International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 3, Issue 2, February 2016

An Optimization Technique for Mining Cybercriminal Network from Online Social Media

[1] Parvathy.G, [2]Bindhu J S
[1] PG scholar [2] Assistant Professor
Department of Computer Science and Engineering
College Of Engineering Perumon, Kerala,India
[1] theparvathygeetha@gmail.com[2] bindhuscholar@gmail.com

Abstract- **Data mining is the process of collecting data from different context and summarizes them into useful information. Data mining can be used to determine the relationship between internal factors and external factors .It allows the users to analyze, categorize and determines the relationships inferred in them. Text mining usually referred to as text data mining can be used be used to extract information from text. Text mining can be used in information retrieval, pattern recognition and data mining techniques. The introduction of social media and social networks has not only changed the opportunities available for us but also we need to be beware about the threats. Recent researches show that the number of crimes are increasing through online social media and they may cause tremendous loss to organizations. Existing cyber technologies are not effective to protect organizations .Existing mining methods concentrate on lexicons in which they can identify only a limited number of relations. Here a genetic algorithm approach is introduced in which latent concepts can be extracted. Genetic Algorithm is a linear search which requires only little information from large search area.. Then these concepts are subjected to extract the semantics which infers the corresponding relationships. Genetic algorithm provides a better solution in which accuracy and time efficiency can be improved. The main contribution of the paper shows that they identify the corresponding cybercriminal networks.**

Index Terms— Latent Dirichlet Allocation, Genetic Algorithm, Laplacian Score, Inferential Language Model

I. INTRODUCTION

The introduction of social media and social networks has not only changed the opportunities available for us but also we need to be beware about the threats. The information available within any sites are valuable to criminals so that they can use the individuals personal information to their advantage. According to the financial losses faced today there is a need for advanced computational intelligence approaches. Existing network mining mainly concentrate on constructed relationship lexicons [1] or manually defined lexico- syntactic patterns [2]. They can identify only a limited number of lexicons. There are increasing evidences showing that the criminals tend to exchange knowledge and transact or collaborative tools through online social media. On the other hand it offers possibility to obtain information about these criminals to create new methods and tools to obtain intelligence on cybercrime activities. Data mining techniques can be applied to assess information sharing and their classification. There is a collective goal of improving the state-of-the-art technology to provide a comprehensive approach to extract relevant information to provide criminal network analysis. Text mining consist of a range of techniques to analyze human languages using linguistic techniques. Text mining is usually the process of deriving information available in text along with the addition of some linguistic features. Concept level approaches can better grasp the implicit meaning associated with each text. Here the main contribution of this paper is the mining of cybercriminal network which can uncover both implicit and explicit meaning of each text based on their conversational messages posted on the online social media.

The concept based approaches are more promising than keyword based which relies on semantics rather than syntax. Concept based methods provides better performance than word based for task like topic modeling [3], opinion mining. The aim of this paper is efficient network mining through concept mining method which can extract more relevant concepts describing cybercriminal relationships

II. LITERATURE SURVEY

The information retrieval task is the retrieval of unstructured information. This information includes images, text etc. All documents are pre-defined and the retrieval system will retrieve documents in standard information retrieval task for satisfying user's needs. In most of the application, collection of documents may be large size, and this document collection needs to be mined. Nowadays lots of information are available in text and they need to be extracted. Text mining often[4] referred to as the process of structuring the input text and deriving patterns within structured data for evaluation and interpretation of output. They include information retrieval, pattern recognition, information extraction etc.

A. *Lexical Affinity*

This method not only identifies effected words, but it also assigns arbitrary words a probable affinity to particular emotions. This approach usually trains probability from linguistic corpora. It has better performance than keyword spotting., but this approach has two main problems: the first is, negated sentences and sentences with other meanings trick lexical affinity because they operate on the word level and second one is, lexical affinity probabilities are often biased towards of a particular genre, dictated by the linguistic corpora source. So it make difficult to develop a reusable, domain independent model.

B. *Keyword Spotting*

This method has increased accessibility and economy. It classifies text based on the presence of unuseful affect words like sad, happy, afraid and bored. However this method is weak in two areas that is it cannot reliably recognize affect negated words and it relies on surface features. Keyword spotting relies on the presence of affected words that are only surface features. Sometimes, a sentence conveys affect only through meaning rather than affect adjectives. Lexical affinity is slightly more sophisticated approach than keyword spotting.

