ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

Design of Suspicious User Profile Identification system Using ACO Algorithm

[1] Asha, [2] Dr. Balkishan
[1] [2] Department of Computer Science and Application, Maharshi Dayanand University, Rohtak, India

Abstract: **- The aim of this paper is to design a system for the identification of suspicious user profiles using Ant Colony Optimization algorithm. The communication technologies and their advancements have greatly influenced our daily lives. The technologies like social networking websites, blogs, chat forums, instant messengers and many more are leading various fascinating trends in today's world. Unfortunately, the increase in suspicious activities is one of the major causes due to the misuse of the technology. The internet is loaded with very large amount of data. Some people make use of the technology to spread rumours, to spread violence, bully other people, spread hate messages or even perform criminal activities like financial frauds etc. and thus increasing the amount of suspicious content over the internet. In this paper, textual data from social networking websites is considered to detect the suspicious user profiles. A suspicious user profile (SUP) detection system is proposed which considers the text based data as input and retrieves the suspicious user profile based on the extracted features.**

Index Terms **-Suspicious User Profile, Swarm Intelligence, Ant Colony Optimization, Social Media Dataset, Online Users, Optimization.**

I. INTRODUCTION

The world of internet technology has facilitated people to interact with each other anytime and anywhere. The use of social networking websites, chat forums, instant messengers and many more has increased and certainly these are the most impressive techniques of communication. People are very keen about the internet technology and its usage. Unfortunately, just like the two faces of the coin the technological use has its dark side too. Some people misuse the technology to intrude other people's life and this has increased the suspicious over the internet. The term suspicious content can be defined as any uninvited web content which may lead to criminal offence [1], fraud, spreading rumours with the motive of misleading people or to grow terror among people, bullying other people or spreading hate messages or any other inappropriate activity. Researchers over the globe are analysing the internet data in order to find out the suspicious content and then to stop the occurring of suspicious activity [2] before any loss. This paper aims to not only find the suspicious content but also the user profiles associated with that content. The text based web data is considered from different sources including social networking websites, blogs etc. to detect the suspicious user profiles. The unstructured web data is pre-processed and features are extracted for the detection. The classification and optimization of results is performed with swarm intelligence based Ant Colony Optimization algorithm. The paper is structured in the form of 5 sections. Section 2 discusses the research work of numerous authors in the field of suspicious activity detection. In section 3 some basic concepts are explained. Section 4 includes the proposed Suspicious User Profile (SUP) detection system and section 5 concludes the paper.

II. RELATED WORD

This section presents the work of some authors in the field of suspicious activity detection. The work of the authors is analysed and is presented in table 1. All of the authors presented have worked or proposed some systems in order to detect the suspicious activity, any abnormal activity or human behaviour, criminal activity etc. In table 1, the analysis is done by presenting the techniques used, dataset considered and the work done by the authors. Khangura et al [3] have proposed a suspicious activity detection system from the textual data collected from the online available chat logs. The proposed system is implemented with the help of support vector machine classifier and then further it is optimized by using the concept of genetic algorithm (GA). In the experimentation, unstructured chat log data is pre-processed and then features are extracted using a rule based engine. The three key features extracted are session, vocabulary terms and users. Further suspicious activities are classified with the help of SVM and genetic algorithm is applied for result optimization and also to detect the user profiles associated with the suspicious activities. The

