IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 1, Ver. VIII (Feb. 2014), PP 93-96
www.iosrjournals.org

Anti-spam Filter Based on Machine Learning Algorithm

Khandekar Amol, Garje Pramod, Rakesh Thakare
(Computer engineering,Pune university,India)

Abstract: *We building one of the filter which help the user relief from the unwanted mails in his inbox of the mail account which is develop using the machine learning algorithm and which also contain the filter.*

In this we creates product which should secure the user from flooding of the unwanted mail we can call it as spam.. There are number of machine learning algorithm from this machine learning algorithm we are used the best of them.. According to the speed and efficiency .In this we are training a system according use of user. So it gives better result while using this system.

In the current trend e-mails are most widely used and business form of communication.

The aim of our project implements unwanted email filtration technique. Because of spam mail create of data changes or bandwidth and processing time of internet services providers and also contains un familiar content.

So we build such product that will satisfy user need and security to user from unfamiliar.

I. Introduction

Email spam has widely used since the early 1990.80% of the spam are seen by virus infected computers. It happens during the advertising on the internet. Spammers collect the information of email address from viruses, websites, customer list and send to other spammer. Our software goal is filter the spam mails by using machine learning algorithm and classify mails into different categories. So we create such a product that will fulfill user need and secure then from hazardous content of unwanted mails (spam mail) that might be grass then system performance. Now days increasing in amount of spam mails that are why it leads into wastage of bandwidth and processing time. So this software tool is very effective for identifying spam mails which are incoming to words the user that will be automatically get classified most of filter available to create a list of legitimates sender list of unwanted mailer and handcraft rules that blocks the mail which contain unfamiliar word or phrase previously we work on anti-spam filtering and the performance of many famous machine learning algorithm including a support vector machine, boosting with C4.5, Naïve Bayes.

The most anti spam filter currently depend on blacklist, white list and hand made rules. That detect for name or email address in the white list side, few filter sent reply to sender not in the white list asking them some simple question to answer it. Whose body looks like legitimate this may leads to conflict or misclassification. There is no of publication, developers of filter which classifies the spam are more aware of power of machine learning. For example anti spam filter best on Navie Bayes was recently get into Mozilla's email client. The filtering system is the best on train message. It has special folder for wanted mails which are legal to user and unwanted or illegal to user (spam mail) are more into special folder. This filter support incremental learning it earns the user can adjust accordingly without considering all message collection for training. Pop file is same that it also present on user computer by using Navie Bayes and incremental learning. It creates, basic filter that are only responsible for categories legitimate message into different categories legitimate message into different categories. It also work as email proxy for POP3 server and message get added its tag. It consider spam one of the most important thing of POP file is its preprocessors. Bayesian filtering is the best on the principle that most things are dependant and things occur in future can be inferred from previously occurred of that thing.

Our system is divided in two parts.