C. *Topic Modelling*

It is way of text mining used to identify the patterns present in a document. Topic modelling can be used to find the abstract topics present in a collection of documents. Topic models are used to discover the hidden topic based patterns present in documents. For this several generative models were introduced.

D. *Generative model have three assumptions:*

- ❖ Each document should have a semantic structure
- ❖ Can infer topics from word document co-occurrence
- ❖ Words are related to topics and topics are related to documents

Generative models have a wide variety of applications in text mining, language processing and information retrieval. In the information research systems [5] apply Latent Semantic Analysis to identify intellectual cores in information systems. It improves the support vector machine and has certain limitations. This analysis shows how the individual ,groups and organizations interact with the IT and is not built on the basis of probabilistic background. It does not deal with words having different meanings. To overcome the matrix reduction problem and polysemy(words having different meanings), they explored a generative model Probabilistic latent semantic model (PLSA)

Here PLSA [6] model can be used for document clustering by employing link supervision between two documents. Here the link between two documents only indicates whether they should belong to the same cluster or not and no additional parameters are evolved here. Probabilistic latent semantic model (PLSA) uses a generative latent class model to perform the probabilities. Here only a qualitative evaluation is performed as only a limited number of concepts are extracted from the documents so that the model suffers from problems of over fitting and computational cost of learning large number of parameters is very high. There is no way to generalize new document or unseen documents.

These problems overcome by LDA(Latent Dirichlet Allocation) [7] as Latent Dirichlet Allocation is a probabilistic generative model in which relevant topics can be extracted from various documents. It is one of the most successful topic model where the probabilities of topics occurring in the document and probabilities of word occurring in the topic can be calculated .First it calculates the number of topics in the documents then calculates specific distribution of topics and then based on this document distribution ,topics are generated then the words for each topic are generated .Latent Dirichlet Allocation can model long length documents compared to another generative models.LDA is a model and the experiments are really good when compared to other know information retrieval techniques. It depends on the word occurrence and the meaning of the concepts whereas the common sense knowledge modeling [3] does not take into account the word co-occurrence and it may be not accurate for large documents

Here LDA is enhanced by Gibbs sampling algorithm [8]. Gibbs sampling is a Markov chain Monte Carlo algorithm that is used to obtain the sequence of probability of words where direct sampling becomes difficult .Gibbs sampling randomly assigns terms to topics. They can be used to approximate the joint distribution and marginal distribution of one or more variables or subset of variables. Here they provide a contextual knowledge extracted from domain specific corpus.

In cyber physical systems the social networks are mined using sentimental analysis [9]. Here Sentimental analysis is used which depends on the attitude of speaker or writer and classifies them using common sense knowledge into positive and negative categories. Topics related to each categories are identified and their contextual polarity is calculated. According to this the word with highest value is taken and their sentence score is calculated. Here we have studied how the cybercriminal activity effect the society but we need to develop a system that helps to secure the social media more efficiently. Social network analysis method uses the source and destination IP addresses of cyber attacks from social media to construct cyber-attack graphs but in our proposed approach it can tap into online social media and utilize the concepts to uncover the relationships.

Existing network mining methods use constructed relationship lexicons or lexicon- syntactic patterns as they can find only limited number of explicit relationships, because they use natural languages[10] that are flexible and unpredictable. Supervised machine learning methods is also a solution but it requires a lot of time and resources. Various computational intelligence methods like artificial neural networks, fuzzy systems, swarm intelligence for intrusion detection were examined [11]using low level network features. Apriori association rule mining method can be used to identify the genes and bound them together such that the genes that belong to the same category are grouped into one. Various natural language techniques has been developed. for describing the relationships between entities and domains and relationship among companies. The Co Miner System[12] was proposed to identify the relationships among companies. Here they use natural language techniques to find the relationships and domains among companies Here only a limited number of input data is used so that the recall value of such a system may be low. Another method for generating a dual probabilistic model for Latent Semantic Indexing[13] is done using Cosine Similarity .Cosine similarity can be used to find the similarity between documents and the similarity between topics present in the document. Cosine similarity is usually measured in vector form depends on the angle between them and not on the magnitude.

