A Network Identification Model on Open Social Network

Amina Beevi S

M. Tech student, Department of Computer Science and Engineering
Al-Azhar College of Engineering and Technology, Perumpallichira, Idukki, Kerala, India

Abstract: ***Today the number of cybercrimes increases very rapidly, this cause large financial loss to many organizations. Existing cyber-security technologies are not effective enough to protect organizations from this cyber attack; this is because existing cyber security solutions are weak in cybercrime forensic and predictions. This paper present development of a novel weakly supervised cybercriminal network mining method to facilitate cybercrime forensic to mine cybercriminal networks from online social media. This method collects the conversational messages posted to online social media by cybercriminals and finds both implicit and explicit relationship among them.***

Keywords: LDA, PLSA, Laplacian score, Cyber attacks, lexicon

1. Introduction

Conventional text mining techniques [3] are inappropriate on finding latent concepts from big data generated on social networks. Need of such concepts place a major role in identifying cyber criminal networks. The classification on the basis of transactional, collaborative nature; has not been applied in cyber world. So here the main aim is efficient network generation through concept mining.

The proposed computational algorithm can effectively extract semantically rich representations of latent concepts describing transactional and collaborative relationships among cybercriminals based on publicly accessible messages posted to online social media. These latent concepts are then applied to bootstrap the performance of inferential language modeling-based relationship classification in texts, inferring the hidden cybercriminal relationships from the text messages and generate corresponding cybercriminal networks and identify the peoples or cybercriminals involved in this network.

2. Motivation and Related Works

The existing network mining methods use lexicons or lexico-syntactic patterns to detect the implicit relationships of entities from free texts. Pre-defined lexicons or lexico-synntatic patterns can only find a limited amount of explicit relationship because they use natural languages [1] and this is very flexible and ambiguous. Supervised machine learning is also a solution for cybercriminal network mining. However such methods require a lot of time and resources to build dataset for effectively train machine learning classifier and also difficult to label messages. Probabilistic latent semantic analysis (PLSA) model [4] is another method to find latent concept among cybercriminals from their messages in online social media, which analyze the blog message and mine latent topics describing cyber crime. However this model suffers from over-fitting and high computational costs of learning large number of parameters. Previously described methods have many disadvantages, so in proposed network identification model LDA-based [2] probabilistic generative model is used to extracts the latent concept describing various cyber criminal relationships, which is used to identify criminal network.

3. Implementation

Designing of probabilistic generative model to extract latent concept describing specific type of cybercriminal relationships for the purpose of bootstrapping the performance of cybercriminal relationship identification is the intuition behind cybercriminal discovery method. A probabilistic generative model for collaborative cybercriminal network mining is illustrated in figure.

