A Spam Filtering Technique using SVM Light Tool

Agrawal Rammohan Ashok[1], Gaurav Shrivastav[2]

[1]M. Tech. CSE, RKDFIST, Bhopal, India

[2]Assistant Professor, CSE, RKDFIST, Bhopal, India

Abstract: ***Spam has become a headache of the today's internet causing many problems to all users on Internet. Spam emails not only consume computing resources, money but are frustrating. A lot of money is being spent/ lost due to spam globally each year. Data mining approaches for content based spam filtering have been seen promising. Filtering is the one of the important technique to stop spam. Now-a-days there is a tremendous need of some trust-worthy and adaptive spam filtering system in the market which should have the ability to react quickly to the real time changes and provide fast and qualitative self-tuning in accordance with a new set of features. This paper explores the use of support vector machines for classifying and filtering spam messages and also to improve the accuracy and time performance. Also a comparative analysis among the algorithms is also being presented here with a view to get the optimized results in spam filtering.***

Keywords: Data Mining, Machine Learning, Spam Classification, Spam Filtering, Support Vector Machine

1. Introduction

Spam is flooding the internet with many copies of the same message or unwanted messages in an attempt to force the message on people who would not otherwise choose to receive it. Mostly spam is commercial advertising. Spam costs the sender very little to send and most of the costs are paid by recipients or the carriers rather than by the sender. E-mail spam is a subset of spam that involves nearly identical messages sent to numerous recipients by e-mail [1]. Day by day the amount of incoming spam increase, scammer attacks are becoming targeted & consequently more of a threat. Below is a graphic that's based on spam data collected by Symantec**'s** Message Labs. It shows that global spam volumes fell and spiked fairly regularly, from highs of 6 trillion messages sent per month to just below 1 trillion, but still the problem is not being solved completely and some measures are to be adopted in order to solve this issues. This graph is based on Symantec's raw spam data [3] computed with the available global spam data.

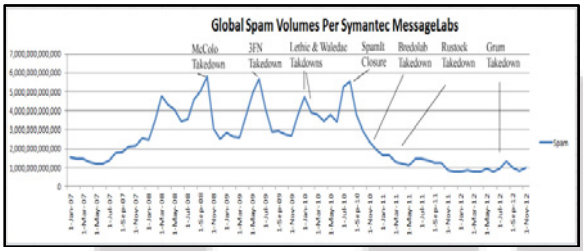


Figure 1: Spam volumes from 01-01-07 to 01-11-12[3]

Spam should be figure out with some new effective approaches otherwise few problems may arise as such:

- Consumes computing resources and time.
- Reduces effectiveness of legitimate advertising.
- Cost Shifting.
- Fraud with Identity theft.

Spam volumes are region dependent and providers you rely on for email and connections to the internet. The following is a Spam data from *Cloudmark*, a San Francisco-based email security firm. Their data (shown in the figure below) paint a very interesting picture of the difference in percentage of email that is spam coming from users of the top three email services: Spam percentage recorded were Yahoo (22%), Microsoft (11%) & Google (6%).

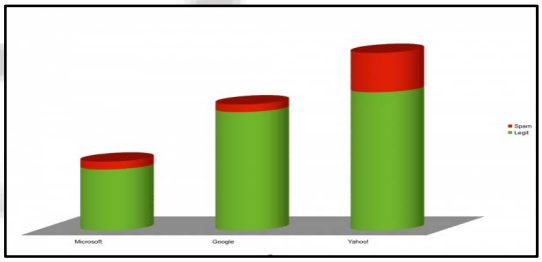


Figure 2: Spam Percentage by Service Providers [3]

2. Literature Survey

Spam mail is sent to a group of recipients who have not requested it. These mails are causing many problems such as filling mailboxes, engulfing important personal mail, wasting network bandwidth, consuming users time and energy to sort through it and many more[11]. According to a series of surveys conducted by CAUBE.AU 1, the number of total spams received by 41 email addresses has increased by a factor of six in two years (from 1753 spams in 2000 to 10,847 spams in 2001)[14]. Therefore it is challenging to develop spam filters that can effectively eliminate the increasing volumes of unwanted mails automatically before they enter a user's mailbox.

