International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

An Data Analytics Applied for Crime Identification over IOT

Pulim. Pradeep Reddy[1], D. Rajesh[2]

[1, 2]Computer Science Engineering, RISE Krishna Sai Gandhi Group of Institutions, Ongole, India

Abstract: ***Data mining is the quest for knowledge in databases to uncover previously unimagined relationships in the data. This paper proposes to apply Decision tree in suspected e-mail detection (e-mails about criminal activities). Deception theory suggests that deceptive writing is characterized by reduced frequency of first person pronouns and exclusive words and elevated frequency of negative emotion words and action verbs. We applied this model of deception to the set of e-mail dataset, then applied ID3 algorithm to generate the decision tree. The decision tree that is generated is used to test the e-mail as suspicious or not. In particular, we are interested in detecting fraudulent and possibly criminal activities from such data.***

Keywords: data mining, deceptive theory, decision tree based classification

1. Introduction

Data mining has recently attracted considerable attention from database practitioners and researchers because of its applicability in many areas such as decision support, market strategy, financial forecasts, etc. Combining techniques from the fields like Statistics, Machine learning, Databases, etc. Data mining helps in extracting useful and invaluable information from database. Detecting unusual communication patterns in various means and channels of communications represents an important class of application directly relevant to security informatics [2].

E-mail has become one of today‘s standard means of communication. The large percentage of the total traffic over the internet is the e-mail. E-mail data is also growing rapidly, creating needs for automated analysis. So, to detect crime, a spectrum of techniques should be applied to discover and identify patterns and make predictions. Data mining has emerged to address problems of understanding ever-growing volumes of information for structured data, finding patterns within data that are used to develop useful knowledge.

As individuals increase their usage of electronic communication, there has been research into detecting deception in these new forms of communication. Models of deception assume that deception leaves a footprint. The work done by various researchers suggests that deceptive writing is characterized by reduced frequency of first-person pronouns and exclusive words and elevated frequency of negative emotion words and action verbs[8]. We apply this model of deception to the set of e-mail dataset and preprocess the e-mail body and to train the system we used ID3 (Iterative Dichotomiser 3) algorithm [6] to generate a decision tree that categorize the e-mail as deceptive or not. Text classification including e-mail classification presents challenges because of large and various number of features in the data set and large number of documents. Applicability in these datasets with existing classification techniques was limited because the large number of features makes most documents undistinguishable. In many document datasets, only a small percentage of the total features may be useful in classifying documents, and using all the features may adversely affect performance. The quality of training dataset decides the performance of both the text classification algorithms and feature selection algorithms. An ideal training document dataset for each particular category will include all the important terms and their possible distribution in the category. To our knowledge, this is the first attempt to apply Decision tree to task of Suspicious e-mail detection (e-mails about criminal activities).

1.1. Motivation

Concern about national security has increased significantly since the terrorist attacks on 11 September 2001. The CIA, FBI and other federal agencies are actively collecting domestic and foreign intelligence to prevent future attacks. These efforts have in turn motivated us to collect the data and undertake this paper work as a challenge. Data mining is a powerful tool that enables criminal investigators who may lack extensive training as data analyst to explore large databases quickly and efficiently. Computers can process thousands of instructions in seconds, saving precious time. In addition, installing and running software often costs less than hiring and training personnel. Also, computers are less prone to errors than human investigators. So this system helps and supports the investigators.

1.2. Organization of the Paper

The paper is organized as follows: Section 2 defines problem statement and related work in this area. Section 3 describes the proposed work and experimental results are presented in Section 4. Section 5 discusses performance measure. Finally, Section 6 concludes the paper and points out some potential future work.

2. Problem Statement and Related Work

It‘s hard to remember what our lives were like without e-mail. Ranking up there with the web as one of the most useful features of the Internet, billions of messages are sent each year. Though e-mail was originally developed for sending simple text messages, it has become more robust in the last few years. So, it is one possible source of data from which potential problem can be detected. Thus the problem

is to find a system that identifies the deception in communication through e-mails.

One of the earlier automated deceptive detection systems, constructed from a record linkage method based on string comparators [5], was proposed by Gang Wang, Hsinchun Chen and Homa Atabakhsh. This method has a restriction that it often requires intensive computation. Xindong Wu and Xing Xingquanzhu developed impact sensitive instance ranking method [12] to identify deception for real world data sets. This method has a restriction that the switching of attribute Ai and class C for attribute prediction APi, the accuracy of APi could be very low. P. S. Kaila and Skillicon developed a method based on the singular value decomposition [8] to detect unusual and deceptive communication in e-mails. The problem with this approach is that it does not deal with incomplete data in an efficient and elegant way and can not incorporate new data incrementally without having to reprocess the entire matrix.

