International Journal of Computer Applications (0975 - 8887)
Volume 134 - No.14, January 2016

A Probabilistic Generative Model for Mining Cybercriminal Network from Online Social Media: A Review

Parvathy G.
PG scholar
Department of Computer Science
College of Engineering Perumon,Kerala, India

Bindhu J.S.
Assistant Professor in CSE
Department of Computer Science and Engineering
College Of Engineering Perumon, Kerala, India

ABSTRACT

Social media has been increasingly utilized as an area of sharing and gathering of information. Data mining is the process of analyzing data from different context and summarizes them into useful information. It allows the users to analyze the data, categorize them and identifies the relationship inferred in them. Text mining often referred to as text data mining can be used to derive information from text. Text analysis can be used in information retrieval, information extraction, pattern recognition, frequency distribution and data mining techniques. An application of this is to scan a set of documents in natural language for predictive classification purposes. Recent researches shows that the number of crimes are increasing through social media that may cause tremendous loss to organizations. Existing security methods are weak in cyber crime forensics and predictions. The contribution of this paper is to mine cybercriminal network which can reveal both implicit and explicit meanings among cybercriminal based on their conversation messages.

Keywords

Latent Dirichlet Allocation(LDA), Laplacian Semantic Ranking, Inferential Language Model, Text Mining

1. 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. Existing cyber technologies are not effective to protect the organizations from various cybercrimes. 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 tends exchange knowledge and transact or collaborative tools through online social media. On the other hand it offers possibility to obtain informationabout 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.

Concept level approaches to natural language processing 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.

2. LITERATURE REVIEW

The information retrieval task is the retrieval of unstructured information. This information includes images, text. All documents are pre-defined and the retrieval system will retrieve documents in standard information retrieval task for satisfying users needs. In most of the application, collection of documents may be large size, and these document collections needs to be mined

2.1 Lexical Affinity

Here it not only identifies effected words, but also assigns arbitrary words a probable affinity to particular emotions. This approach trains probability from 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.

2.2 keyword Spotting

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

2.3 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 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.

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, 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.A basic PLSA model is illustrated in figure 1.Here D is the document index variable, C is a words topic drawn from documents topic distribution, W is a word drawn from the word distribution of this word topic.

These problems are overcomed by LDA [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 a 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 an intensively studied model and the experiments are really impressive 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

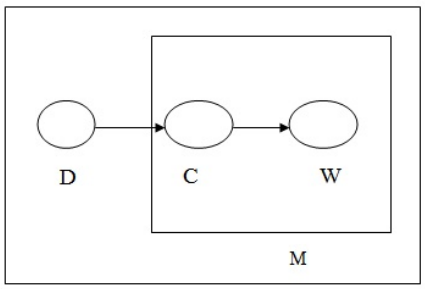

Fig. 1. Basic PLSA Model

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 provides a contextual knowledge extracted from domain specific corpus

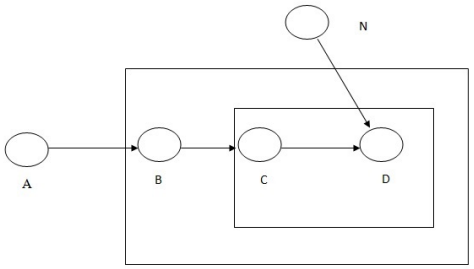

Fig. 2. Basic LDA Model

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.

The main difference between the proposed language model and other existing language model is that it uses context sensitive text mining and we use summation instead of multiplication which can be used to combine the probabilities of terms. Then a classification based ranking method can be done in which the frequency values are determined which by normalization can determine the final relationship scores.

3. COMPARISON OF DIFFERENT TEXT MINING METHODS

Comparison of different generative models are depicted on Table 1.

Table 1. Comparison Of Different Text Mining Methods

Method	Characteristics	Limitations
Latent Semantic Analysis	• Dimensionality is reduced using singular value decomposition • Captures semantics of words • Straightforward statistical background	• Difficult to determine the number of topics • Difficult to label a topic
Probabilistic Latent Semantic Analysis	• It models each topic in a document as a sample Each word is generated from a single topic • Each document generates a mixture of topics • Reduces dimensionality to topic level	It does not develop a probabilistic model

4. RECOMMENDED APPROACH FOR MINING CYBER CRIMINAL NETWORK

Here different generative models were discussed for text mining and all these models can be effectively used for text mining for different applications but they do not built a probabilistic model considering the distribution of words. Latent Dirichlet Allocation enhanced with Gibbs Sampling technique generates a probabilistic model which considers the sequence of probability of words and contextual information of the related corpus.

The main generative process in LDA include

—Choose N: A document is sequence of N words

—Choose Θ where theta is the multinomial distribution.

—For each of the N words perform the following steps.

1. Choose a topic
2. Choose a word from multinomial probability conditioned on the topic selected.

This model would be extremely useful in many critical applications like mining cybercriminal network. A laplacian semantic inference[14] method can be incorporated in this model to infer the semantics of mined concepts. This approach can be efficiently used for mining cybercriminal networks to classify online messages to criminal and non criminal and to infer a particular relation. This approach can be further enhanced by incorporating genetic algorithm to classify criminal messages . Genetic algorithm uses an optimization method in which a better solution can be obtained from a set of candidate solutions.

5. EXPERIMENTAL ANALYSIS

The topic solutions to LDA shown in Table 2 .Here the solution to LDA represents the components of camera and users.Here it shows the topics present in documents and the frequencies for each word.If topic 1 is about flash ,topic 2 is about picture then the topic selection is done by selecting the principal words belonging to each topic according to their probabilities LDA also provide better modelling of documents. Each words are loaded into appropriate topics such as battery and pictures. LDA allows long length documents and facilitates labeling of topics

Table 2. The topics solution of LDA

T0	T1	T2
camera(0.0455)	camera(0.0566)	camera(0.0640)
great(0.0243)	flash(0.0152)	batteries(0.0244)
pictures(0.0235)	good(0.0106)	good(0.0167)
mode(0.0112)	battery(0.0114)	price(0.0185)
canon(0.0120)	just(0.0132)	zoom(0.0078)
picture(0.0099)	canon(0.0083)	batteries(0.0112)
pictures(0.0100)	flash(0.0118)	picture(0.0085)

6. CONCLUSION

This paper analysis various generative models and network mining techniques that can be used to uncover the cybercriminal network. The ability to mine social media to extract relevant information is a crucial task. Since a great proportion of information contained in social media are in unstructured form, there is a need state-of-the art tool to collect and apply intelligence methods .Existing cyber technologies are not effective and they are weak in cybercrime forensics. Here a novel context sensitive text mining method is recommended by which latent concepts are extracted and these latent concepts are subjected to extract the semantics which describes the cybercriminal relationships. This system can be enhanced by genetic algorithm. Genetic Algorithm is a robust search method requiring little information to search effectively in a large or poorly-understood search space.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. 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.

7. 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,IEEEIntell.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