An Ontology based Surveillance Approach for Detecting Suspicious Messages using Data Mining

Granthana.KS
PG Student, Dept. of CSE
APS College of Engineering, Bangalore
Karnataka, India

Somasekhar.T
Senior Lecturer, Dept. of CSE
APS College of Engineering, Bangalore
Karnataka, India

Abstract: **Nationwide, as clients have become increasingly technology-dependant, there is an increasing number of clients using IM as a primary source of communication. Despite the popularity that instant messaging is gaining in the mainstream social culture, most research only looks at general online interaction and its consequences on face to face social interactions. Detecting and identifying any phishing (suspicious) messages in real-time, is really a complex and dynamic problem involving many factors and criteria. Innumerable error and suspicious messages are sent through Instant Chat Messengers (ICM) which is untraced, leading to hindrance network communications and cyber security. We propose a work that discover and predict such messages that are sent using IM. Further, these instant messages are put under surveillance that identifies the type of suspected cyber threat activity by culprit along with their personnel details. This thesis proposes an Ontology-based approach using Ontology Based Information Extraction technique (OBIE), Association rule mining (ARM) a data mining technique with set of pre-defined Knowledge-based rules for detecting suspicious messages using Data Mining. The experimental results obtained will aid to take prompt decision for eradicating cyber crimes.**

Keywords- Instant chat Messengers (ICM); Ontology based Information Extraction(OBIE); Association Rule Mining(ARM); Knowledge based rules.

I. INTRODUCTION

The Internet has revolutionized the computer and communications world like nothing before. The invention of the telegraph, telephone, radio, and computer set the stage for this unprecedented integration of capabilities. The Internet is at once a world-wide broadcasting capability, a mechanism for information dissemination, and a medium for collaboration and interaction between individuals and their computers without regard for geographic location. The Internet represents one of the most successful examples of the benefits of sustained investment and commitment to research and development of information infrastructure. Beginning with the early research in packet switching, the government, industry and academia have been partners in evolving and deploying this exciting new technology [1].

Computer crime, or cybercrime, is any crime that involves a computer and a network.[The computer may have been used in the commission of a crime, or it may be the target. Net crime is criminal exploitation of the Internet [2].

Dr. Debarati Halder and Dr. K. Jaishankar (2011) define Cybercrimes as: "Offences that are committed against individuals or groups of individuals with a criminal motive to intentionally harm the reputation of the victim or cause physical or mental harm to the victim directly or indirectly, using modern telecommunication networks such as Internet (Chat rooms, emails, notice boards and groups) and mobile phones (SMS/MMS)". Such crimes may threaten a nation's security and financial health. Issues surrounding these types of crimes have become high-profile, particularly those surrounding hacking, copyright infringement, child pornography, and child grooming. There are also problems of privacy when confidential information is intercepted or disclosed, lawfully or otherwise [2].

The E-crime department must be improvised with the development of technology to find criminals. Many of the Instant Messaging Systems (IMS) developed restricted their limit for sending messages, video and audio conferencing. They are not well equipped to detect online suspicious messages [3].

While surveying various architectures of Instant messengers helped us to develop a new Framework. WordNet, is a lexical database, contains a huge amount of information consisting of (155287 words organized in over 117000 Synsets for a total of 207000 word-sense pairs) words that is useful for our study for scanning and filtering the text messages stored in TDB (Text Database) [4]. WordNet is used as features for classification of words from unstructured text. Similarly, WordNet Ontology based on information extraction technique is discussed in [5]. Our Contribution includes improving the existing IMS using data mining technique of Associative rules [6], Ontology based information retrieval technique (probabilistic models), which is guided with pre-defined Knowledge based rules and ARM. Early detection of suspicious messages from instant messaging systems is possible with our proposed Framework to identify and predict the type of cyber threat activity and trace the criminal details [3].

II. PROBLEM STATEMENT AND RELATED WORK

Cybercrime is a fast-growing area of crime. More and more criminals are exploiting the speed, convenience and

anonymity of the Internet to commit a diverse range of criminal activities that know no borders, either physical or virtual [4].