D. Puniskis [12] in his research applied the neural network approach to the classification of spam. His method employs attributes composed of descriptive characteristics of the evasive patterns that spammers employ rather than using the context or frequency of keywords in the message. The data used is corpus of 2788 legitimate and 1812 spam emails received during a period of several months. The result shows that ANN is good and ANN is not suitable for using alone as a spam filtering tool.

In [13] email data was classified using four different classifiers (Neural Network, SVM classifier, Native Bayesian Classifier, and J48 classifier). The experiment was performed based on different data size and feature size. The final classification result should be '1' if it is spam otherwise

'0'. This paper shows that simple J48 classifier which make a binary tree, could be efficient for the dataset which could be classified as binary tree.

In [2] author provides a comprehensive analysis of various classifiers using different software tools viz. WEKA, RapidMiner which were to be implemented on a common dataset for implementing the supervised learning approach for spam classification.

3. SPAM Filtering Techniques

Depending on used techniques spam filtering methods [1] are generally divided into two categories:

1. Avoid spam distribution in their origins methods
2. Avoid spam at destination point methods.

Spam must be detected initially by using some spam detection [10] techniques in order to lower the costs of implementation and then filtered to improve the results using some filtering techniques as:

- Hiding contact information
- Looking for an filtering software

3.1 Theoretical Approach for Spam Filtering

[1] Depending on used theoretical approaches spam filtering methods are divided into traditional, learning-based and hybrid methods. In traditional methods the classification model or the data (rights, patterns, keywords, lists of IP addresses of servers), based on which messages are classified, is defined by expert. The data storage collected by experts is called as the *knowledge base*. There are also used trusted and mistrusted senders lists, which help to select legal mail. Actually it makes sense only creation of the "white" list, because spammers use fictitious e-mail addresses. This technique can't represent itself as a high-grade anti-spam filter, but can reduce considerably amount of false operations, being a part of e-mail filtration system based on other classification methods. In learning-based methods the classification model is developed using "Data Mining techniques". There are some problems from the point of view of data mining as changing of spam content with time, the proportion of spam to legitimate mail, insufficient amount of training data are characteristic for learning based methods. Some Traditional methods [1] are:

- Acceptance of sender as a spammer.
- Verification of sender mail address & domain name.
- Content Filtering.

3.2 Acceptance of Sender as a Spammer

These methods rely on different blackhole lists of IP and e-mail addresses. It is possible to apply own blackhole and white lists or to use RBL services (Real-time Blackhole List) and DNSBL (DNS-based Blackhole List) for address verification. Advantage of these methods is detection of spam in early step of mail receiving process. Disadvantage is that the policy of addition and deletion of addresses is not always transparent. Often the whole subnets belonging to providers get to the Black lists. For such systems it is actually impossible to estimate the level of false positives (the legitimate e-mail wrongly classified as spam) on real mail streams.

3.2.1 Verifying sender mail address & domain name:
This is the simplest traditional method of filtration if DNS request's name is same with the domain name of sender. But spammers can use real addresses so that current method becomes ineffective. In this case it may be verified with possibility of sending the message from current IP address. Firstly, the Sender ID technology [4] can be used where sender's e-mail address is protected from falsification by means of publishing the policy of domain name use in DNS. Secondly, there can be used SPF (Sender Policy Framework) technology [5], where DNS protocol is used for verification of sender's e-mail address. The principle is that if domain's owner wants support SPF verification, then he adds special entry to DNS entry of his domain, where indicates the release of SPF and ranges of IP addresses from where may become an email from users of current domain.

Traditional method's disadvantages are:

- Necessary to update the knowledge base regularly.
- Dependence on update suppliers.
- Low Security level.
- Dependence on natural language of correspondence.
- Low level of detection because of general models of classification.