integrated approach of SVM and GA has achieved a performance accuracy of 74.3%. Another suspicious activity detection system was introduced by the authors Chiu et al [4] which identifies the presence of suspicious blocks in the multimodal dataset. The authors find it more beneficial to find out the suspicious block and then the suspicious activities present in those blocks. In this way authors are able to find group of suspicious activities belonging to a particular single source. The experiments are conducted based on two different datasets and also using two algorithms. The dataset considered for the experimentation consist of real world and synthetic data. The two datasets are implemented with the CrossSpot and CrossSpot + algorithms and the experiment results have achieved high precision rate even in the case of incomplete or partially filled data. The authors Patil et al [5] have proposed a prototype to encounter suspicious or abnormal human behavioural activities. The proposed system considers the live data from the camera and compares the normal human behavioural aspects like running etc. with the abnormal aspects and thus detects the change in normal behaviour using SURF technique. The authors have suggested that this prototype is of great use in many of the public and private places including public malls, airports, railway stations etc. The authors Jiang et al. [6] have proposed a suspicious and fraud information detection system using the Twitter's hashtag dataset. A metric based approach is used to find out the data block containing suspicious data based on Erdos-Renyi-Poisson model. The experimentation is performed using CrossSpot algorithm. further, the results obtained using CrossSpot algorithm are compared to the result values obtained by Singular Value Decomposition (SVD) and High Order Singular Value Decomposition (HOSVD) and the comparison results shows that CrossSpot has shown efficient results. The authors Alami and Beqqali [7] have analysed text based web data to find out the suspicious text as well as the user profiles with suspicious behaviour. The dataset used for the experimentation is Twitter dataset containing millions of user profiles and tweets. Initially, the data is pre-processed by removing stop words and to find out the similarity between the user's posts the concept of Normalized Compression Distance (NCD) is used.

Tayal et al [8] have worked on a crime detection and crime identification system (CDCI) which detects the criminal activities of Indian cities. Initially, unstructured crime data is collected from different crime web sources and then the unstructured data is pre-processed. Further, the data is clustered into two with the help of k-means clustering technique. These clusters are formed according to the similarity in data. The use of Google maps is also done to make the visualization more clear and effective. Finally the data is classified with the help of K-Nearest Neighbour (KNN) algorithm to get the effective results. The map representation contains three coloured dots which are blue, yellow and red in colour. The blue colour dots in the map represents those Indian cities where numbers of crimes are less than 50. The yellow dots represents those cities where crime number is between 50-100 and red dots are the cities where crime number is the most i.e. greater than 100. Another crime identification system is proposed by the authors Gowri et al [9]. The authors have considered the web based email and chat data to detect the crime based data with the objective to help the security departments and agencies. The gathered data is cleaned and pre-processed and also a dictionary is prepared which contains suspicious words in order to effectively check the presence of suspicious words in the online collected data. The experiment is conducted with the help of R programming language.

Table 1: Analysis of work of different authors on suspicious activity detection

Author and Year	Dataset	Technique	Work done
Khangura et al [3] (2017)	• Chat logs data	• Support Vector Machine and • Genetic Algorithm	A suspicious activity as well as associated user profile detection system from chat data with vocabulary term, users and sessions as feature vectors.
Chiu et al [4] (2017)	• UTD-MHAD dataset	• CrossSpot algorithm • CrossSpot + algorithm	A block data based suspicious activity detection system to find activities belonging to particular single source.
Patil et al [5] (2017)	• Live Camera data	• Speeded Up Robust Features (SURF)	A prototype for abnormal behaviour detection from live camera data using SURF descriptor.

Jiang et al. [6] (2016)	• Twitter's hashtag dataset	• Erdos-Renyi-Poisson (ERP) model • CrossSpot algorithm.	A metric based approach based on ERP model to detect the data block of suspicious words by implementing using CrossSpot algorithm.
Alami and Beqqali [7] (2015)	• Twitter dataset	• Normalised Compression Distance	A NCD based suspicious content and user profile detection system from Twitter text based dataset.
Tayal et al [8] (2015)	• National Crime records Bureau • Committee to protect Journalists	• K-means clustering • KNN classifier	A CDCI system to detect the number of crimes in Indian cities with blue (<50), yellow (50-100) and red (>100) colour dots representation indicating the number of crimes in respective city using Google maps.
Gowri et al [9] (2014)	• Email data • Chat messages	• R programming language	A dictionary is developed consisting of suspicious words in order to find the presence of suspicious words in the email or chat data with the objective to help security departments and agencies.

III. PROCEDURE FOR PAPER SUBMISSION

In this section the basic idea about the concept of Swarm Intelligence and swarm intelligence based Ant Colony Optimization technique is stated.

