Genetic Algorithm Based Method for Identification of Cybercriminal Networks from Online Social Media

Neeraja Bhaskar[1], Revathy N[2]
[1]PG Scholar, Dept. of computer science and engineering
[2]Assistant Professor, Dept. of computer science and engineering
TKM Institute of Technology, Kollam, India

***Abstract*— Text mining originated from data mining and it is the knowledge discovery from textual data to uncover useful but hidden information. In the past two decades social community mining and social network analysis is an important research area, but little work is performed for the automated discovery and analysis of cybercriminal networks. This paper focused on the techniques to analyze the online social media messages to uncover the various cybercriminal relationships and visualization of the cybercriminal networks, which facilitate cybercrime forensics. This paper introduces a novel solution called GA based LDA, which uses Genetic Algorithms (GA) to determine a near optimal configuration for LDA in the context of cybercriminal network mining. The experimental results shows that the proposed method outperforms the existing Latent Dirichlet Allocation (LDA) based context sensitive Gibbs sampling algorithm.**

***Index Terms*—Genetic Algorithm, Latent Dirichlet Allocation, Cybercriminal Network, Relationship Inference, Seeding Relationship Indicators.**

I. INTRODUCTION

There was a rapid growth in the number of cybercrimes when compared to the previous years as the advancement of information technologies. The existing cyber security technologies such as Intrusion Prevention Systems and anti malware software are weak in cybercrime forensics and prediction. The existing cyber security solutions rely on low level network traffic features and software coding signatures to identify cybercrimes. Since the cybercriminals can constantly change their attack tactics, which leads difficulty in cybercriminal detection. So to combat the rapidly growing trend of cybercrimes here applying advanced computational methods to identify the underground cybercriminal networks.

More evidences have shown that cybercriminals tend to exchange cyber attack knowledge or even transact cyber attack tools through the dark markets established in online social media. Such a trend offers opportunities for cyber security analysts and researchers to tap into the online social media by analyzing the conversational messages to develop better insights about cybercrimes and communities of cybercriminals. So the effectiveness and efficiency of cybercrime forensics can be enhanced by means of automated cybercriminal network mining method.

The remainder of the paper is organized as follows: Section II provides an overview of the recent related works; Section III illustrates the proposed cybercriminal network mining methodology; Section IV describes implementation details; Section V describes the evaluation procedure and their results and Section VI offers the concluding remarks.

II. RELATED WORKS

The work closest to the proposed method is the application of a LDA based probabilistic generative model to mine latent topics describing two kinds of cybercriminal relationships (transactional and collaborative) based on their conversational messages posted to online social media [1]. Figure 1 illustrates the main steps of the existing cybercriminal network mining methodology. Context Sensitive Gibbs sampling algorithm is developed to implement the LDA topic model. The topic modeling module used to extract the relevant concepts (cybercriminal relationships). LDA [2] is a conceptual clustering method which automatically groups semantically related terms together to form cybercrime related concepts. The latent topic modeling results are randomly ordered, so Laplacian ranking (topic re-ranking method) constructed to identify two sets of aggregated concepts. Then the aggregated concepts used to infer the relationship label of a test message. The problem with this method is the identified terms in each clusters are not relevant. So the number of terms produced in each aggregated concepts are not sufficient for infer a hidden cybercriminal relationships in the test messages. The proposed method can discover more relevant cluster members and the generated cybercriminal network is more accurate. In Tsai and Chan [3], a probabilistic latent semantic analysis (pLSA) model used to mine latent topics describing cybercrimes from blog messages. It suffers from the problem of over-fitting and extraordinary computational cost of learning a large number of model parameters. The proposed method not only discovers latent topics about cybercrimes, but it also uncovers the cybercriminal networks from this topics. Lau, Xia and Li [4], proposed LDA based social media analytical model for cybercrime forensic. This model can only discover latent concepts related to cybercrimes, not any further processing are performed on that concepts.