III. PROPOSED METHOD

The proposed method aims to distinguish the messages into criminal and non criminal by using genetic algorithm which builds up a dynamic topic model. Here more corpora can be extracted from social media by which the accuracy and time efficiency of the system can be improved. Genetic algorithm is a linear which requires little information from large search area. The provide a better solution from a set of candidate solutions. Genetic algorithm belongs to evolutionary algorithm which generates solutions to optimization problems using techniques initialization ,selection ,cross over ,mutation and best fit. The flowchart is depicted in figure 1

The working of genetic algorithm is as follow:

First a population is created from a group of individuals and then these individuals are evaluated. The evaluation is performed and each individual are given a fitness score based on which they are evaluated .Two individuals are selected based on their fitness score, higher the score greater the chance to be selected. This process continues until a best solution is obtained from a set of candidate solutions. Genetic algorithm takes the advantage of giving greater weight to individuals with best fitness score and concentrate the search in regions which leads to select the best topics. Genetic algorithm provides a heuristic search to solve optimization problems. Here Genetic algorithm provides a better solution in which more concepts can be extracted and time efficiency can be improved.

(1)Here first a graph is generated using each words in a document as the vertex of graph G. Assign weights for each edge depending on the frequency of times each word occurs.
(2)Check while the disconnected components of G are less than the number of topics given remove the edge with minimum weight such that they form n clusters
G={G1,G2,G3.....GN}.
(3)For each graph add components of graph G to cluster C

Here the clusters obtained from the above are given as input. In each clusters there are a set of k words.

In initialization randomly K words are selected from the clusters .In selection operator 2N solution are formed they give preference to better solutions allowing them to pass on their cluster to next sequence. The goodness of each concept is determined by its fitness score. Fitness score can be determined by the similarity of members in the cluster In cross over operation generates a sequence by combining any two best solutions from the previous generation. Mutation can be applied to randomly chosen cluster changing the weight to a new random number and changing the word with any other word in the corpus, and ensures a newly introduced word into the sequence

IV. FLOWCHART

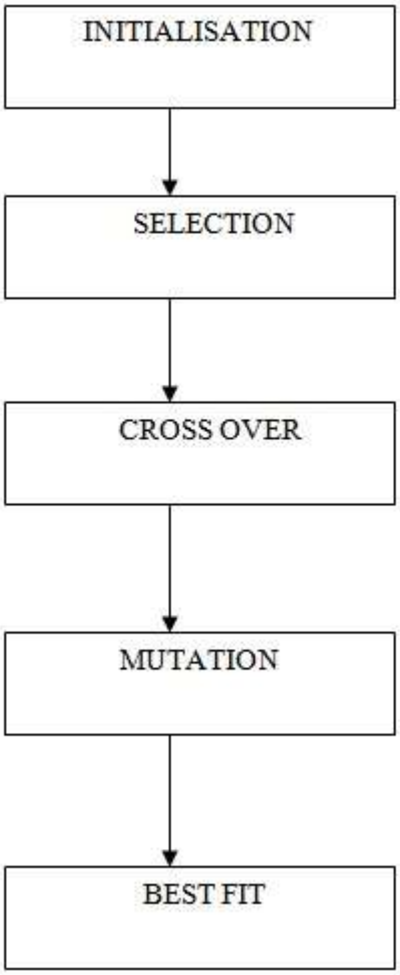

Figure 1: Flowchart Of Genetic

A. Algorithm

```
Randomly initialize members in the cluster
Determine the fitness of each solution
Repeat
        Select words from cluster C
        Perform cross over on words creating a new sequence N+i
        Perform mutation of sequence N+i
        Determine fitness score of sequence N+i
Until best topics are found
```

V. EXPERIMENTAL ANLAYSIS

The graph in figure 2 shows the comparisons between LDA method and Genetic algorithm.LDA is implemented using context sensitive Gibbs sampling algorithm. From the graph we concluded that the GA based method find more latent topics than LDA method. These helps to identify precious cybercriminal relationships .And from these relationships find efficient cybercriminal network.