A. Swarm Intelligence

Swarm Intelligence is a nature based discipline and its methodologies are inspired from the natural species like ants, bees, bats etc. According to the authors Bonabeau et al [10] the methodologies in swarm intelligence are collective groups of intelligence with simple agents. The agents work in collaboration with each other. Each agent work individually or has its own way of performing task but the collective work of all agents is well organised and synchronised as well [11]. Any swarm intelligence system is based on the following three fundamental principles.

1. Feedback: The nature of the solution is decided on the basis of feedback provided. A positive feedback results as a good solution whereas negative feedback corresponds to the elimination of the poor solution.
2. Multiple Interactions: The agents have multiple interactions with each other in order to find the best possible solution.
3. Randomness: This property exhibits to the testing and creation of new solutions.

B. Swarm Intelligence Techniques

Since the concept of swarm intelligence is based on the natural species, its techniques and their working is also based on the social species. There are numerous swarm intelligence based techniques based on the animals and insects [12]. Figure 1 shows some of the popular swarm intelligence techniques.

C. Ant Colony Optimization

In this paper, the implementation of the proposed system is conducted using swarm intelligence based ant colony optimization (ACO) technique. ACO is a graph based probability algorithm, where the system agents are ants and initially the concept was introduced by the authors Maniezzo, Colorni and Dorigo [13] [14] [15]. The main idea of the algorithm is based on the natural behaviour of ants. They exhibit parallel search individually to find out the local solution of the problem data [16]. After performing the local search step the value of the pheromone trail is updated. Due to evaporation the trace values are decreased and the trace value at the component which is used to construct the solution is increased by placing the pheromone [17]. Due to the synchronized and collective behaviour of ants, they interact with each other and thus generate a combined optimal solution based on the heuristic information

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

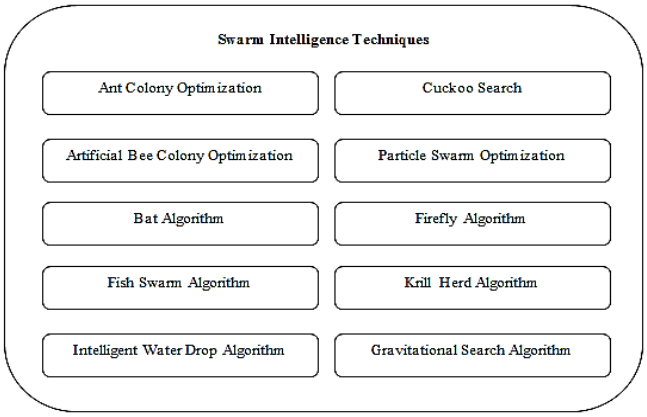

Figure 1: Swarm Intelligence Techniques

IV. SUSPICIOUS USER PROFILE (SUP) DETECTION SYSTEM

This section explains the working of the Suspicious User Profile (SUP) detection system. The proposed system not only finds out the suspicious activity but also the user profiles associated to those activities. SUP detection system works in four different modules listed as follows:

1. Text Dataset Module
2. Pre-processing Module
3. Feature extraction and selection Module
4. Ant Colony Optimization

Figure 2 presents the flow char representation of the SUP detection system in accordance to the working of the different modules.