Lau and Xia [5], proposed an approximation method of Gibbs sampling, a Markov chain Monte Carlo algorithm, has

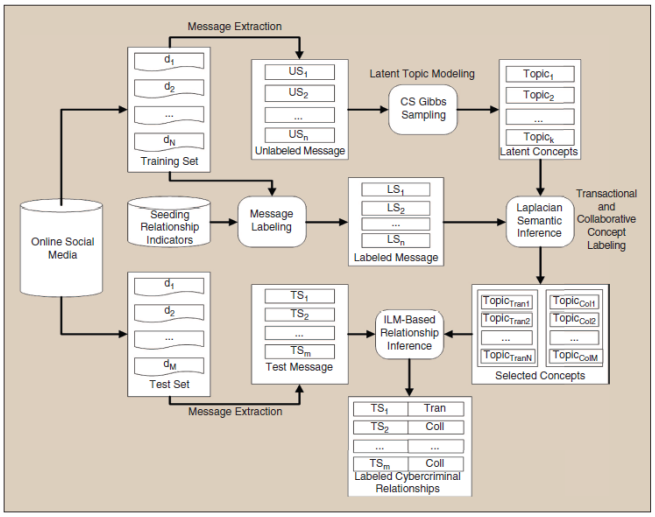

Figure1: Existing system architecture

been applied to estimate the conditional probabilities of LDA-based models. This method is also not sufficient for the mining of cybercriminal networks.

Dhanya, Jathavedan and Sreekumar [6], proposed two clustering methods for text clustering. spectral bisection and genetic algorithm (GA)is used for the clustering. Then it evaluates the clustering efficiency between GA and k-means algorithm. The proposed method effectively use GA based clustering for the mining cybercriminal networks. The method proposed by Grant and Cordy [7] estimate the optimal number of latent topics in source code. It generates series of Latent Dirichlet Allocation models with varying topic counts. The method proposed by Mustafa, Hajeer, Dipankar Dasgupta and Sugata Sanyal [8] genetic algorithm is used as a data mining method where the ultimate goal is to determine clusters of network communities in a given online social network data types including comments, emails, chat sessions, etc. and can form clusters according to one or more topics. Annibale Panichella, Bogdan Dit and Rocco Oliveto [9] introduce a new method LDA-GA which use GA to determine near optimal configurations for LDA in three different SE tasks: traceability link recovery, feature location, and software artifact labeling.

The main contributions in this paper are the development and evaluation of a novel weakly labeled cybercriminal network mining method based on standard LDA model and Genetic algorithm(GA). GA is commonly used for data mining applications such as clustering, classification and feature extraction. Here GA is used for cluster mined latent concepts to form aggregated concepts. These aggregated concepts are then used to infer a hidden cybercriminal relationship in the online social media messages. So the overall computational methods can uncover relationships among cybercriminals. The existing system can discover two types cybercriminal relationships i.e., transactional and collaborative. The proposed method discovers the networks showing transactional and collaborative relationships.

III. PROPOSED SYSTEM

The basic intuition behind the proposed method is to mine latent concepts describing specific types of cybercriminal relationships. The concepts are extracted using a generative model to bootstrap the performance of cybercriminal relationship identification. Figure 2 illustrates the main steps of the proposed methodology. Here messages are retrieved from an online forum are used for the processing. First, the conversational text messages collected from an online social media are preprocessed using stop word removal. The extracted messages are then fed into a GA based module to mine relevant cybercrime related concepts.

The latent topics from the messages are mined using GA based LDA method. Based on the pre-defined relationship indicators the mined concepts are classified into two relationship categories. The relationship indicators with their categories are stored in a database. The proposed method uses mainly two relationship categories i.e., transactional and collaborative. Two sets of aggregated concepts are formed. These concepts are stored for infer the hidden cybercriminal relationships from the arbitrary messages. Since this method requires small set of seeding indicators, it's a weakly supervised relationship mining method.