New trends in cybercrime are emerging all the time, with costs to the global economy running to billions of dollars [4].

In the past, cybercrime was committed mainly by individuals or small groups. Today, we are seeing criminal organizations working with criminally minded technology professionals to commit cybercrime, often to fund other illegal activities. Highly complex, these cybercriminal networks bring together individuals from across the globe in real time to commit crimes on an unprecedented scale [4].

Criminal organizations turning increasingly to the Internet to facilitate their activities and maximize their profit in the shortest time. The crimes themselves are not necessarily new – such as theft, fraud, illegal gambling, sale of fake medicines – but they are evolving in line with the opportunities presented online and therefore becoming more widespread and damaging [4].

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. Similarly, architecture detection of suspicious messages from IM for text messages using Data mining technique has been proposed.

III PROPOSED FRAMEWORK FOR SUSPICIOUS MESSAGES

In this Section, we explore the operational phases of proposed Framework as shown in Fig. 1. The Suspicious proposed Framework as shown in Fig. 1. T he Suspicious Message Detection (SMD) algorithm initiates the steps to capture the instant messages that are communicated between the users and then , stores them into database for identifying suspicious messages.

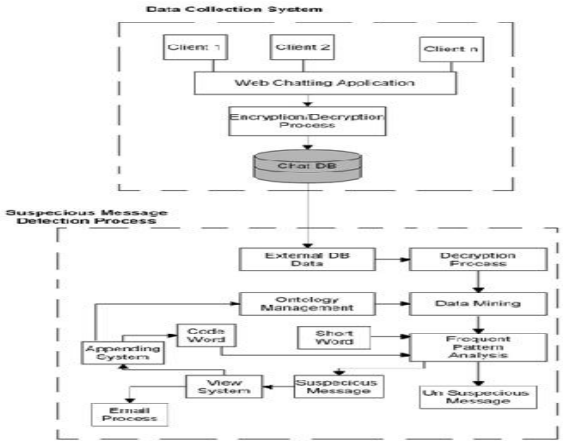

Fig. 1. Proposed Framework to detect suspicious messages from Instant Chat Messenger (ICM).

The architecture of instant messenger to detect suspicious messages is shown in the figure 1. It consists of two phases:

1. Phase 1: This phase is named as data collection system. In this phase, users/clients communicate with their friends, colleague's etc using instant messenger i.e. a chat application. The messages exchanged are collected (data) and stored in a database in an encrypted format.
2. Phase 2: This phase is used to detect any suspicious messages are communicated between users. First thing we need to do is extract data (messages) from the database which is in an encrypted format. Once we decrypt the data, we will obtain an original message content. After decryption, we apply data mining techniques such as ontology management and replacement, frequent pattern analysis, association rule mining and also we are to check any code words or short words in the messages. Once messages pattern are detected, we will find out whether it is suspicious or unsuspicious messages. If it is suspicious, we can report to the crime department with the user id and other details.

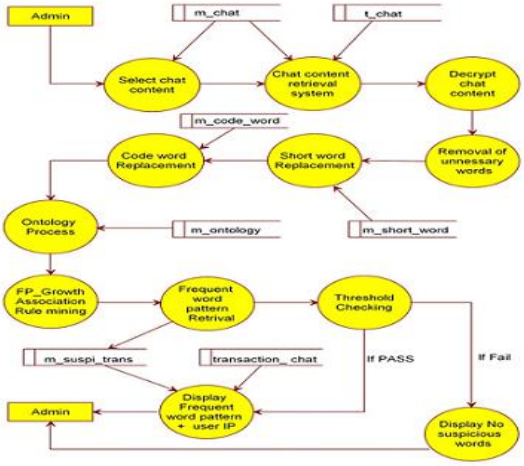

Fig 2: shows the Schematic cum algorithmic representation for proposed Framework for tracing criminals from instant text messages named as SPD algorithm.