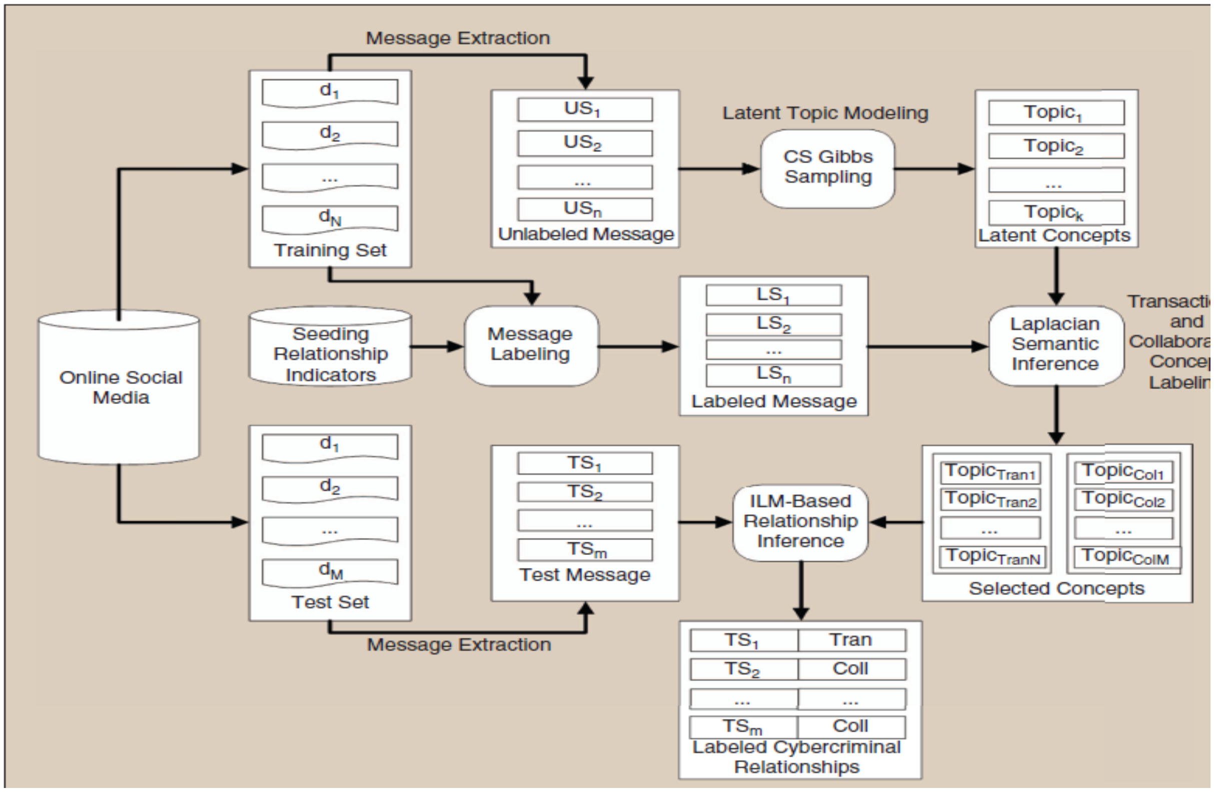

Figure 1: Probabilistic generative model for cybercriminal network mining

Here, collection of documents denoted by d_1 to d_N is taken from online social media. From the collection of messages, generate conversational messages US_i that refer to at least two users are extracted from a collection of unlabelled documents. These extracted unlabelled message is given to context sensitive Gibbs Sampling method for Latent topic modeling and generate latent concepts.

The latent concept discovered by LDA-based[2] latent semantic mining are unlabeled, then the classical LDA method is extended to infer the semantic label such as collaborative or transactional, of a mined concept. For that Laplacian scoring is applied, which is mainly for feature selection, to latent topic selection. If a candidate latent concept is semantically close to some messages characterized by a specific semantic label then the latent concept has a semantic label.

Labeled message LS_i was extracted by applying some generic seeding relationship indicators to an unlabeled corpus *D*.Here we mainly focus on collaborative and transactional relationship, there for two different Laplacian ranking are constructed. Which identify two set of latent concepts with the respective semantics. Based on the outputs from the Laplacian scoring method we select the high ranked and consistent latent concepts to represent the specific type of cybercriminal relationship. For each of these relationship type, aggregated concept is constructed which is used for inferring the relationship label of an arbitrary message.

After Laplacian based latent concept labeling, the aggregated transactional concepts and collaborative concepts are applied to infer the relationship label of an arbitrary message that refer to at least two cybercriminals. If the message content *d* is more likely to generate the transactional concept, the message will be considered to describe transactional cybercriminal relationship. If the message content *d* is more likely to generate the collaborative concept, the message will be considered to describe collaborative cybercriminal relationship. A novel inferential language modeling method is used to estimate the probability of *d* generating a specific cyber criminal relationship label. If the mined concept is incomplete then each term should be smoothed with respect to entire cybercrime message corpus *D* by means of maximum like hood collection language model.

A novel inferential language model consisting maximum like hood estimation of the term with respect to *d*, and context-sensitive text mining based smoothing for effective performance. The maximum like hood document language model is defined based on the frequency of term appearing in the message *d*. The set of term association or information flows are mined by applying the context sensitive text mining method to the unlabeled cybercrime corpus *D*. The strength of term association derived based on the context-sensitive text mining method.

