Advanced Science and Technology Letters
Vol.122 (ASP 2016), pp.236-239
http://dx.doi.org/10.14257/astl.2016.122.45

Analysis of Enterprise Email Big Data to Detect Careless Email Activities That May Cause Security Problems

Sung-min Kim[1] and Young-guk Ha[1,1]

[1] Konkuk University, Seoul, Republic of Korea
allmax100@naver.com, ygha@konkuk.ac.kr

Abstract. In modern business, email has become the most commonly used means of communication whose popularity is attributed to its simplicity of usage and low cost. However, there have occurred a lot of cases where a business's security got in trouble by a worker's careless email use. This paper suggests a method to help detect such problematic use of email by analyzing email data. The method is designed to find email messages that do not seem to have asked for a reply but were replied by someone, which we suggest can be a clue to the writer's carelessness, and involves document-vectorization through basic bag-of-words model and word2vec technique, which is a state-of-the-art method to create document vectors out of text documents. Enron email dataset was used as input data for an experiment which shows overall classification results. The results show which email messages are to be watched first to find careless email messages.

Keywords: email, big data, careless activity detection, security, word2vec, doc2vec.

1 Introduction

Recently, many businesses or organizations have spared no effort in developing ways to brace themselves for security threats posed by their insiders [1-3]. Such threats can stem from someone's purpose of inflicting damage on a business, but in some cases, they can also be mistakes out of someone's carelessness, which may lead to serious trouble and catastrophic damages to the business. In general, workers' mistakes are considered to come from their carelessness. If an employer can recognize which employees are likely to be careless at present and to cause trouble, then the employer is able to take measures to handle the dangers and prevent further damages by notifying them about their dangerous status. In this research, we suggest email messages are the best resource to watch since it is the media that workers always use at the workplace and contain information about their work as well as mental condition. Some of email activities of employees can be a useful clue to their

[1] Corresponding author.

ISSN: 2287-1233 ASTL

carelessness. One of email activities that we decided can be a useful clue is sending back a reply to an email message whose sender didn't want to get a reply, which we suggest can reflect his or her carelessness.

2 Enron Email Dataset

Enron email dataset is one of the favorite resources for researchers trying to improve existing email tools, or analyzing email usage. [4] It is from an actual company, Enron, which was an American energy, commodities, and services company based in Houston. [5] The email dataset includes email messages from 168 users who worked in senior management of Enron. It has 619,446 messages, and they are organized into folders. [6] But some messages are shared in multiple folders, and some are computer generated. After getting rid of the duplicated messages, we got around 192,798 distinct messages. Each email message contains several meta properties including information about its writer and sender, subject, date, etc.

3 Finding Careless Email Messages

We suggest replying to an email message that did not ask for any reply is an important clue to the writer's carelessness. Email messages can be divided into two types: messages whose sender expected to get a reply from the receiver, and messages whose sender did not expect any reply. This chapter explains an experiment where we tried to classify messages into the two types, which is the first step to find messages that did not ask for any reply message.

3.1 Distinguishing Email Messages with Replies

For every email messages whose subject does not start with 'RE:' or 'FW:'–called a regular message–, we tried to find another message whose subject starts with 'RE:'– we call it a reply message– and, if 'RE:' is removed, is the same as the regular message's subject. Besides considering the subjects of messages, we also took other factors into account, such as email addresses, and the time at which those messages were sent. Following these steps, we got 6,772 regular messages each of which has at least 1 reply message paring with the regular message.

3.2 Building Feature Vectors for Email Messages and Email Types

After the paring process, we recognized 6,772 messages with at least one reply message and other 186,018 messages without any reply. First, we created document vectors for each email message according to basic bag-of-words model [7] and the Doc2Vec technique [8]. And we summed vectors of messages that have reply messages to build a feature vector of 'Type R', which indicates messages having

replies, and other messages' vectors to build a feature vector of 'Type N', which indicates messages not having replies.

3.3 Classifying Email Messages

Using the feature vectors of the two types, we tried to classify all the 192,790 messages into the two types. The followings are the steps to classify the messages.

1. Make a document vector for each email message using Doc2Vec.
2. Make a document vector for each email type by summing all the emails of the type.
3. Compare each document's vector with each feature vector to decide the similarity between them.
4. The type of a document is decided according to which type's vector is more similar to the document. If its similarity to type R is bigger, then it is classified into 'Class R', otherwise it is classified into 'Class N'.

4 Experiment Result

With the introduced steps, we got 71,487 messages classified into Class R and other 121,303 messages classified into Class N. Among Class N messages, 6,772 messages were Type R. These messages can be considered to be the ones that are more likely to have not asked for any reply than the other Type R messages. Table 1 shows the overall experiment statistics.

Table 1. Overall experiment statistics

Message Type	Number of Message
All Messages	192,790
Type R Messages	6,772
Type N Messages	186,018
Messages Classified into Class R	71,487
Type R Messages in Class R	4,270
Type N Messages in Class R	67,217
Messages Classified into Class N	121,303
Type R Messages in Class N	**2,502**
Type N Messages in Class N	118,801

5 Conclusion

In this paper, we suggest that email messages exchanged among workers are a useful clue to the workers' carelessness, which could lead to unintentional insider threats. For the experiment, we used Enron email dataset that includes around 150 users who worked with Enron, and about 500,000 messages. Getting rid of some computer

generated files, we got around 190,000 distinct email messages. With the pre-processed dataset, we found 6,772 messages that received at least one reply message. They were used to build a feature vector of Type R which means a type of messages that received a reply. Other messages were used to build a feature vector of Type N that means a type of messages that did not receive a reply. Using the two feature vectors, we classified all the email messages into Type R or Type N. The classification results show that among all the messages without replies, 2,502 messages are more likely to have not asked for any reply than the other messages.

Acknowledgement

This work was supported by the ICT R&D program of MSIP/IITP. [B0101-15-0559, Developing On-line Open Platform to Provide Local-business Strategy Analysis and User-targeting Visual Advertisement Materials for Micro-enterprise Managers]

References

1. Johnson, N. B.: DHS cyber effort shifts to insider threats, 2013. http://www.federaltimes.com/article/20131216/DHS/312160014/DHS-cyber-effort-shifts-insider-threats.
2. Lombardo, T.: SECNAV launches plan to battle 'insider threats', 2013. http://www.navytimes.com/article/20130907/NEWS/309070005/SECNAV-launches-plan-battle-insider-threats-.
3. Vijayan, J.: DARPA launches insider threat detection effort for military, 2010. http://www.computerworld.com/article/2515328/security0/darpa-launches-insider-threat-etection-effort-for-military.html.
4. Cohen, W. W.: Enron Email Dataset, 2009. https://www.cs.cmu.edu/~./enron/.
5. Wikipedia. Enron Email Dataset, 2014. http://en.wikipedia.org/wiki/Enron.
6. Bryan Klimt, Y. Y.: Introducing the Enron Corpus, 2004. http://ceas.cc/2004/168.pdf.
7. jemdoc. Bag-of-words representation of text. https://inst.eecs.berkeley.edu/~ee127a/book/login/exa_vecs_bag_of_words_rep.html.
8. kifi, "From word2vec to doc2vec: an approach driven by Chinese restaurant process", http://eng.kifi.com/from-word2vec-to-doc2vec-an-approach-driven-by-chinese-restaurant-process/