3.2.2 Traditional Content Filtering:
This technique was used to filter the contents of the received mails. A mail may be pdf file, an excel document, an image or even an mp3 file. The problems with content filtering were:

- **Low cost to evasion:** Spammers can easily alter features of an email's content.
- **Customized emails are easy to generate:** Content-based filters need fuzzy hashes over content, etc.
- **High cost to filter maintainers:** Filters must be continually updated as content-changing techniques become more sophisticated.

3.3 Learning Based Methods

An actively developed intellectual method based on few data mining algorithms for e-mail filtration divide the object to some categories using classification model previously defined on the base precedential information. Assume spam filtration is defined by the function

$$f(m,\zeta) = \begin{cases} m_{spam}, & \text{if the message } m \text{ is considered as spam} \\ m_{leg}, & \text{if the message } m \text{ is considered as legitimate mail} \end{cases}$$

Where m is a classified mail, m ζ is a vector of parameters m_{spam} and m_{leg} is spam and legitimate e-mail respectively. Many spam filters based on classification using machine learning techniques. In learning-based methods the vector of parameters ζ is a result of classification trainings on previously collected e-mails.

$$\zeta = Z(M).$$
$$M = \{(m_1, y_1), (m_2, y_2), \cdots, (m_n, y_n)\},$$
$$y_i \in \{m_{spam}, m_{leg}\},$$

Where m1, m2… mn are previously collected messages, y1, y2 are the corresponding labels and, Z is the training function. The following types are belonged to learning-based methods.

1) ***Image-based spam filtering.*** Spammers embed the message into the image and then attach it to the mail. Some traditional methods based on analysis of text-based information do not work in this case. Image filtering process is costly and time-consuming work. In the paper [9] it is proposed three-layer (Mail Header Classifier, the Image Header Classifier and the Visual Feature Classifier) image-spam filtering. In the First layer it is applied Bayesian classifier and SVM classifier in the remaining layers.
2) ***Bag of words Model***. The bag-of-words model is a simplifying assumption used in natural language processing and information retrieval. In this model, a text (such as a sentence or a document) is represented as an unordered collection of words, disregarding grammar and even word order [1, 6]. In spam filtering two bags of words are considered. One bag is filled with word found in spam e-mails, and the other bag is filled with words met in legitimate e-mails. Considering e-mail as a pile of words from one of these bags, there used Bayesian probability to determine to which bag this e-mail belongs. *K-Nearest* neighbor, SVM, boosting classifiers may also be applicable to the bag of words.
3) ***Collaborative spam filtering:*** This is gathering spam reports between P2P users or from mail server [1, 4, 5]. The collaborative centralized spam filtration is more economic in comparison with personal approach, but only under condition of presence of adequate procedures of the analysis of false operations and operative reclassification of not correctly classified messages.
4) ***Social networking against spam:*** This is a one of the latest methods where the information extracted from social networks is used to fight spammers [1, 5]. So in case of learning-based methods user defines the classification model himself, so that the majority disadvantages of traditional methods are solved successfully; intellectual methods are autonomous, independent on external knowledge base, doesn't require regular update, multilingual, independent of natural language, able to study new types of spam user-aided. There is advantage as construction of personalized mail classification model, where user himself defines which mail is legal or which one is a spam. Therefore learning-based methods have higher rank in spam determination. In many spam filtration systems based on the learning-based methods the Baye's theorem, Marcov's chain and others are successfully applied.

3.4 Smart Spam Filtering Method

One of the latest approaches in spam filtering is hybrid filtration system which is a combination of different algorithms, especially if they use some of the unrelated features to produce a solution. In this case it can be applied various filtering techniques and get the advantages of the traditional and learning-based methods.