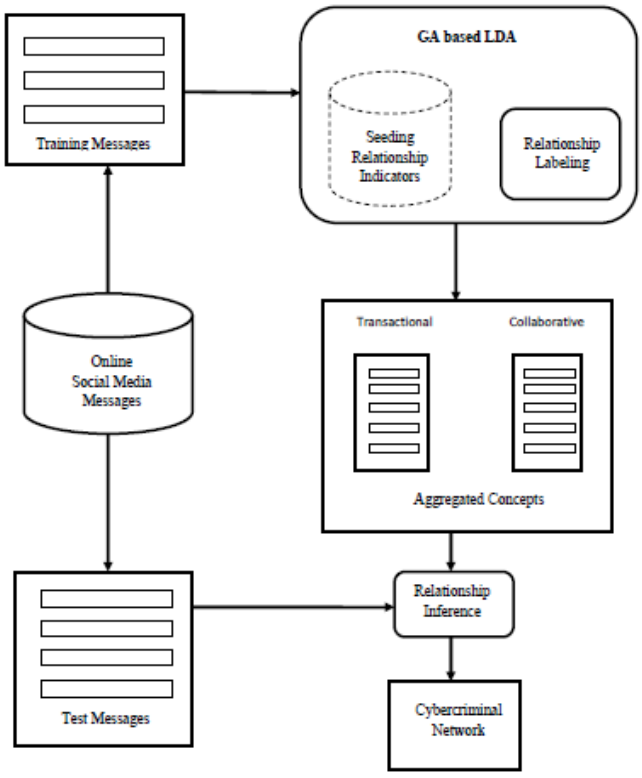

Figure 2: Proposed system architecture

A. Latent topic modeling

Some recent approaches to modeling document content are based upon the idea the probability distribution over words in a document can expressed as mixture of topics and each topic is a probability distribution over words [2]. Here we use one such model Latent Dirichlet Allocation which is an Information Retrieval (IR) model that allows fitting a generative probabilistic model from the term occurrences in

documents. LDA maps the documents from the space of terms into smaller space of topics. Then the latent topics can be clustered based on their shared topics.

LDA can be regarded as a topic based clustering method, which can be used to cluster documents in the topics space using the similarity between topic distributions. The best clustering results depends on the different LDA configurations. LDA requires a set of parameters as input. They are:

- k, which is the number of topics that the topic model extract from the data. This is same as the number of clusters in a clustering algorithm.
- n, which is the number of iterations.
- α, which affects the topic distributions per document.
- β, which influences the term's distribution per topic.

B. GA based LDA

In the proposed approach, vary all the parameters to find an optimal configuration for LDA by using GA. Different LDA configurations gives various clustering models of the documents. So all clustering models obtained by configuring LDA are not good.

Based on the assumption that the higher the clustering quality produced by LDA, higher the accuracy of LDA when used for cybercriminal network mining task, this paper present an approach to efficiently identify the best LDA configuration P= [α, β, k, n], that maximizes the overall quality of the clustering. For solving such an optimization problem this paper applies Genetic Algorithms, which is a stochastic search technique based on the mechanism of a natural selection and natural genetics. The introduction of GA by Holland [10] in the 1970s, this algorithm has been used in a wide range of applications where optimization is required and finding an exact.

Here the chromosomes (individuals or solutions) are represented as an array with four floats. Each elements represents denotes k, n, α and β. Thus a chromosome represents a particular LDA configuration and the population is represented by a set of different LDA configurations.

The chromosomes in the initial population are subjected basic GA operations such as selection, crossover and mutation of unique chromosomes. The fitness value is applied each chromosome for measuring the quality of different clustering models that are produced by various LDA configurations. The process is iterated and the chromosome with the highest fitness value is taken as the solution. The novel GA approach can be briefly summarized as:

1. Create chromosomes with four elements (LDA parameter values), by generating different LDA configurations
2. Build population using set of chromosomes.
3. Randomly select pair of chromosomes for the next phase.
4. Perform crossover by generating new population.
5. Perform mutation, which randomly changes one of the four LDA parameter values of an individual, with a different parameter value.
6. Replace repeating elements with missing elements in all chromosomes. This will create the new population
7. Remove the duplicating chromosomes.
8. Cluster the messages using unique chromosomes from the final population, using LDA. That extracts latent topics related to the cybercrime domain.
9. Evaluate the cluster quality based on the fitness value.
10. Remove chromosomes with zero fitness value.
11. Do steps 4 to 10 repeatedly.
12. Select the chromosome with the highest fitness value which is the optimized LDA configuration.

IV. IMPLEMENTATION

The work is implemented as web application. To evaluate the effectiveness of proposed cybercriminal network mining method, first we need to collect cybercrime related messages from online social media. The training messages are collected via an API called Topsy, which retrieve relevant tweets. For the dynamic test messages, here made use of an online forum. Any registered users can post messages to the forum and start a thread of communication. The proposed method focuses on two types of cybercriminal relationships namely transactional relationship and collaborative relationship. Transactional relationship refers to buying or selling cyber-attack tools between two parties, and a collaborative relationship denotes the sharing of information or tools between cybercriminals and it does not involve any money exchange between the communicating parties. The terms that refer to cybercriminal relationships are clustered according to the seeding relationship indicators. Some of them are shown in the Table 1.

A sample segment of the mined cybercriminal network generated by the proposed method is shown in the figure 3. The cybercriminal networks are plotted using graph sharp tool.