Classification is an important data mining problem. The input is a dataset of training records (also called training database), wherein each record has several attributes. Attribute with numerical domains are numerical attributes and attributes whose domains are non non-numerical are categorical attributes. There is also a distinguished attribute called the class label. This classification aims at building a concise model that can be used to predict the class label of future, unlabeled records. Many classification models including Naive Bayes, Decision tree, Support vector machine, and Neural networks have been proposed.

[13] compared a cross-experiment between 14 classification methods, including Decision tree, Naive Bayesian, Neural networks, Linear square fit, Rocchio. KNN is one of the top performers, and it performs well in scaling up to very large and noisy classification problems. [9] showed a good performance reducing the classification error by discovering temporal relations in an e-mail sequence in the form of temporal sequence patterns and embedding the discovered information into content-based learning methods. Approach to Anomalous e-mail detection is considered. [15] showed approaches to detect anomalous e-mail and involved the deployment of data mining techniques. [4] proposed a model based on the Neural network to classify personal e-mails and the use of principal component analysis as a preprocessor of NN to reduce the data in terms of both dimensionality as well as size. [11] and [14] developed an algorithm to reduce the feature space without sacrificing remarkable classification accuracy, but the effectiveness was based on the quality of the training dataset. In the classification experiment for spam filtering, Decision tree showed better result than NB, NN, or SVM classifier [10].

3. Proposed Work

In this paper, we present a novel data mining based decision tree algorithm to detect e-mail concerning criminal activities. It is developed specifically for detecting deceptive communication in e-mail. The architecture of the proposed system is as follows.

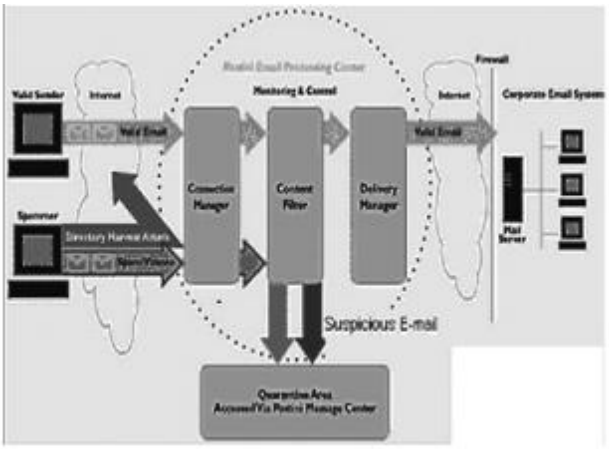

Figure 1: Proposed Suspicious e-mail detection system.

The architecture shown above is used to detect the suspicious e-mails. Connection manager is used to give the connection between the e-mail sender and the Processing center. The Content filter plays the important role i.e., it uses the preprocessing and classifying algorithm such as Decision tree, etc. to separate the suspicious e-mails. This output is delivered to the investigator with the help of Delivery manager.

The proposed method is implemented in JDK1.5 because Java is a high performance language for technical computing. In implementation, there are three parts: E-mail preprocessing, Building decision tree and Validation.

3.1 Database Used in Experiment

The Microsoft Access database is used to store the e-mail messages. The dataset contains the folder information for each of the suspicious and normal e-mail. Each message present in the folders contains the sender and receiver e-mail address, date and time, subject, body, text and some other e-mail specific technical details. We created MS Access database for the dataset to store the e-mail message, our database contains two tables. The first table contains the information of the e-mail message the sender, subject, text and other information. The second table contains the recipient‘s information. It contains the e-mail address of the recipient and the type (To, Cc, Bcc) in which message was sent to the recipient.

3.2 Text Classification Architecture

In Figure 2 we present a simple architecture of text classification systems. There is a pool of documents which represents the content at hand that can either be stored on disk, or could come from data streams or the web. There are standard preprocessing steps applied to this document corpus, followed by an appropriate choice of token models, representation methods, and labeling systems. Classification models are chosen to operate on train-validation-test splits, and classifiers are learned and stored.

