CSCI 4511 writing 3

Zirui Chen

March 2022

Morden spam filtering techniques can be generally categorized into two categories, namely reputation based filtering and content-based filtering.[1] Reputation based filtering can be further categorized as origin based, social networks, traffic analysis and protocol based filtering. Content-based filtering can also be further categorized as heuristics fingerprint based and machine learning based.

Black-list and white-lsit is one widely adopted method of reputation based filtering. The idea is to set up an Real time IP address balck-list. [1]Every Time user receives a message, the filter will firstly compare the IP address of the sender with IPs in balck-list. If there is a match, the email will be automatically sent to trash. However, this technique is inherently flawed since there are many ways to get around it. One of the techniques is the use of botNets. Botnets are capable of sending massive amounts of spam messages with different IP addresses. This will significantly increase the number of IP addresses on the black-list. On the other hand, putting messages into the trash without taking a look at the content will increase the odds of false positives. False positives in this case means that legit messages are wrongfully categorized as spam messages. The cost of having false positives is much more expensive than false negatives.

Another approach under reputation-based filtering is traffic analysis.[2] Under the presumption that spam emails from one spammer usually have the same content, messages can then be categorized under the similarity of message content. If one category has lots of similar messages, a method can be applied to filter such messages. One method used during the IEEE 2006 conference is called Bulk Mil Traffic Classification. The result indicates that it has 70.4 percent of accurate rate to filter out spam mails in a fast and cheap manner.

Content-based filters are believed to be the method that can achieve the highest accuracy. In morden machine learning based content filters, various machine learning techniques are used in combination with natural language

processing techniques. Before any machine learning methods can be applied, data must be tokenized, stemmed and cleaned. Several natural language processing techniques can be used to help feature extraction, which would significantly increase accuracy.

The most commonly used content-based filtering method is Bayesian classifier. This classifier method implements a statistical method called maximum posteriori probability(MAP).[3] Firstly, we need to obtain a data set, which should clearly indicate the data type(spam or ham). Secondly, the data set should be randomly split to 20 percent:80 percent. The 80 percent shall be used as the training set and 20 percent as the test set. Bayesian can then be used to classify the data in a training set. The model shall be tested using the test data. Results should be evaluated by four scores, namely Precision, F-score,recall and accuracy.[4] As discussed earlier, NLP can play an important role in feature extraction. There are several approaches such as the N-gram model[5], pos tagging and message entropy.[6]The bayesian classifier can be fed with data that is processed by different MPL methods to achieve a better model. The Bayesian classification has an accuracy of over 95percent when properly implemented. However, it is also subject to many vulnerabilities such as bayesian poisoning. The idea behind bayesian poisoning is to poison the model by feeding the classifier wrongfully marked data(e.g. Sending legit messages and reporting them as spam).[7]

Another simple approach of content-based filtering is called k nearest neighbors.[3] The idea is roughly the same as the Bayesian classifier, the only difference is that the modeling process. This algorithm has a predefined number of centers, and the center will be automatically updated based on the data. Once modeling is completed, an outgoing message will be evaluated, if a large portion of it is in k neighborhood, then it would be categorized as spam.

One of the most recent techniques is called a support vector machine a support vector machine is a framework that minimizes risk.[8] It's developed by Vapnik based on optimal classification. This framework, in its essence, is to find a hyperplane that separates spam and ham classes. Under the circumstance that the datas are not linearly spepartelable, the hyperplane will be found in higher dimensions.[3] This framework is deemed to be efficient and accurate.

One natural language processing technique used for spam detection is called maximum entropy. Entropy means the amount of information composted in a piece of information. The idea is to find the "appropriate probability distribution that maximizes the entropy".[3]

We must note that in morden spam detection systems, all aforementioned techniques work together, in whole or in parts, to ensure the best accuracy

of spam detection. To deliver the state of the art spam filter would require millions of data and years of effort. My project would focus on how different NLP processing techniques would affect the accuracy of sample machine learning classification.

References

[1] Alexy Bhowmick Shyamanta M.Hazarika. Spam review detection techniques: A systematic literature review. 443, 2017.

[2] Ni Zhang, Yu Jiang, Binxing Fang, Xueqi Cheng, and Li Guo. Traffic classification-based spam filter. In *2006 IEEE International Conference on Communications*, volume 5, pages 2130–2135, 2006.

[3] Ahmed Khorsi. An overview of content-based spam filtering techniques. 2007.

[4] Shafi'I Muhammad Abdulhamid, Muhammad Shafie Abd Latiff, Haruna Chiroma, Oluwafemi Osho, Gaddafi Abdul-Salaam, Adamu I. Abubakar, and Tutut Herawan. A review on mobile sms spam filtering techniques. *IEEE Access*, 5:15650–15666, 2017.

[5] Tunga Gu¨ngo¨r Ali C¸ ıltık. Time-efficient spam e-mail filtering using n-gram models. 2007.

[6] Urko Zurutuza José María Gómez Hidalgo AEnaitz Ezpeleta. Does sentiment analysis help in bayesian spam filtering? 9648, 2016.

[7] Enrico Blanzieri · Anton Bryl. A survey of learning-based techniques of email spam filtering. *Springer*, 2008.

[8] Hamid A. Jalab Thamarai Subramaniam and Alaa Y. Taqa. Overview of textual anti-spam filtering techniques. 2010.