4. Support Vector Machines

Support Vector Machines [7, 8] has been recently proposed by Dr. V. Vapnik as an effective statistical learning method for pattern recognition. SVM based on statistical learning theory has proved many advantages and is different from previous nonparametric techniques such as nearest-neighbors. It operates on induction principle called "*Structural risk minimization"*, which can overcome the problem of over fitting and local minimum and gain better generalization capability. Kernel function when applied, doesn't increase the computational complexity, furthermore overcomes the curse of dimensionality problem effectively. SVM has demonstrated *higher generalization capabilities [8]* in high dimensional space and spare samples. Its essence is to map optimal separating hyper plane that can correctly classify all samples. SVM has proved to be one of the most efficient kernel methods. Unlike many learning algorithms, SVM leads to good performances without the need to incorporate prior information. Moreover, the use of positive definite kernel in the SVM can be interpreted as an embedding of the input space into a high dimensional feature space where the classification is carried out without using explicitly this feature space. Hence, the problem of choosing architecture for a neural network application is replaced by the problem of choosing a suitable kernel for a SVM. Many SVM uses kernel functions but we are to incorporate some efficient and time-saving approaches to stop the spam traffic. It's a great deal of implementing such kernel functions. One such way is through classification and the building the incoming data in the SVM light format directly which will result in time – saving at the end side, thus optimizing the performance.

4.1 Classification through SVM

The idea of SVM classification [15] is the same as that of the preceptron: find a linear separation boundary wT x + b = 0 that correctly classifies training samples (such that a boundary exists). The difference from the perceptron here is that we don't search for any separating hyper plane, but for a very special maximal margin separating hyper plane, for which the distance to the closest training sample is maximal. Definition: Let X = {(xi, ci)}, xi 2 Rm, ci 2 {−1, +1} denote as usually the set of training samples. Suppose (w, b) is a separating hyper plane (i.e. sign(wT xi+b) = ci for all i). Define the margin mi of a training sample (xi, ci) with respect to the separating hyper plane as the distance from point xi to the hyper plane: mi = |wT xi + b| kwk. The margin m of the separating hyper plane with respect to the whole training set X is the smallest margin of an instance in the training set: m = min I min. Finally, the maximal margin separating hyper plane for a training set X is the separating hyper plane having the maximal margin with respect to the training set.

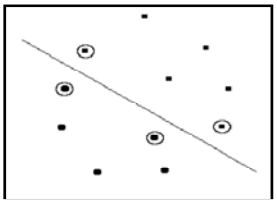


Figure 3: Maximal margin separating hyper plane

The above figure 3 shows the maximal margin separating the hyper plane. Here the circle represents the support vectors. Now here since the hyper plane given by parameters (x, b) is the same as the hyper plane given by parameters (kx, kb), we can safely bound our search by only considering canonical hyper planes for which min I |wT xi + b| = 1. It is possible to show that the optimal canonical hyper plane has minimal kwk, and that in order to find a canonical hype rplane it suffices to solve the following minimization problem: minimize ½ wTw under the conditions.

$$c_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \quad i = 1, 2, \ldots, n$$

Using the Lagrangian theory the problem may be transformed to a certain dual form: maximize

$$L_d(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j c_i c_j \mathbf{x}_i^T \mathbf{x}_j$$

with respect to the dual variables $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ so that $\alpha_i \geq 0$ for all I and $\sum_{i=1}^{n} \alpha_i c_i = 0$.

5. Applying Clustering using SVM

Further on we will designate by classification the process of supervised learning on labeled data and we will designate by clustering the process of unsupervised learning on unlabeled data. Dr. Vapnik [VAP 01] presents an alteration of the classical algorithm which is used for unlabeled training data. Here finding the hyper-plane becomes finding a maximum dimensional sphere of minimum cost that groups the most resembling data. In some of the approaches, we can find a different clustering algorithm based mostly on probabilities. For document clustering we will use the terms defined above and we will mention the necessary changes for running the algorithm on unlabeled data. There are more types of usually used kernels that can be used in the decision function.