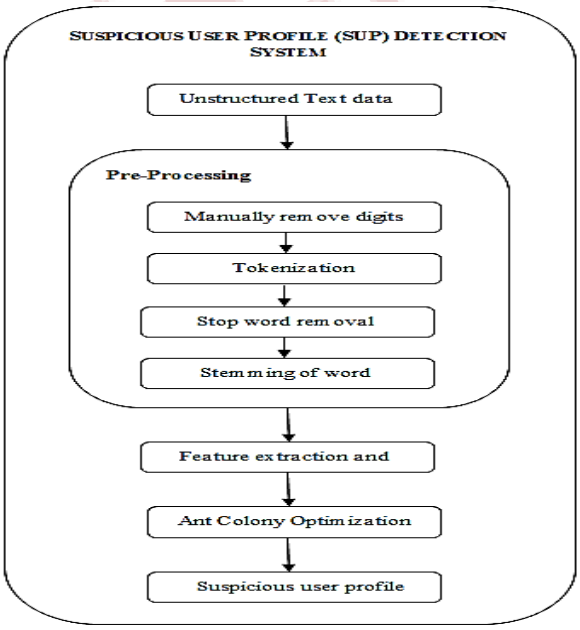

Figure 2: Flow chart of the SUP detection system

A. Text Dataset Module

This is the first module of SUP detection system. In this module, the text based dataset is collected from various social networking websites, chat forums, blogs and comments etc. the dataset contains the data in the form of comments, posts, messages etc. and the information about the associated user profile. Table 2 shows the sample dataset collected from various online resources. This dataset is raw in nature so it needs to be processed before further operation.

B. Pre-processing Module

In this module the unstructured text dataset is pre-processed for further processing. Initially, the numerical digits present in the text are removed manually if they are of no use. The data is tokenized and then the words having no linguistic meaning also known as stop words are removed for the dataset. The final process in this module is stemming. Stemming is performed to reduce the word to its base word.

C. Feature Extraction and Selection

The third module is the feature extraction and selection module. Features play a very important role in the process of classification. The statistical features like term frequency (TF) weighting, term frequency inverse document frequency (TF-IDF) are extracted and then further fed as input to the classification algorithm.

D. Ant Colony Optimization

The last module is the classification module. The algorithm used for the suspicious user profile identification is swarm intelligence based ant colony optimization (ACO) algorithm. The algorithm is based on the natural working phenomenon of ants. The working of the ACO algorithm is explained as below:

Algorithm

```
Initialise the parameters ()
x, y = 0
for y = 1 to colonies do
  Ant a0  // create sub-colony and release the agents
    While !(sub-colony termination condition) do
          x = x + 1
          manage_pheromone()
          manage_demon_action()
          determine_solution_quality()
          manage_selection_procedure()

      end while
      y = y + 1
      Sbest  // optimal solution candidate
      Update pheromone()
end for
```

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

Table2: Sample Text dataset

User_ID	User_Name	Text_Data
1.	Amit Verma	Just started a new blog Tricksmode.com for latest tricks... hope you all will enjoy article there,-)
2.	Carmen	He knew from experience the tormenting expectation of terror and death
3.	Jamie	Where were you last night? I kept waiting for the whole night.
4.	Robbie	You can never buy love..... but you still have to pay for it.
5.	Brad	People were killed by suicide bomber.
6.	Madison Malone	Harry Potter is 20 years old today! Hufflepuff still sucks!

E. Working of SUP Detection System

The working of the proposed SUP detection system is explained with the help of an example. The following example is randomly considered from the text dataset in order to give the detailed working description of the SUP detection system. The SUP detection system accepts sentence based text input and delivers output by detecting the suspicious word present in the text and also specifies the user profile associated as shown in table 3.

Table 3: Working of SUP Detection System

Input_Text	This is not good at all!!!I do not feel safe here!!!I am terrified to death....
Tokenization	'This' 'is' 'not' 'good' 'at' 'all' '!!!' 'I' ' do' 'not' 'feel' 'safe' 'here' '!!!' 'I' 'am' 'terrified' 'to' 'death' '....'
Stop word removal	'not' 'good'' do' 'not' 'feel' 'safe' 'here' 'terrified' 'death'
Stemming	'not' 'good'' do' 'not' 'feel' 'safe' 'here' 'terrify' 'death'
Suspicious word present or not?	Yes
User Profile	User_id = 138; User_profile = Gupta Neena

V. CONCLUSION

This paper discusses the design and development of Suspicious User Profile (SUP) detection system. The SUP detection system takes text based data in the form of chat messages, comments, posts, status updates etc. collected from various online available resources. The data is pre-processed and then the proposed SUP detection system accepts the input data and signifies the presence of the suspicious word. Suspicious word could be any word which signifies the presence of any terror, terrorism, fraud, cyber bullying, financial fraud, violence etc related activity. If the input text data contains the suspicious word then the associated user_id along with user_name is delivered as output. The working of the SUP detection system is shown with the help of an example. On the basis of the results evaluated which are shown in the example indicates that the proposed SUP detection system is efficient to detect the suspicious words and associated user profiles.