The working pseudo code of our Framework is demonstrated using Schematic-cum-algorithmic representation is shown in Fig.2. Clients/users communicated via web chatting application (instant messenger), are stored in a database which is in an encrypted format. From the database, admin select the chat content and then decrypt the chat content so that the original chat content is viewed. Then data mining techniques are applied. Steps of algorithm are illustrated as follows:

1. First, Removal of unnecessary words in the chat messages like prepositions, conjunctions.
2. Ontology management and replacement which is done through OBIE method

3. Then detects, if any short words[26] or code words are in the messages.
4. Once the above three process is done, we check out for any frequent pattern/words are in the content, then we apply FP_GROWTH algorithm (fig 4) with the help of association rule mining to obtain support and confidence value of the chat content. Then obtained value is compared with the threshold value.
5. If we find any suspicious, then we will view the email id of the user who has sent messages. And displayed to admin.

The OBIE plays a crucial role that predicts an d maps the domain (topic) to which these suspicious words belong. Metadata is the essential component, that maintains information of all databases used, users information to whom the message belongs and other relevant information pertaining to Framework (time, date, receivers and senders details, etc.). Just like a log of history maintained by most of IMS. The pre-defined rules of Table I, specifically *rule 1* is given to OBIE, using associative rules, are framed carefully by analyzing brainstorming session of real time datasets that are taken from FBI and CBI investigations of solved cases and GTD.

Table I. Shows *3 rules* to be satisfied by OBIE Model while extracting and

RULE 1 (Pre-defined Knowledge based rules)	
Type of threat activity (Domain)	**Stem words to be detected in a given Context**
Murder →	kill, assault, assassinate, eliminate, gun dagger, knife, stab, location, money
Kidnap →	Hijack, capture, seize, abduct, usurp, grab, gun, take_hostage, location, amount, kill,property
Terrorist attack →	Bomb, vehicle, location, suicide_attack, bag, holy_place, laptop, demolish, payment , cash
Drug supply & Smuggling →	Packet, brown_sugar, cash, cocaine, hashish, M.tabs, Methoquoline, opium, charas, location, injection,Morphine, LSD STR/ECA, dibucaine
Match Fixation→	Location, luxurious_flat, cash, hotel, bet, gifts (car), virgin_girl,bank, payment, loose_game
Corruption charges →	Luxurious_flat, money, bank, cheque, deposit, diamond, avoid_tax, laptop, offshore_account
Robbery & theft→	Jewelry_shop, place, bank, night, gun, knife, location, vehicle, break, night, keys, locker
sexual harassment→	Phone_messages, beautiful, come, payment, spend_night, location, jewelry_items, park hotel, car_gift, body_parts, property, help,Job
RULE 2 (undetected words)	
Ambiguous and undetected words to be checked and corrected automatically based on nearness of stem words using ontology taxonomy constructed [5] by *rule 1*. [In OBIE this *rule 2* is applied to Information Extraction Module (TPDB) before sending to knowledge database (KDB)] mapping the stem words to Domain.	

IV PRELIMINARIES

A. Word Extraction From Unstructured Text Using Ontology

1) Information Theory principle using probability

Ontology makes use of probabilistic topic models, to organize words under a topic (domain) into a meaningful hierarchy. The *GSHL algorithm*, builds the concept hierarchy based on the principle of information theory. Kullback-Leibler, divergence (*DKL(P||Q))*(also known as relative Entropy) is given first to underpin the principle for establishing relationships between topics. Following the Gibbs inequality, KL divergence value is to be: $D_{KL} > 0$.

2) Topic (Domain) Hierarchy Construction Algorithm

Having, defined the information principle, for establishing a relationship between topics, the next step is to organize topics into hierarchies. Wang Wei and et al. developed Global Similarity Hierarchy Learning (GSHL) algorithm. This *GSHL*, recursively searches for most similar topics of the current "root" topic and removes those that do not satisfy the condition on difference of K_L divergence. GSHL, start with an initial topic as the root node and look for top *n* most similar topics according to (*dis*)similarity measures. The parameters used in algorithm are as follows:

N — The total number of topics (domains i.e. root words).
M_c — The maximum number of sub-nodes (words) for a particular node (root node, i.e. domain/topic).
TH_s and TH_d — The thresholds for similarity and divergence measures.
TH_n — The noise factor, defined by the difference between two K_L divergence measures D_{KL} *(P||Q)* and
$D_{KL}(Q||P)$.
I — Maximum number of iterations.