We use the kernel function to transpose the training data from the given input sequence to a higher dimensional feature space and also to separate this data in the new obtained state. The basic idea is to calculate the norm of the differences between the two vectors. The most frequently used are the linear kernel, the polynomial kernel, the Gaussian kernel and the sigmoid kernel. We can choose the kernel according to the type of data that we are using. The linear and the polynomial kernel run best when the data is well separated. The Gaussian and the sigmoid kernel work best when data is overlapped but the number of support vectors also increases. For clustering the training data will be mapped in a higher dimensional feature space using the Gaussian kernel. In this space we shall try to find the smallest sphere that includes the image of the mapped data. This is possible as data is generated by a given distribution and when they are mapped in a higher dimensional feature space they will group in a cluster. After computing the dimensions of the sphere this will be remapped in the original space. The boundary of the sphere will be transformed in one or more boundaries that will contain the classified data. The resulting boundaries can be considered as margins of the clusters in the input space. Points belonging to the same cluster will have the same boundary. As the width parameter of the Gaussian kernel decreases the number of unconnected boundaries increases. When the width parameters increases there will be overlapping clusters. We may make use of some of the Gaussian Kernels for the effective, efficient algorithms that maximizes the cluster classifications throughput while using a limited number of resources. Also there are various clustering methods that can be applied over a number of domains, each having its own pros and cons. The use of the best techniques shall be applicable.

5.1 Use of SVM in Clustering Problems

- SVM for Binary Classification
- Multiclass Classification
- Clustering & Sequential Minimal Optimization
- Probabilistic Outputs for SVM

5.2 Advantages of Clustering using SVM

- High performance & Large capacity
- High availability
- Incremental growth.
- Improved efficiency

5.3 Applications of Clustering

- Scientific computing & Making movies
- Commercial servers (web/database/etc)

6. Evaluation Metrics: Spam Classification & Filtering

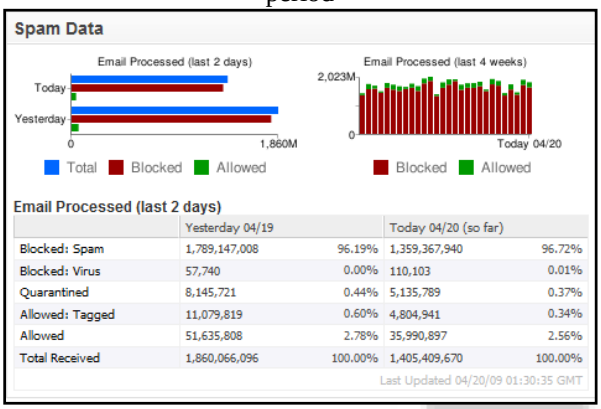
In [1] the author has seven freeware anti-spam software products for testing. Each software was being installed on Windows XP platform on different personal computers with POP3 mail server. The following table shows the summary of result which was obtained using the real post traffic.

Table 1: Testing different spam filtering software's

Anti-spam software	FPP	FNP	FP	Detected		FN
				TP	TN	
Matador 1.0.0	4.9%	4.7%	35	410	256	256
SpamBrave	1.1%	7.0%	8	400	283	283
SpamArrow	2.5%	5.8%	18	405	273	273
Espresso 1.06.94	2.9%	20.9%	21	340	270	270
Spam Bully	2.1%	16.3%	15	360	276	276
Spam Fighter	1.9%	19.1%	14	348	277	277
Qurb 2.0	5.8%	2.3%	42	420	249	249

TH-True Negatives, TP-True Positives, FN-False Negatives, FP-False Positives.

Table 2: Survey of Spam Data collected in a specific time period



Generally In the field of text classification using the SVM tool, the performance evaluation on spam filtering makes use of related indexes. The followings are definitions about two common indexes: Recall Ratio & Precision Ratio of information retrieval in spam filtering [6]. Recall Ratio is the ratio of the amount of spam that has been filtered to the amount of E-mails that should be filtered. Precision Ratio is the ratio of the amount of spam that has been filtered to the amount of E-mails that have been filtered.