REFERENCES

[1] Murugesan, M. Suruthi, R. Pavitha Devi, S. Deepthi, V. Sri Lavanya, and Annie Princy. "Automated Monitoring Suspicious Discussions on Online Forums Using Data Mining Statistical Corpus Based Approach." Imperial Journal of Interdisciplinary Research 2, no. 5 (2016).

[2] Yen, Ting-Fang, Alina Oprea, Kaan Onarlioglu, Todd Leetham, William Robertson, Ari Juels, and Engin Kirda. "Beehive: Large-scale log analysis for detecting suspicious activity in enterprise networks." In Proceedings of the 29th Annual Computer Security Applications Conference, pp. 199-208. ACM, 2013.

[3] Khangura, M. Dhaliwal, M. Sehgal. "Identification of Suspicious Activities in Chat Logs using Support Vector Machine and Optimization with Genetic Algorithm." International Journal for Research in Applied Science & Engineering Technology (IJRASET), VII(5), 2017.

[4] Chiu, Carter, Justin Zhan, and Felix Zhan. "Uncovering Suspicious Activity From Partially Paired and Incomplete Multimodal Data." IEEE Access 5 (2017): 13689-13698.

[5] Patil, Yogesh V., Omkar A. Salvi, Paresh R. Waghmare, Akshay D. Kondilkar, and Vijayalaxmi P. Kadroli. "Abnormal Behaviour Detection on Live Streaming." (2017).

[6] M. Jiang, A. Beutel, P. Cui, B. Hooi, S. Yang, and C. Faloutsos, "Spotting Suspicious Behaviors in Multimodal Data: A General Metric and Algorithms," IEEE Trans. Knowl. Data Eng., vol. 28, no. 8, pp. 2187–2200, 2016.

[7] Alami, Salim, and OMAR EL BEQQALI. "DETECTING SUSPICIOUS PROFILES USING TEXT ANALYSIS WITHIN SOCIAL MEDIA." Journal of Theoretical & Applied Information Technology 73, no. 3 (2015).

[8] Tayal, Devendra Kumar, Arti Jain, Surbhi Arora, Surbhi Agarwal, Tushar Gupta, and Nikhil Tyagi. "Crime detection and criminal identification in India using data mining techniques." AI & society 30, no. 1 (2015): 117-127.

[9] GOWRI ,G.S.ANANDHA MALA ,G.DIVYA , " SUSPICIOUS DATA MINING FROM CHAT AND EMAIL DATA " , International Journal of Management and Applied Science (IJMAS) , pp. 75-79, Volume-2,Issue-2 (2014)

[10] Bonabeau, Eric, Marco Dorigo, and Guy Theraulaz. Swarm intelligence: from natural to artificial systems. No. 1. Oxford university press, 1999.

[11] Hinchey, Michael G., Roy Sterritt, and Chris Rouff. "Swarms and swarm intelligence." Computer 40, no. 4 (2007).

[12] Panigrahi, Bijaya Ketan, Yuhui Shi, and Meng-Hiot Lim, eds. Handbook of swarm intelligence: concepts, principles and applications. Vol. 8. Springer Science & Business Media, 2011.

[13] Dorigo, Marco. "Optimization, learning and natural algorithms." Ph. D. Thesis, Politecnico di Milano, Italy (1992).

[14] Maniezzo, A. C. M. D. V. "Distributed optimization by ant colonies." In Toward a practice of autonomous systems: proceedings of the First European Conference on Artificial Life, p. 134. Mit Press, 1992.

[15] Dorigo, Marco, Vittorio Maniezzo, and Alberto Colorni. "The ant system: An autocatalytic optimizing process." (1991).

[16] Maniezzo, Vittorio, and Antonella Carbonaro. "Ant colony optimization: an overview." In Essays and surveys in metaheuristics, pp. 469-492. Springer US, 2002.

[17] Pedemonte, Martín, Sergio Nesmachnow, and Héctor Cancela. "A survey on parallel ant colony optimization." Applied Soft Computing 11, no. 8 (2011): 5181-5197.