The main differences between proposed language model and other existing inference-based language models are that we use context-sensitive text mining to extract term associations and use a summation instead of multiplication operator to combine the probabilities of deduced term to smoothen the maximum likehood document language model.

The proposed inferential language model can perform a ranking-based classification of the relationship labels of message. Here the messages are ranked according to their probabilities of having a specific relationship label. After relationship classification, a frequency count for a specific type of relationship is developed for each pair of cybercriminals. These frequency values are then subject to liner normalization to develop the final relationship scores for all the pairs [6].

If a pair has both a transactional and collaborative relationship at the same time, the system simply assigns the more specific transactional relationship to the pair. Lastly a cybercriminal network is composed based on the identified relationships among all the valid pairs pertaining to a specific period. Ones the cybercriminal relationship between each pair of users is identified, a cybercriminal network [5] are generated for a given period. From this network we can easily find cybercriminals involved in a particular criminal activities.

If we give a proper set of seeding relationship indicators then it is enough to discovering any type of cybercrime related relationship. Here we use a small set of seeding relationship indicators as input that do not need expensive manual labeling of a large number of training messages.

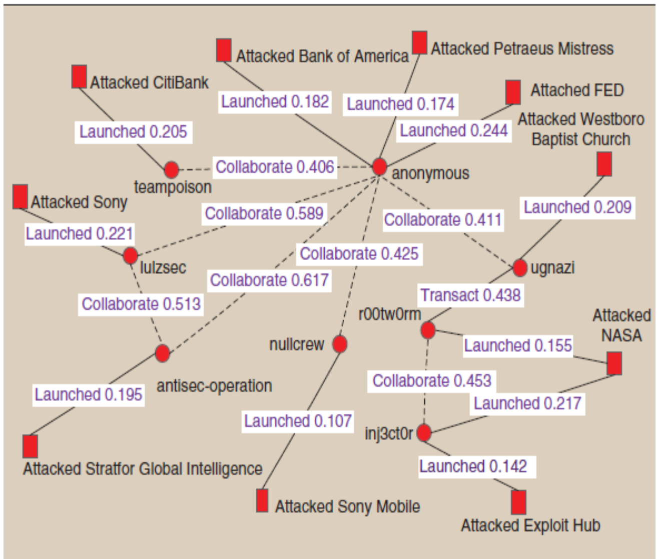

Figure 2: An example of mined cybercriminal network

4. Conclusion

In this paper, I introduce a novel weakly supervised cybercriminal network mining method to facilitate cybercrime forensics to mine cybercriminal networks from online social media. This is one of the emerging network mining methods to identify cybercriminal relationships and also protect organizations from various cybercriminal attacks. Today there is a rapid growth of number of cybercrimes, the existing technologies are not effective enough to protect from these cyber attacks, but the proposed method significantly outperforms than the existing techniques.

References

[1] D. Maynard, V. Tablan and C. Ursu, (2001): Named entity recognition from diverse text types. In Proc. Conf. Recent Advances Natural Language Processing.

[2] D. M. Blei and M. I. Jordan (2003): Latent dirichlet allocation. J. Mach. Learn. Res, 993–1022.

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

[4] D. Ramage, C. D. Manning, and S. T. Dumais (2011): Partially labeled topic models for interpretable text mining.In Proc. 17th ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining, San Diego, CA, 457– 465.

[5] H. Du and S. J. Yang (2011): Discovering collaborative cyber attack patterns using social network analysis. In Proc. 4th Int. Conf. Social Computing, Behavioral-Cultural Modeling Prediction, 129–136.

[6] P. J. Denning and D. E. Denning (2010): “The profession of IT: Discussing cyber attack,” *Commun ACM*, vol. 53, no. 9, pp. 29–31

Author Profile

Amina Beevi S received the B.Tech degrees in Computer Science and Engineering from Kerala University at MES Institute of Technology and Management in 2013. And now she is pursuing her M.Tech degree in Computer Science and Engineering under MG University, Kerala in Al-Azhar College of Engineering and Technology.