The parameters *THs*, *THd*, *THn* are user-specified constants, which are tuned to obtain desirable precision and accuracy values. Specifically, in our experiment, we have found that setting *THs*, *THd*, and *THn* within some narrow range results in only a slight variation of precision values. The pairwise measures of Cosine similarity, JS divergence, and KL divergence are collectively denoted as the ***Ms*** matrix. The algorithm will terminate according to the conditions specified in the while loop. The pseudo code for *GSHL* algorithm is shown in Fig. 3.

Algorithm 1. *GSHL (root)*
Require: Initialize *V*, *Ms*, *I*, TH_s, TH_d, TH_n, and M_c.
Ensure: A terminological ontology with "broader" and "related" relations.

```
1  Initialize V, Ms, I, THs, THd, THn, and Mc;
2  while (i < I and V is not empty) do
3    Add current root into V ;
4    Select most similar Mc nodes (words) of
5        root word (topic)
6    from Ms;
7    Add similar nodes into Vtemp;
8    Remove nodes in Vtemp against//similarity and divergence
9                                    //difference condition
10   for (all nodes n ,in Vtemp) do
11     if (Sim(ni, root) > Sim(ni, Sibling(root))) then
12       Assert broader relations between root and
13           topic ni;  //relationship between topics is broader
14     else default
15      Assert related relation between root and
16           topic ni;  //relationship between topics is nearer
17     end if
18     Move topic ni  from Vtemp to V ;
19     Increment i by 1;
20   end for
21   Remove current root from V ;
22 end while
```

Fig. 3. Shows the terminological ontology algorithm to find the root word from domain (topics) using threshold value.

3) OBIE Model for Root word (Domain) Extraction

Ontology represents knowledge as a set of concepts within a domain and finds the relationships among those concepts. It is used to reason about the words within that domain (topic) and describe the domain. The hidden suspicious words are explored from these messages and domain *(Murder, Kidnap, Terrorist attack, Drug supply and smuggling, Match fixation, corruption charges, Robbery & theft,* and *Sexual harassment)* is found using ontology in our framework.

3) FP-tree construction to find frequent pattern in the text:

Algorithm 2: FP-Growth

Input: A database DB, represented by FP-tree constructed according to Algorithm 1, and a minimum support threshold ?.

Output: The complete set of frequent patterns.

Method: call FP-growth(FP-tree, null).

Procedure FP-growth(Tree, a) {

(01) if Tree contains a single prefix path then // Mining single prefix-path FP-tree {

(02) let P be the single prefix-path part of Tree;

(03) let Q be the multipath part with the top branching node replaced by a null root;

(04) for each combination (denoted as ß) of the nodes in the path P do

(05) generate pattern ß ∪ a with support = minimum support of nodes in ß;

(06) let freq pattern set(P) be the set of patterns so generated;

}

(07) else let Q be Tree;

(08) for each item ai in Q do { // Mining multipath FP-tree

(09) generate pattern ß = ai ∪ a with support = ai .support;

(10) construct ß's conditional pattern-base and then ß's conditional FP-tree Tree ß;

(11) if Tree ß ≠ Ø then

(12) call FP-growth(Tree ß , ß);

(13) let freq pattern set(Q) be the set of patterns so generated;

}

(14) return(freq pattern set(P) ∪ freq pattern set(Q) ∪ (freq pattern set(P) × freq pattern set(Q)))

}

Fig 4: FP-tree construction to find frequent word pattern in the text messages from the database

3) DES algorithm to encrypt and decrypt messages that are sent using ICM:

To encrypt and decrypt a file using the Data Encryption Standard Algorithm, one should perform the following steps:

