International Journal of Recent Engineering Research and Development (IJRERD)
Volume No. 02 – Issue No. 03, ISSN: 2455-8761
www.ijrerd.com, PP. 46-47

Spam detection system based on data mining for websites

Ishrath Nousheen
CSE Dept , Nawab Shah Alam Khan College of Engg & Tech,
Hyderabad, India,

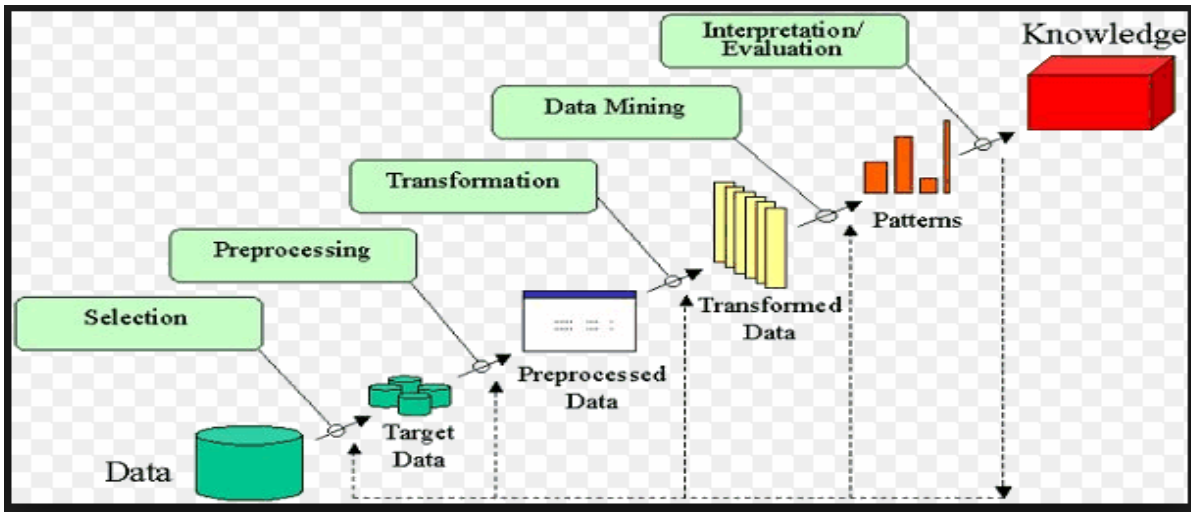
Abstract: As we observe now a days that all transactions and communications, whether general or business are performed through-mails. More time and cost can be saved through emails. It is the most effective tool for communication. As we experiencing in mailing that, these are also affected by spam. Spam is an irrelevant or unsolicitated messages sent over the internet, for the purpose of phishing, spreading malware, etc. I am going to explain about GAD algorithm.

Index terms: Classifier, Feature Selection, E-mails, GAD algorithm

1. Introduction

The very much favorite means of communications is E-mailing. It is a way of communication as it saves a lot of time and costs. It provides a way for internet users to easily transfer information globally. But sometime the mails can be attacked by active or passive attacks.

Data mining is about finding insights which are statistically reliable, unknown previously and actionable from data (Elkar 2001). We will be using the term fraud. Here the term fraud refers to the abuse of a profit organization system without necessarily leading to direct legal consequences. Fraud can become a business critical problem if it is very prevalent and if the prevention procedures are not fail-safe(1).



2. Background

Earlier algorithm for checking the spam were like Regression Analysis, Baye's theorem & keyword matching were used for uncovering patterns, this algorithm were incapable of handling large data set can be applied on small data sets. In olden days spam checking was done manually or generic algorithms were used [2].

As data sets have increased their sizes so we have to implement the data mining using Machine learning [3] data processing technique, we can easily disclose the hidden pattern and spam.

3. GAD Clustering Algorithm

Clustering is a data mining technique widely used in number of applications. Many studies have been conducted on it. there are some preprocess methods such as sampling , sub space and compression. They have been compressed in order to reduces the data to smaller size to achieve speedup. Here I am introducing the GAD algorithm. It performs fast clustering .GAD is most general solution to exploit activity detection for fast clustering.

3.1 Notations Let *D* be number of dimensions, *N* be the number of patterns, *K* be the number of centers. Suppose there are *I* iterations in the algorithm.[5] At each iteration *i*, for a pattern *p*, we have NC(*i,p,,j*) represents pattern p's *jth* nearest center. In real implementation, the value of NC(*i,p,j*) is the center's id.

We usually denotes the General Activity Detection as a function of four parameters:

GAD(M,S,m,B)

Where M denotes Search Methods, S denotes Activity States, *m* denotes the number of nearest centers, and B denotes Boundary

Algorithm we are analyzing the GT, GT is an exact clustering algorithm. It is faster than k-means and gets the nearest results. GT saves the each pattern's nearest center. The idea is that, if a patterns previous nearest center is static or moves closer to the pattern. This process is repeated for each iteration.

At iteration *i*, NC (i, p, 1) = C1, NC(i, p, 2) = C2. At iteration i+1, C1 and C2 are active, and C3 is static. Since Dist(i+1,p,C1) < D-NC(i, p, 2), CGAUDC searches from active centers to determine the pattern's nearest centers, C3 is ignored. The result is: NC(i + 1, p, 1) = C1, NC(i + 1, p, 2) = C2. The nearest center is correct but the 2nd nearest center is wrong. At iteration i+2, C1 and C2 are active, C3 is static. Since Dist(i+2, p, C1) < D-NC(i + 1, p, 2), CGAUDC searches from active centers and ignores C3 again. The result is: NC (i + 2, p, 1) = C1, NC(i + 2, p, 2) = C2. Both the nearest and the 2nd nearest center are wrong.

The problem of GT (and CGAUDC) is the that they only consider the first nearest (and the second nearest) center, which is not able to fully explore the power of activity detection, as shown in Section 5.1. We propose GAD (General Activity Detection) to consider any m number of nearest neighbors. Such extension is not a simple task, CGAUGC directly extends GT to consider 2 nearest neighbors but fails to get the exact result. To solve the problem, we introduce the idea of m-Boundary to make sure we can extend to any m without getting error. Our exact GAD algorithm is faster than GT and CGAUDC because it is able to achieve the low-bound of activity detection.

4. Conclusion

This paper demonstrates the prevention of mails from spam or malware. It also deals with the prevention from attacks. Here I have explained about the GAD algorithm. This particular algorithm performs very fast clustering, which is more helpful for us. Some of the notations I have explained in detail about GAD algorithm.

References

[1]. Clifton phua, vincent lee , kate smith & ross gaylera, "Comprehensive Survey of Data Mining-based Fraud Detection Research".

[2]. Ritesh Kumar, Shital Ghadge, G.S. Navale, "Spam Detection using Approach of Data Mining for Social Networking Sites".

[3]. Xin jin, Cindy Xide Lin,Jiebo Luo, jiawei Han "A Data Mining based Spam Detection System for Social Media Networks", PVLDB, 2011 .

[4]. Harshal S. Multani , Amrita Sinh Marod, Vinita Pillai , Vishal Gaware "Spam Detection in Social Media Networks: A Data Mining Approach"

[5]. Xin Jin Sangkyum Kim Jiawei Han Liangliang Cao Zhijun Yin ,"GAD: General Activity Detection for Fast Clustering on Large Data"