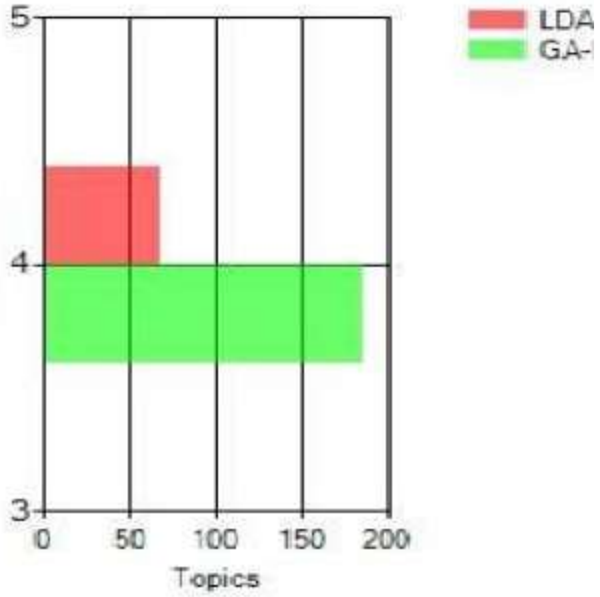

Figure 2: Comparison of LDA and Genetic Algorithm

VI. CONCLUSION

Latest security systems are weak in cybercrime detection so there is rapid growth in the number of crimes through online social media. The contribution of the paper is the development of a dynamic topic model using genetic algorithm. The laplacian score algorithm can effectively extract semantics of the corresponding concept describing into criminal and non criminal concepts. These concepts are then applied to inferential language model to infer the corresponding relationship. Genetic algorithm provides a better solution in which more concepts can be extracted by which time efficiency and accuracy can be improved. By mining the network security intelligence in social media not only facilitates the cyber attack but also has an intelligence to predict the cyber attack before they can be launched.

REFERENCES

[1] R.Xia, C.Zong, X.Hu and E.Cambria,Feature ensemble plus samples selection:A comprehensive approach to domain adaptation for sentiment classification ,IEEE Intell .Syst.,vol 28,no.3,pp.10-18,2013

[2] R.Li,S.Bao,J.Wang,Y.Yu and Y.Cao, Cominer: An effective algorithm for mining competitors from the web, Data Mining, in Proc.Int. Conf. Data Mining,2006,pp. 948-952

[3] D.Rajagopal, D.Olsher, E.Cambria and K. Kwok(2013): Commonsense topic modeling In Proc.ACM Int. Conf.Knowledge Discovery Data mining, Chicago

[4] Sangno Lee, Jeff Baker,Jaeki Song : An empirical comparison of four text mining methods Proceedings of the 43rd Hawaii International Conference on System Sciences 2010

[5] A. Sidorova, N. Evangelopoulos, J. Valacich and T. Ramakrishnan, Uncovering the intellectual core of the information systems discipline, MIS Quarterly, 32 (2008), pp. 467-482..

[6] Lingfeng Niu ,Yong Shi :Semi-Supervised PLSA for Document Clustering:2010 International Conference On Data Mining Workshops

[7] M Blie and M.I Jordan(2003): Latent Dirichlet Allocation.J.Mach.Learn Res,993-1022

[8] S. Geman and D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian relation of images, IEEE Trans. Pattern Anal. Mach. Intell., vol. 6, no. 6, pp. 721741, 1984

[9] Mining Social Network Data for Cyber Physical System: Manjushree Gokhale, Bhushan Barde, Ajinkya Bhuse, Sonali Kaklij: (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (2) , 2015, 1490-1492

[10] D.Maynard,V.Tablan and C.Ursu,(2001): Name enitity Recognition from diverse text types, In Proc,Conf.Recent Advances Natural Language processing.

[11] S.X .Wu and W.Banzhaf,The use of computational intelligence in intrusion detection systems:A review. Appl.Soft.Comput.,vol 10. No.1.pp.1-35,2010

[12] Y.Xia,W.Su,R.Y.K.Lau and Y.Lie,Discovery latent commercial networks from online financial news article, Enterprise inform.Syst.,vol 7,no.3,pp.303-331,2013

[13] Chris H.Q.Ding A Similarity Based Probability Model for Latent Semantic IndexingProc Of 22nd ACM SIGIR99 Conference, pp.59-65

[14] R. Y. K. Lau, D. Song, Y. Li, C. H. Cheung, and J. X. Hao, Towards a fuzzy domain ontology extraction method for adaptive e-learning, IEEE Trans. Knowl. Data Eng., vol. 21, no. 6, pp. 800813, 2009.

[15] Y. Song, S. Pan, S. Liu, M. X. Zhou, and W. Qian, Topic and keyword re-ranking for LDA-based topic modeling, in Proc. 18th ACM Conf. Information Knowledge Management, 2009, pp. 17571760.

[16] J.Y.Nie,G.Cao and J.Bai,Inferential language models for information retrieval,ACM Trans .Asian Lang.Inf.Process.,vol 5,no.4,pp.296-322,2006

[17] Dynamic Social Network Analysis of a DarkNetwork: Identifying Significant Facilitators, Siddharth Kaza, Daning Hu, and Hsinchun Chen, Fellow, IEEE

[18] Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering: Mikhail Belkin and Partha Niyogi