1)Create a KeyGenerator for the DES algorithm and generate a secret key.
2)Create an IvParameterSpec object, which is an implementationofthe AlgorithmParameterSpec Interface, a specification of cryptographic parameters.
3) Create two Cipher objects, one to implement the encryption and the other one for the decryption. Both Ciphers must be initialized to encryption/decryption mode, with the key and the algorithm parameters defined above.
4) Create a FileInputStream to read the file to be encrypted and a FileOutputStream to write the encrypted file.
5) Read data from a FileInputStream into a byte array.
6) Encrypt the byte array. Create a new CipherOutputStream using the encryption cipher and the byte array. The CipherOutputStream encrypts data before writing it out to an OutputStream,as shown in the write_encode(byte[], Outputstream output) method of the example.
7) Create a FileInputStream to read the above encrypted file that will now be descrypted.
8) Decrypt the file. Create a new CipherInputStream using the decryption cipher and the byte array .The CipherInputStream will read in the byte array and decrypt each byte before returning it. This is demonstrated in the read_decode(byte[], InputStream input) method of the example.

Fig 5: shows DES algorithm for encrypting and decrypting text messages

V. EXPERIMENTAL RESULTS

A. Evaluation method for datasets

We used Precision metric [20] to evaluate our Framework. The extracted suspicious words efficacy are based on two factors, the number of actual words available in the pre-defined database, to that of the number of words from user generated extracted testbeds.

$$\text{Precision } (P) = \frac{\text{Correctly Extracted}}{\text{Total Extracted Correctly}}$$

$$\text{Recall}(R) = \frac{\text{Correctly Extracted}}{\text{Total no. of Possible Words}}$$

B. Preparation of datasets and results obtained

The Terrorist Attack (Domain) dataset is taken from Global Terrorism Database (GTD) which has recorded information on terrorist events around the world since 1970 to till date. The complete representation related to terrorist attacks is found using *CODEBOOK[21]*. We obtained dataset using brainstorming session from domain experts using GTD that consists of 59787 rows, size of 30MB, and 7 columns

taken out of 99 column s, named it as User Generated Content (UGC) i.e. UGC-testbed-1 and tested with our Framework. The outputs obtained are shown in Table II.

Table II. Outputs Obtained from UG C-testbed-1 Dataset

Terms	Framework Out ut
Total Extracted Correctly	**1779**
Correctly Extracted	**1703**
Total Possible words extracted	**1732**
Precision	**95.72**
Recall	**98.32**

Dataset used are manually created b y brain storming session from domain experts using GTD, as we could not able to get re al suspicious contents that are stored in hi story from IM and SNS, due to authorization restriction [10].

C. Comparison of our framework with existingIMS

Currently none of Instant Messengers has the ability to detect suspicious messages during online chat. The features based on which our Framework is compared with IM(ICM) are shown in Table III.

Table III. Comparison of our Framework with (IM, SNS & Apps)

Features	IM	Proposed Framework
Cyber threat Activity Detection	Static Detection (time consumed)	Dynamic Detection
Report Generation for E-crime department	No Report	Report with details (Email-id, Phone No., etc.)
Ontology support	No	Yes
Dynamic Location Mapping based on ISP and IP address	No	Yes (using R2D Wrapper)
Efficiency	Very Good	Moderate (as online messages are monitored & stored)
Database & Data Mining support	No	Yes
System Architecture	Easy to Design	Complex to design
Code words and short words	Not detected	Detected
Encryption and decryption	No	Yes

V. CHALLENGES AND FUTURE WORK

Framework aids the E-crime department to identify suspicious words from cyber messages and trace the suspected culprits. Currently existing Instant Messengers lack these features of capturing significant suspicious patterns of threat activity from dynamic messages and find relationships among people, places and things during online chat, as criminals have adapted to it [10]. The User Generated Content (UGC) testbed is proven to be useful, for monitoring terror and suspicious crimes in cyberspace which provides national and international security. We used simple English terms like *kill, murder, etc*. But, in practical scenarios these words are in specific coding language, for example "*picnic*" is used instead of *"kill"*.