Table 1: Showing seeding relationship indicators

Transactional	Collaborative
Sell	Chat
Money	Join
Account	Learn
Transfer	Bomb
Buy	Help
Bank	Hack
Pay	Attack

Figure 3: A sample segment of the mined cybercriminal network

Each node represents a cybercriminal or cybercriminal group and the edge represents the information about their relationship strength. Two cybercriminal networks can be drawn for each transactional and collaborative relationship.

V. PERFORMANCE EVALUATION

The comparative study between the existing and proposed approaches reveals that LDA based GA able to identify LDA configurations that lead to better accuracy. Figure 4 show the ROC curve of the baseline and the experimental systems for transactional and collaborative relationship classification under the online forum corpus. The area under curve shows the probability of a classifier correctly identifies a true positive case, for a comparative evaluation. It shown that the GA based LDA experimental system performs better than the CSLDA baseline system.

VI. CONCLUSION

The novel LDA based GA method find the best configuration for LDA, which is able to produce good results for specific cybercriminal network mining tasks. The advantage of GA with respect to the other search algorithm is its intrinsic parallelism, i.e., having multiple solutions (individuals) evolving in parallel to explore different parts of the search space.

Semantic rich and relevant aggregated concepts, describing relationships are generated with less computational cost are obtained by using the proposed approach. The experimental results show that the novel GA based LDA outperforms the existing CSLDA based cybercriminal network mining method.

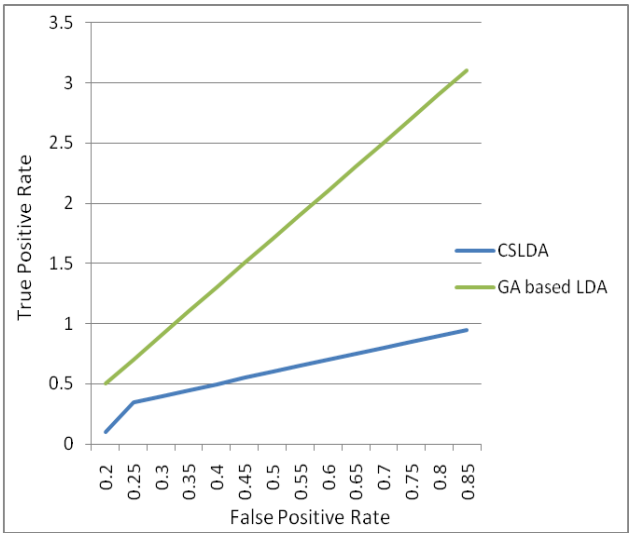

Figure 4: Performance evaluation graph

REFERENCES

[1] Raymond Y.K. Lau, Yunqing Xia and Yunming Y, "A Probabilistic Generative Model for Mining Cybercriminal Networksfrom Online Social Media",IEEEComputational intelligence magazine, February 2014.

[2] D. M Blei, A.Y.Ng, and M.I. Jordan, "Latent dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, Mar. 2003.

[3] F. S. Tsai and K. L. Chan, "Detecting cyber security threats in weblogs using probabilistic models," in Proc. Pacific Asia Workshop Intelligence Security Informatics (Lecture Notes in Computer Science), 2007, vol. 4430, pp. 46–57.

[4] Raymond Y.K. Lau, Yunqing Xia, and Chunping Li, " Social Media Analytics for Cyber Attack Forensic", International Journal of Research in Engineering and Technology (IJRET) Vol. 1, No. 4, 2012 ISSN 2277 – 4378.

[5] Raymond Y. K. Lau and Yunqing Xia, "Latent Text Mining for Cybercrime Forensics", International Journal of Future Computer and Communication, Vol. 2, No. 4, August 2013.

[6] P.M,Jathavedan M,Sreekumar A, "Implementation of Text clustering using Genetic Algorithm", International Journal of Computer Science and Information Technologies, Vol. 5 (5) , 2014, 6138-6142.

[7] S. Grant and J. R. Cordy, "Estimating the optimal number of latent concepts in source code analysis," in Proc. of the 10th International Working Conference on Source Code Analysis and Manipulation (SCAM'10), 2010, pp. 65–74.

[8] Mustafa H. Hajeer , Alka Singh , Dipankar Dasgupta and Sugata sanyal , Clustering online social network communities using genetic algorithms, 2013.

[9] Annibale Panichella, Bogdan Dit and Rocco Oliveto, "How to Effectively Use Topic Models for Software Engineering Tasks? An Approach Based on Genetic Algorithms", 2013.

[10] J. H. Holland, "Adaptation in Natural and Artificial Systems". University of Michigan Press, 1975.