ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Detecting Anomalies by Online Techniques Using Spam Detection

R. Dharani M.E (CSE) [1], S. Subashini M.Tech (IT) AP/CSE [2]

Kathir College of Engineering, India[1, 2]

ABSTRACT: In data mining and machine learning anomaly detection has been an important research topic. Intrusion or credit card fraud detection is the many real world applications are require effective and efficient frameworks to identify deviated data instances. Most anomaly detection methods are typically implemented in batch mode, and thus cannot be easily extended to large-scale problems without sacrificing computation and memory requirements. An online oversampling principal component analysis is an existing algorithm to address this problem. The proposed system consists of a spam detection method to detect the spam data in the user login. This approach is used to analyses the data with same text. Using stopping words and stopping terms methods the spam data's are secured with the algorithm as effective and efficient in the Networks.

KEYWORDS: inverse document frequency, Anomaly Detection, stop words removal, streaming words.

I. INTRODUCTION

Anomaly detection (or outlier detection) is the identification of items, events or observations which do not conform to an expected pattern or other items in a dataset. Typically the anomalous items will translate to some kind of problem such as bank fraud, a structural defect, medical problems or finding errors in text. Anomalies are also referred to as outliers, novelties, noise, deviations and exceptions. The context of abuse and network intrusion detection, the interesting objects is often not rare objects, but unexpected bursts in activity. This pattern does not adhere to the common statistical definition of an outlier as a rare object, and many outlier detection methods (in particular unsupervised methods) will fail on such data, unless it has been aggregated appropriately. Instead, a cluster analysis algorithm may be able to detect the micro clusters formed by these patterns. The importance of anomaly detection is due to the fact that anomalies in data translate to significant, and often critical, actionable information in a wide variety of application domains.

ANOMALIES

Anomalies are patterns in data that do not conform to a well defined notion of normal behavior. Figure 1 illustrates anomalies in a simple two-dimensional data set. The data has two normal regions, N1 and N2, since most observations lie in these two regions. Points that are sufficiently far away from these regions, for example, points o1 and o2, and points in region O3, are anomalies. Anomalies might be induced in the data for a variety of reasons, such as malicious activity, for example, credit card fraud, cyber-intrusion, terrorist activity or breakdown of a system, but all of the reasons have the common characteristic that they are interesting to the analyst. The interestingness or real life relevance of anomalies is a key feature of anomaly detection.

ANOMALY DETECTION IN TEXT DATA

Anomaly detection techniques in this domain primarily detect novel topics or events or news stories in a collection of documents or news articles. The anomalies are caused due to a new interesting event or an anomalous topic. The data in

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

this domain is typically high dimensional and very sparse. The data also has a temporal aspect since the documents are collected over time. A challenge for anomaly detection techniques in this domain is to handle the large variations in documents belonging to one category.

II. DETECTING ANOMALIES ON EXPERIMENTAL SETUP

All of the unsupervised anomaly detection techniques we have developed are applicable to both small pieces of text (like a short paragraph) and to large pieces of text like documents or books. We chose to experiment on the task of finding sections in a document that are anomalous, but there is nothing inherent in the methods to suggest that they must be used detect anomalies within a document; they could equally well be used to detect whole documents that are anomalous with respect to a collection. The task is always about finding text that does not belong or is unusual with respect to its surroundings and it is just a matter of scope as to what those surroundings. Many thousands of different experiments were run to detect anomalous segments in documents. Describing these with a view to investigating what types of anomaly are easiest to detect, the effect text size has on anomaly detection, the impact of Standardization, and the anomaly detection technique that performs best. The detection of anomalies focus primarily on detecting when the author, genre, writing style, or topic is anomalous. In these experiments we take a document that contains an anomaly and feed it to our anomaly detection program. This program returns a list of all segments ranked by how anomalous they are with respect to the whole document. If the program has performed well, then the truly anomalous segment should be at the top of the list (or very close to the top). Our assumption has been that human wishing to detect anomaly would be pleased if they could find the truly anomalous segment in the top 3 or 5 segments marked most likely to be anomalous, rather than having to scan the whole document or collection. This may not be the case in situations where there is no reason to believe that anomalies exist. Test documents are artificially created by taking a document" made up of random segments from a single source and inserting a randomly chosen segment from a different source. In this scenario, the source which makes up the majority of a document is the normal population while the single inserted segment is anomalous with respect to that population. Our anomaly detection procedures are then run over this artificially created test document with the goal of identifying the inserted segment from the different source as an anomaly. We created thousands of documents in this manner from different sources and using different sized segments, but always inserted a segment from one source into a collection of segments from a different sources.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