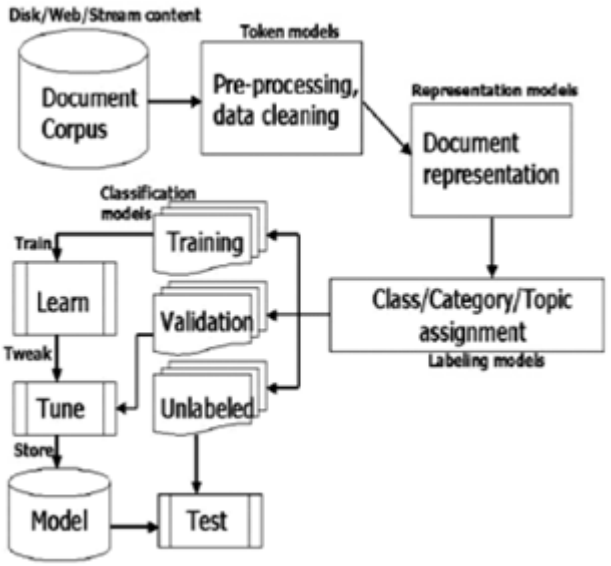

Figure 2: The standard text classification set up.

3.3. E-mail Preprocessing

E-mail preprocessing involves the process of transforming the training dataset into a representation suitable for the decision tree – ID3 (Iterative Dichotomiser 3) algorithm. This stage extracts the informational words from the data set. It consists of the following two steps: 1. Removal of non-discriminative words 2. Suffix stripping.

1) Removal of Non-discriminativeWords
In e-mails, certain words are most frequent and are not discriminative of a message contents, such as prepositions, pronouns and conjunctions. Examples of such words are —w", —that", —and"," this", etc. Some widely used conversational English words, such as —I'm", —isn't", —can't", etc. are of less importance. Elimination of these terms is performed in this step. Based on the theory of deception, a deceptive e-mail will have highly emotional words and action verbs. So, such words are set as keywords and extracted from the input dataset and the most frequent, but less deceptive words are eliminated in this step. Examples for highly emotional words and action verbs are —lifeless", —anger", —kll", —attack", etc.

2) Suffix Stripping
Suffix stripping is a process of removing the commoner morphological and inflexional endings from words in English. Its main use is a part of a term normalization process that is usually done when setting up information retrieval systems. The Porter stemming algorithm (or _Porter stemmer') is used to perform this process. Ignoring the issue of where precisely the words originate from we can say that a document is represented by a vector of words, or terms. Terms with a common stem will usually have similar meanings,

for example:
Assassinate
Assassinated
Assassinating

Frequently, the performance of an IR system will be improved if term groups such as this are conflated into a single term. This may be done by removal of the various suffixes -ED, -ING, -ION, -IONS to leave the single term assassinate. In addition, the suffix stripping process will reduce the total number of terms in the IR system, and hence reduce the size and complexity of the data in the system, which is always advantageous. Hence, those words which are extracted from the previous steps are suffix stripped to increase their efficiency.

Unfortunately e-mails are usually very noisy and simply applying text-mining tools to them, which are usually not designed for mining from noisy data, may not bring good results. Prior to indexing and classification, a number of preprocessing steps were performed.
1) E-mails were converted to plain-text from .mbox files.
2) Headers and HTML components were removed.
3) Body of the message was extracted.
4) The message body was tokenized into words, stop words were removed, and words were converted into lower case.

Figure 3 shows an example e-mail, which includes many typical noises (or errors) for text mining. Lines 1 and 2 are a header; lines from 4 to 8 are a signature. All of them are supposed to be irrelevant to text mining. Only line 3 is actual text content.

1. On Wednesday February 2007 13:39:42-0500, —X"
2. YYY@Domain.Com
3. Today there will be bomb blast in parliament house and the US consulates in India at 11.46 am. Stop it if you could. Cut relations with the U.S.A. Long live Osama Finladen Asadullah Alkalfi.
4. —
5. ---
6. Best regards
7. X
8. ---

Figure 3: Example of e-mail message.

Figure 4 shows an ideal output of cleaning of the e-mail in Figure 3 within it; the non-text parts (header, signature and quotation) have been removed. The text has been normalized. Specifically, the extra line breaks have been eliminated.

1. bomb
2. blast

Figure 4: Cleaned e-mail message.

In this paper, we formalize the e-mail-cleaning problem as that of non-text data filtering and text data normalization. By _filtering' of an e-mail we mean a process of removing the parts in the e-mail which are not needed for text mining, and by _normalization' of an e-mail we mean a process of converting the parts necessary for text mining into texts in canonical form (like a newspaper style text).

Header, signature, quotation (in forwarded message or replied message), program code, and table are usually irrelevant for mining, and thus should be identified and removed (in a particular text mining application, however, we can retain some of them when necessary).