1. Training mail system.
2. Classification mail system

II. System overview

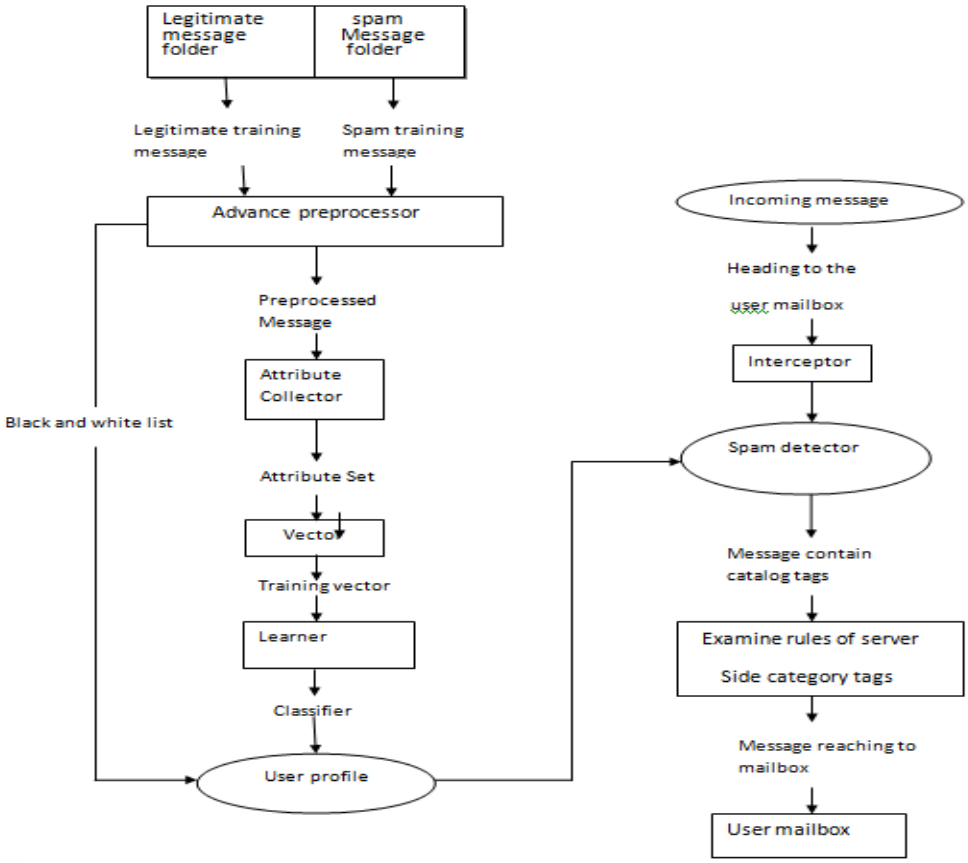

Fig. System overview

Training mail system has its own collection of spam mail which has its own collection of we regularly updates this collection and its contents in built 2500 spam message are sent blindly without interest of user the efficiency of filter get increase when it becomes more older. We got number of useful attributes for filtering spam that are characteristic of legal mails to be included in training mail system.

The advance preprocessor scan training mails and remove attachment, duplicates, HTML tags. It also replace token by unique number to make publicly available.

The advance preprocessor creates list of legal message with the addresses the user has receive wants messages from and a list of illegal messages the address of the sender of all spam mails. For this system retain only the first five mails from each sender. Mostly those senders unlikely send illegal messages and their addresses also stored into address book of user that means the individual sender.

Attribute collector identifies the good attribute into the vector representation seen it is not practical in terms of storage and computation to consider pool of candidate attribute. Attribute are collected according to their gain major of information. It has been view very effective in practice we uses 1-gram attribute only as our result does not provide any proof. On benefit of m-grams for m>1 collection of attributes to retain can be set at will. The minimum number of attributes according to their accuracy classification speed and training will changes for per user and is tedious to identify without specific that using hundreds of attributes is a better compromise.

Ones the attribute have been collected vector system convert that message which are for training into vectors all collection of attributes are in number with actual value of attributes.

A been define as Occ(d)/L(d) where Occ(d) is number of occurrence in document of the token represented by A and L(d) length of s. calculated in occurrences of token. The learning component gets the training vector along with the value for the parameter (lambda) the cost of misclassifying a spam message has legal message as spam the learning algorithm is provided by weka can be choose any one of algorithm in learning component. The default choice is support vector machine (SVM). Implementation it has better speed and accuracy. The component which use for learning are produces a specific classifier of user are resides with legal list of mail or while list and illegal list of user or black list in the user model.

III. First evaluations:

We seen into number of parameter of anti spam filtering and learning experimentally there effect on couple of bench mark derived by different user. The attribute vector included from parameter study, attribute collection, size of them and training mails set and learning algorithm which are Navie Bayes, logit boost a boosting and support vector machine algorithm are use by learner. The cost is and very important factor. All the experiments were done under a cost sensitive framework.

Table 1:

	Lambda=1			Lambda=9		
	Pr	Re	WAcc	PR	RE	WAcc
1 grams						
Naïve Bayes	91.50	95.73	93.43	85.14	91.21	94.78
Flexible Bayes	94.20	87.75	94.14	94.35	70.61	95.02
Logit boost	93.30	91.80	92.64	97.85	76.67	96.23
SVM	95.97	90.20	96.50	98.70	78.33	97.06

In above table weighted (%) accuracy, recall and precision of the machine learning algorithm. Considered for cost scenarios and attributes that gives better result.

IV. Second evaluation:

In this evaluation we judge that how actually machine learning algorithm in mail filter will work in real word. The system is use in PUS collection for training purpose this collection has 2313 legal message and 1826 illegal message. They are included in our filter at same time. The filter was configuring of number of scenario in which lambda=1 is also scenario where it simply detects spam mails according to user priority for incoming message. This spam mails moves into the special folder. Which may be sometimes leads to misclassification of legitimate message. Number of attributes are retain was up to 520 base on this evaluation result, no black list was use.

Support vector machine implementation was selected to train the filter. This gives good performance on pu3. It is the more popular instant of KERNEL learning of class method. The I/P data into some higher dimensional feature space is main idea to map in the learning algorithm problem separable.

Formula

$$F(y) = Wi.bi(y) + Wo$$

The new vector spam for dimensionality is n and the non-linear functions that map the original attributes to the new ones is denoted by bi(y). This above formula can be written another form also in which bi(y) will present in only dot product. The KERNEL function can be put behalf of their dot product which can the computational problem of working on high dimensional future space. The number of kernel classes are will know such as polynomial, Gaussian and application dependent parameter associated to it and some time it is important for the performance of support vector machine. In the research we have solution that the support vector machine is linear but it can perform non linear ones also. So the experimental result does not provide any proof for improvement in performance by using the polynomial KERNELs of various degrees. A linear support vector machine model was included.