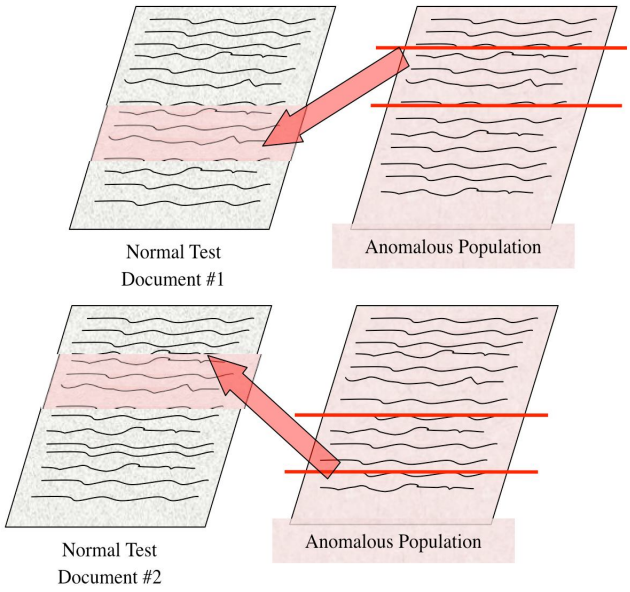

Fig 1.1 Detecting anomalies on documents or text

DIFFICULTIES

A botnet is a collection of internet-connected computers whose security defences have been breached and control ceded to a malicious party. Each such compromised device, known as a "bot", is created when a computer is penetrated by software from a malware distribution; otherwise known as malicious software. The Internet is plagued by malicious activity, from spam and phishing to malware and denial-of-service (DoS) attacks. Much of it thrives on armies of compromised hosts, or botnets, which are scattered throughout the Internet. However, malicious activity is not necessarily evenly distributed across the Internet: Some networks may employ lax security, resulting in large populations of compromised machines, while others may tightly secure their network and not have any malicious activity. If Attackers tries to login with the guessing password attack then also the session carried out to the original application. The spam message is blacklisted but senders IP address is not monitored to block the spam message forwarded from particular IP. Botnets consist of groups of compromised machines used for malicious purposes on the Internet which is not identified. While data transmission the packet loss by which network is not identified and not blacklisted.

MAIL SERVER CREATION

Mail server creation is the first module in this project. Initially the mail server is created to communicate with the difference persons through email server. The mail server environment includes the option for sending email through recipient through composer option and they can receive the mail from various recipients. The mail server also has the option for viewing sent mails, spam mails and deleted mails.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

ADMIN

This is the main module, which will maintain the overall control of the users from both remote login and known machine. Admin and server are the main part of the communication in the user's side. The server responses to the users query if only they have accessibility rights. Here the administrator has the information about the attacks and the blacklisted information. After identifying the message as spam, the administrator verifies the message comes from which IP address. If the same spam sender sending the repeated spam message then the particular content will added to the spam database table and they cannot able to forward the spam messages.

USERS

Users are the end persons who are making or initialize the communication with the server. Normally in this work users can split as

- Legitimate users and
- Attackers

The attackers are mainly from the remote system logins and they are uses the compromised systems.

- Known machine
- Unknown or remote login system.

WEB ACCESS SECURITY

Due to their ubiquitous use for personal and/or corporate data, web services have always been the target of attacks. These attacks have recently become more diverse, as attention has shifted from attacking the front end to exploiting vulnerabilities of the web applications.

ACCESSIBILITY AND VERIFICATION

This module will get the cookie storage for the checking purpose while the user login from remote systems. If the user is first time then the cookie will store the data and can't verify things for access.

REMOTE LOGIN LIMITATION

Once the protocol finds the numerous of wrong guess for a single as well for multiple login then the process will make some limitation for the remote login identity.

COOKIE VERIFICATION

This module keeps the copy of every single user login in the remote login as well the known machine, further it will helps to verify the data when wrong entry implies.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

SPAM DETECTION

PREPROCESSING

The first step towards handling and analyzing textual data formats in general is to consider the text based information available in free formatted text documents. Initially the pre-processing is done with existing spam document by following process.

Removing stop words and stem words

The first step is to remove the un-necessary information available in the form of stop words. These include some verbs, conjunctions, disjunctions and pronouns, etc. (e.g. is, am, the, of, an, we, our) and Stemming words e.g. 'deliver', 'delivering' and 'delivered' are stemmed to 'deliver'.

Term frequency

The data is given a suitable representation based on words or terms defined in the text. Different data representations methods i.e. term frequency (TF), inverse document frequency (IDF) and term frequency and inverse document frequency methods may be used at this level of information processing.

This is very important module in this project. While composing new mail, server performs pre-processing and clustering to identify whether the user is sending spam content. If the content is identified as spam then the content will be blocked by the server so the message will be blocked and cannot be transmitted to recipient. This process reduces traffic and zombies attack in the network.

REPORTS

It is the outcome of admin checking and the cookie detail for the whole process; here the admin can refer the limitation to suite for particular remote login system and filtering the spam sender activity list.

III. CONCLUSION AND FEATURE ENHANCEMENT

The spam mail is viewed by the user if any hackers are attack the original data at a time. By mails the anomaly or outliers are detected by the user and also with the help of admin. Feature enhancement is improved in proposed system itself.

REFERENCES

1.Yuh-Jye Lee, Yi-RenYeh, and Yu-Chiang Frank Wang, Member, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, Vol. 25, No. 7, July 2013
2. Lee, W., Xiang, D.: Information-Theoretic Measures for Anomaly Detection. Proceedings of the IEEE Symposium on Security and Privacy, IEEE Computer Society Press, Oakland,CA (2001) 130-143
3. Abe, N., Zadrozny, B., and Langford, J. 2006. Outlier detection by active learning.In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM Press, New York, NY, USA, 504-509.
4. M. Markou and S. Singh, "Novelty detection: a review part 1: statistical approaches," Signal Processing, vol. 83, pp. 2481 – 2497, Decemeber2003.
5. Axelsson, S., Research in Intrusion Detection Systems: A Survey, Technical Report No. 98-17,Dept. of Computer Engineering, Chalmers University of Technology, Gteborg, Sweden, 1999.
6. C.C.Aggarwal Re-Desining functions and Distance Based Applications for High Dimensional Data.ACM SIGMOD Record,March 2001.
7. BAI, Z.-J., CHAN, R. AND LUK, F. Principal component analysis for distributed data sets with updating. In Proceedings of International workshop on Advanced Parallel Processing Technologies (APPT), 2005.