7. SVMLight TOOL

SVMLight is an implementation of Support Vector Machine in C Language.

7.1 Training Step

The following syntax is used for training.

svm-learn [-option] train_file model_file

where,
train_file contains training data. The filename of train_file can be any filename. The extension of train_file can be defined by user arbitrarily.
model_file contains the model built based on training data by SVM.

7.2 Format of input files (Training data)

Training data is a collection of documents. Each line represents a document. Each feature represents a term (word) in the document. The label and each of the feature value pairs are separated by a space character. Feature value pairs must be ordered by increasing feature number value e.g., tf-idf.

7.3 Testing Step

The following syntax is used for testing.

svm-classify test_file model_file predictions

The format of test_file is exactly the same as train_file. It needs to be scaled into same range. We use the model_file based on training data to classify test data, and compare the predictions with the original label of each test document.

7.4 Evaluations of Performance

- Accuracy (AC) is the proportion of the total number of predictions that were correct.
  AC = (a + d) / (a + b + c + d)
- Recall is the proportion of positive cases that were correctly identified.
  R = d / (c + d) {Actual positive case no}
- Precision is the proportion of the predicted positive cases that were correct.
  P = d / (b + d) {Predicted positive case no}

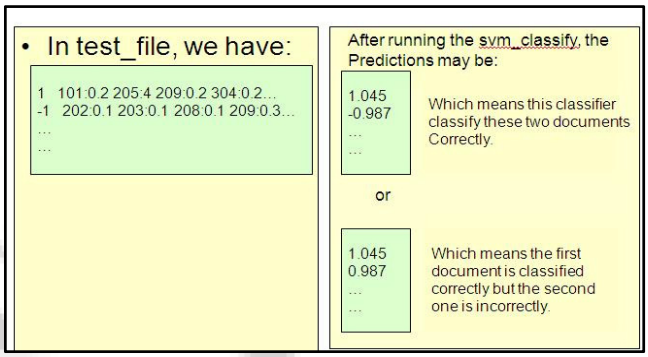


Figure 4: An Example using SVMLight

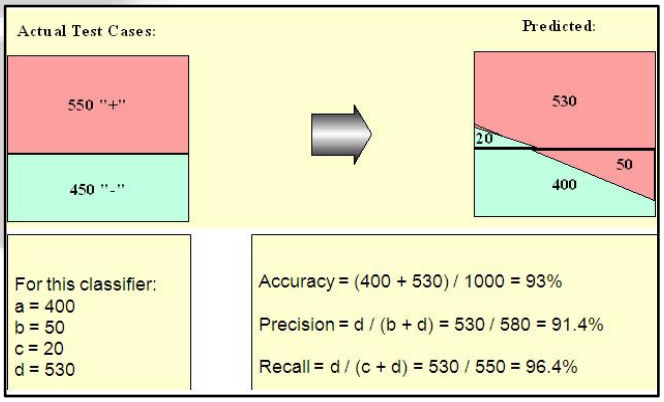


Figure 5: Actual & Predicted Test cases

8. Implementation

The primary concern will be the use of compression models in spam classification/filtering. Collections of some data sets in date warehouse and then perform data mining on those sets of data them by training and supervised learning method. Training the data involves a series of inputs sequences and the desired outputs which will eliminate the spam from the desired messages. In the second step, learning a function from training data set available to the user and the objective is to predict output from the function which is based upon any valid input after having seen a number of training examples. The steps followed here are as follows:

1. Acquisition of data and Dataset description
   a. Preparations of the required datasets.
   b. Training the datasets.
2. The Spam based dataset(has 541 instances)
3. Each instance in the file relates to a separate email.
4. Each email is represented using some real attributes like integers and subsets selectors.
5. Generating the svm_light format directly based on the above features (using the SVM-Light tool)

6. 500 instances were used for training and 41 were used for testing.
7. Training the classifiers simultaneously.
8. Testing the Classifiers based on the above trained classifiers.

