3GPP2 partners, "Short Message Service over IMS: 3rd Generation Partnership Project 2," developed under 3GPP2, published in 2007.
[3](2012).[Online].Available:[Online].https://wikis.oracle.com/display/CommSuite7RR92909/Developing+an+Instant+Messaging+Architecture.
[4] Jer Lang Hong, "Data Extraction for Deep Web Using WordNet," published by IEEE Transactions on systems, man and cybernetics, 2011.
[5] Daya C. Wimalasuriya, and Dejing Dou,"Ontology-Based InformationnExtraction: An Introduction and a Survey of Current Approaches,"Journal of Information Science,Volume 36, No. 3, pp. 306-323, 2010.
[6] David W. Cheung, and et al., "Maintenance of discovered association rules in largedatabases: an incremental updating technique," published by IEEE in 1996
[7] Tree-based Partitioning of Data for Association Rule Mining Shakil Ahmed, Frans Coenen, and Paul Leng
Department of Computer Science, The University of Liverpool.
[8] Michael Robertson, Yin Pan, and Bo Yuan, "A Social Approach to Security: Using Social Networks to help detect malicious web content," published by IEEE in 2010.
[9] Appavu, and et al.,"Data mining based intelligent analysis of threatening e-mail," published by Elsevier in knowledge-based systems in 2009.
[10] Framework for Surveillance of Instant Messages in Instant messengers and Social networking sites using Data Mining and Ontology Proceeding of the 2014 IEEE Students' Technology Symposium
[11] "A Framework for Mining Instant Messaging Services "John Resig Ankur Teredesai Data Mining Research Group.
[12][online] http://en.wikipedia.org/wiki/Stop_words
[13]E. Thambiraja,G. Ramesh, and Uma Rani, "A Survey on Various Most Common Encryption Techniques," published by IJARCSSE Journal volume 2 issue 7, pp. 226-233, 2012
[14] (2013). [Online].http://www.start.umd.edu/gtd/dowloads/codebook.pdf

[15]Zheng R, Li J, Chen H, Huang Z., "A framework for authorship identification of online messages: writing-style features and classification techniques". Journal of the American Society for Information Science and Technology, February, 57(3), pp.378–93, 2006.
[16] M.Mahmood Ali, and Lakshmi Rajamani, "Framework for Surveillance of Emails to Detect Multilingual Spam and Suspicious Messages," IEEEWorkshop on Computational Intelligence: Theories, Applications and Future Directions , IIT Kanpur, India, pp. 42-56, 2013.