4. Conclusions and Further Work

E-mail is an important means for communication. It is a possible source of data from which potential problem can be detected. In this paper, we have employed decision tree-based classification approach to detect e-mails in relation to criminal activities. All the e-mails were classified as suspicious (1) or not (0). From this experiment, we can find that a simple decision tree classifier can provide better classification result for suspicious e-mail detection. In the near future, we plan to incorporate other techniques such as different ways of feature selection, and classification using other method. One major advantage of the decision tree-based classifier is that it doesn‘t assume that terms are independent and its training is relatively fast. Furthermore, the rules are human understandable and easy to be maintained. The proposed work will be helpful for identifying the deceptive email and will also assist the investigators to get the information in time to take effective actions to reduce criminal activities. In the future, we would add following features to the paper: automatic reply to the incoming e-mails that are found deceptive and enabling our application to work in mobile environment.

A problem we faced when trying to test out new ideas dealing with e-mail systems was an inherent limitation of the available data. Because we only have access to our own data, our results and experiments no doubt reflects some bias. Much of the work published in the e-mail classification domain also suffers from the fact that it tries to reach general conclusion using very small data sets collected on a local scale.

References

[1] S. APPAVU ALIAS BALAMURUGAN, R. RAJARAM, S. SENTHAMARAI KANNAN, A Novel Data mining approach to Detect Deceptive Communication in Email Text. Proceedings of the National Conference on Advanced Computing, MIT, Chennai, India, (2007), pp. 179–188.
[2] A. BADIA, M. M. KANTARDZIC, Link Analysis Tools for Intelligence and Counterterrorism. Proceedings of the IEEE International Conference on Intelligence and Security Informatics, Atlanta, GA, (2005), pp. 49–59.
[3] W. COHEN, Learning rules that Classify Email. In proc.of the AAAI Spring Symposium on Machine Learning in Information Access, (1996).
[4] B. CUI, A. MONDAL, J. SHEN, G. CONG, K. TAN, On Effective Email Classification via Neural Networks. In Proc. of DEXA, (2005), pp. 85–94.
[5] G. WANG, H. CHEN, H. ATABAKHSH, Automatically Detecting Deceptive Criminal Identity. Comm. ACM, (2004), pp. 70–76.
[6] J. HAN, M. KAMBER, Data Mining Concepts and Techniques. Morgan Kaufmann Publishers, 2004.
[7] J. TANG, H. LI, Y. CAO, Z. TANG, Email Data Cleaning. Proceedings of KDD, Chicago, USA, (2005).
[8] P. S. KEILA, D. B. SKILLICORN, Detecting Unusual and Deceptive Communication in Email. Technical Reports, (June 2005).
[9] S. KIRITCHENKO, S. MATWIN, S. ABU-HAKIMA, Email Classification with Temporal Features. Intelligent Information Systems, (2004), pp. 523–533.
[10] S. YOUN, D. MCLEOD, A Comparative Study for Email Classification. Proceedings of International Joint Conferences on Computer, Information, System Sciences and Engineering, Bridgeport, CT, (2006).
[11] S. SHANKAR, G. KARYPIS, Weight Adjustment Schemes for a Centroid based Classifier. Computer Science Technical Report TR00-35, (2000).
[12] X. WU, X. XINGQUANZHU, Data Acquistation with Active and Impact Sensitive Instance Selection. 16th IEEE interactive conference, (2004).
[13] Y. YANG, An Evaluation of Statistical Approaches to Text Categorization. Journal of Information Retrieval, Vol. 1, No. 1/2, (1999), pp. 67–88.
[14] Y. YANG, J. PEDERSEN, A Comparative study on Feature selection in Text Categorization. In ICML, (1997), pp. 412–420.
[15] Z. HUANG, D. D. ZENG, A Link Prediction Approach to Anomalous Email Detection. Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Taipei, Taiwan, (2006).

Author Profile

Pulim.Pradeep Reddy Obtained the B.Tech. Degree in Information Technology (IT) from Malineni Lakshmaia Engineering College, Kanumalla , Singarayakonda . At present persuing the M.Tech in Computer Science (CS) at RISE Gandhi Group of Institutions, Ongole.

D. Rajesh obtained the B.Tech Degree in Information Technology (IT) from ASR Engineering College in 2005 and obtained the M.Tech Degree in Software Engineering (SE) from CVSR Engineering College in 2010. At present working as Associate professor .He has 10 years of teaching experience and working in Computer Science and Engineering(CSE) Department at RISE Gandhi Group of Institutions, Ongole.