Table 2:

Days used	212
Message received	6732(Avg. 31.75 per day)
Spam message received	1623(Avg. 7.65 per day)
Legitimate message received	5109(Avg. 24.10 per day)
Legitimate to spam ratio	3.15
Classified legitimate message correctly(L->L)	5057
Classified legitimate message incorrectly(L->L)	52 (Avg. 1.72 per week)
Classified legitimate message correctly(S->L)	1450
Classified legitimate message incorrectly(S->L)	173 (Avg.5.71 per week)
Precision	96.54% (PU3:96.43%)
Recall	89.34% (PU3:95.05%)
WAcc	96.66% (PU3:96.22%)

Table 2 : real evaluation result of filter, by using support vector machine with the 520 1-gram attributes for lambda =1. In this support vector machine learning algorithm messages are trained on pu3.

In above table we shown the result of evaluation and in this do not have the list of unwanted message (black list) was used. All mails that were categories as spam were detected by classifier of learner. Precision was very much similar to its corresponding score gain with 10 fold cross validation on PU3. Recall was lower (89.33% as opposed to 95.4%). Analysis of misclassified mail belong shields more focus on this issue. Overall performance of filter is quite good, thought this filter has scope for improvement. The mail user rules that message tagged as spam are moved to the special folder late on average some spam message per week (5.70). In the inbox than in single day he received (7.65). This filter it moves approximately to legal message or legitimate message are not correctly moved to special folder which week. For research purpose it checks that folder at the each week end.

Which does the misclassification much easier to know? He also failed that in many times the misclassification legitimate message were in different to him. For e.g.(newsletter , subscription, verification etc). An observation conform that misclassified message that are analyzed. It should be noted that filter was never retrained during the evaluation.

The system to keep both learning model and legitimate message list updated it leads to better result. Now we moved towards the misclassified message analysis form the starting with 173 misclassified spam messages. The message has very little text or non text. Those messages are known as hard spam. Which also contains mostly hyper links, message those hides the text behind image in attachment, massage contain in spam message. The elaborate preprocessor is responsible for catching the many of this message. In this project we see that there is arm race between spammer and developer of spam filter. In the future filter may be develop to incorporate optical character reorganization to avoid to message send as images, and they to follow links to web pages to handle the spam message which are only contains hyperlinks and without text. Approximately 8% of the misclassification spam message where doing advertisement of pornographic sites. In this message they take care of lighting friendly words with no hyperlinks to those sites.

The misclassified spam message those were written in other language except German language if filter has been trained during 6 months. This would have allowed filter to select as attribute from non-English words becoming common. The collection of spam message is get add the non English message.

Some spam message are misclassified which contained very unusual contain are 3% and it is very difficult to filter this type of spam message and train the system this message are more related to scientific research for e.g. processing natural language platform, which may be and induction of an attempt made by sender of spam mail (spammer to the target user groups) personalize message of this particular and kind are very difficult to detect. Those are very much similar to the vocabulary and contain of the legitimate message but it is more interesting to read.

V. Conclusion

We presented filter, a filter is best on machine learning algorithm for dictating spam mail of text category. Its real word evaluation that plays important role in anti spam filtering by conforming through machine learning algorithm.

There is arm race between filter developer and spammer. Retraining and spammer retraining is required regularly but more advance preprocessing during its six months evaluation period. Some functionality is missing we are planning for large scale of the current state implementation of system.

In the large run different approaches of filtering will mix, adding to be successful completion more than one algorithm by combining more than one algorithm to improve the efficiency is seems to be promising.

References

[1]. I. Androutsopoulos, G. Paliouras, and E. Michelakis. Learning to filter unsolicited commercial e-mail. Technical report, National Centre for Scienti_c Research \Demokritos", 2004.

[2]. H. D. Drucker, D. Wu, and V. Vapnik. Support Vector Machines for spam categorization. IEEE Transactions On Neural Networks, 10(5):1048{1054, 1999.

[3]. J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting. Annals of Statistics, 28(2):337{374, 2000.

[4]. J.M. Gomez Hidalgo and M. Mana Lopez. Combining text and heuristics for cost-sensitive spam filtering. In Proceedings of the 4th Computational Natural Language Learning Workshop, pages 99{102, Lisbon, Portugal, 2000.

[5]. T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In Proceedings of the 10th European Conference on Machine Learning, pages 137{142, Heidelberg, Germany, 1998.

[6]. G.H. John and P. Langley. Estimating continuous distributions in Bayesian classifiers. In Proceedings of the 11th Conference on Uncertainty in Arti_cial Intelligence, pages 338{345, Montreal, Quebec, 1995

[7]. D.D. Lewis. Naive (Bayes) at forty: The independence assumption in information retrieval. In Proceedings of the 10th European Conference on Machine Learning, pages 4{15, Chemnitz, Germany, 1998.

[8]. T.M. Mitchell. Machine learning. McGraw-Hill, 1997.