Issues and challenges of our Framework are:

1. The suspicious words sent in Steganography techniques are not detected and hence neglected as ignore words.
2. Support for Multilingual languages to be included [24]. The media may also actively participate in transmitting messages to terrorists and criminals indirectly, via news papers and TV channels (Text, Audio and Video) unknowingly [25].
3. Integration with HADOOP to solve Big Data problems.

If the proposed Framework integrated with existing IM at Server-side, for surveillance will change the world of cyberspace to rest in peace without cyber crime [10].

REFERENCES

[1] http://www.internetsociety.org/internet/what-internet/history-internet/brief-history-internet.
[2] http://en.wikipedia.org/wiki/Computer_crime
[3] "Framework for Surveillance of Instant Messages in Instant messengers and Social neworking sites using Data Mining and Ontology" by Mohammed Mahmood Ali, Mohammed Mahmood Ali, Lakshmi Rajamani.
[4] http://www.interpol.int/Crime-areas/Cybercrime/Cybercrime.
[5] (2012).[Online]. Available: http://www.fbi.gov/sandiego/press-releases/ 2012/ic3-2011-internet-crime-report-released.
[6] 3GPP2 partners, "Short Message Service over IMS: 3rd Generation Partnership Project 2," developed under 3GPP2, published in 2007.
[7] (2012). [Online]. Available: [Online]. https://wikis.oracle.com/display/CommSuite7RR92909/Developing+an+Instant+Messaging+Architecture.
[8] (2012).[Online].Available: http://www.ontologyportal.org/
[9] Daya C. Wimalasuriya, and Dejing Dou,"Ontology-Based Information Extraction: An Introduction and a Survey of Current Approaches," Journal of Information Science,Volume 36, No. 3, pp. 306-323, 2010.
[10] M. Mahmood Ali, and L. Rajamani, "Framework for surveillance of instant messages, " published by inderscience in IJITST, vol. 5, 2013.
[11] 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.
[12] (2012). [Online]. Available: http://www.digitaltrends.com/social-media/facebook -scans-chats-and-comments-looking-for-criminal-behavior/
[13] Appavu, and et al.,"Data mining based intelligent analysis of threatening e-mail," published by Elsevier in knowledge-based systems in 2009.
[14] M. Mahmood Ali, and L. Rajamani, "Phishing Detection in Instant Messengers using Data Mining Approach," proceedings of ObCom 2011, published by Springer-Verlag Berlin Heidelberg 2012, part I, CCIS 269, pp. 490–502, 2012.
[15] Sunitha Ramanujam, and et al., "A Relational Wrapper for RDF Reification," E. Ferrari et al. (Eds.): TM 2009, IFIP AICT 300, pp. 196–214, IFIP International Federation for Information Processing 2009.
[16] (2012). [Online]. Available: http://www.webconfs.com/stop-words.php
[17] M.W.Du, and S.C.Chang, "An Approach to Designing Very Fast Approximate String Matching algorithms," IEEE journal, 1994.
[18] Y. Zhai and B. Liu, "Web data extraction based on partial tree alignment," in Prococeeding of ACM, 2005.
[19] Jer Lang Hong, "Data Extraction for Deep Web Using WordNet," published by IEEE Transactions on systems, man and cybernetics, 2011.
[20] C.D. Manning, P. Raghavan, and H. Schutze, Introduction to Information Retrieval, Cambridge Univ. Press, 2008.
[21] (2013).[Online].http://www.start.umd.edu/gtd/dowloads/codebook.pdf.
[22] (2012). [Online]. Available: http://dir.yahoo.com/Society_and_Culture/Crime/Types_ of_Crime/
[23] 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.
[24] M. Mahmood Ali, and Lakshmi Rajamani, "Framework for Surveillance of Emails to Detect Multilingual Spam and Suspicious Messages," IEEE Workshop on Computational Intelligence: Theories, Applications and Future Directions , IIT Kanpur, India, pp. 42-56, 2013.
[25] Chaditsa Poulatova, "The Media: A Terrorist Tool or a Silent Ally," published by IEEE in 2011.
[26] http://chatworddictionary.com