9. Conclusion & Future Enhancements

This problem differs from classical text categorization tasks in many ways. Cost of misclassification is highly unbalanced. Messages in an email stream arrive in order and must be classified upon delivery. It seems to be very common to deploy a filter without training data. Here we try to implement a new kernel function that would rather helps in achieving efficiency and solving many clustering problems. We also try to implement the kernels function that would even help in spam detection and classification/filtering upon detection at the origins. These unique characteristics of the spam classification and filtering and clustering approaches problem are reflected in the design of our experiments and the choice of measures that were used for classifier evaluation. Here after our analysis about the overall scenario of spam classification/filtering, we conclude that various classifiers like “Decision Tree” & “Graphs Models” classifiers have large memory requirement. Also the number of features for spam classification/filtering is havoc & large and hence in order to evaluate these attributes the use SVM has proved to be a good classifier because of its sparse data format and acceptable recall and precision value. SVM is regarded as an important application of “kernel methods implementation”, one of the key areas in machine learning. In Future, we have planned to propose and implement some more new algorithms using *kernel* functions that will enhance the efficiency and performance over any of the available datasets at real-time and removing e-mail spam significantly.

10. Acknowledgement

I would like to thank Shri. Gaurav Shrivastav, for his inspiring discussions, guidance and mutual encouragement provided to me at all times!!!!

References

[1] Saadat Nazirova, Institute of Information Technology of Azerbaijan National Academy of Sciences, Azerbaijan. Received April 11, 2011; revised May 8, 2011; accepted May 15, 2011
[2] R.Deepa Lakshmi and N. Radha, “Supervised Learning Approach for Spam Classification Analysis using Data Mining Tools” International Journal on Computer Science and Engineering, Vol. 02, No. 08, 2010, 2760-2766
[3] Brian Krebs, Krebs on Security, http://krebsonsecurity.com/2013/01/spam-volumes-past-present-global-local/
[4] Microsoft Sender ID Framework. http://www.microsoft.com/mscorp/safety/technologies/senderid/default.mspx.
[5] Sender Policy Framework. http://www.openspf.org/Introduction.
[6] Bag of Words Model. http://en.wikipedia.org/wiki/Bag_of_words_model_in_computer_vision.
[7] N. Cristianini and J. S. Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, 2000.
[8] V. Vapnik. , The Nature of Statistical Learning Theory, Springer, New York, 1995.
[9] T.-J. Liu, W.-L. Tsao and C.-L. Lee, “A High Performance Image-Spam Filtering System,” Ninth International Symposium on Distributed Computing and Applications to Business, Engineering and Science, 10-12 August 2010, Hong Kong, pp. 445-449. doi:10.1109/DCABES.2010.97
[10] M. Sasaki & H. Shinnou, “Spam Detection Using Text Clustering,”IEEE Proceedings of 2005 International Conference on Cyberwords, Singapore, 23-25.
[11] Duncan Cook, Catching Spam before it arrives: Domain Specific Dynamic Blacklists, Australian Computer Society, 2006, ACM.
[12] D. Puniškis, R. Laurutis, R. Dirmeikis, An Artificial Neural Nets for Spam e-mail Recognition, electronics and electrical engineering ISSN 1392 – 1215 2006. Nr. 5(69)
[13] Youn and Dennis McLeod, “ A Comparative Study for Email Classification, Seongwook Los Angeles” , CA 90089, USA, 2006.
[14] Bekker, S 2003, Spam to Cost U.S. Companies $10 Billion in 2003, ENT News, viewed May 11 2005, http://www.entmag.com/news/article.asp?EditorialsID=5651>.
[15] SVM Application List.
[16] http://www.clopinet.com/isabelle/Projects/SVM/applist.html
[17] Wikipedia, “Spam”. http://en.wikipedia.org/wiki/Spam_(electroni).
[18] Wikipedia, “E-mail spam”. http://en.wikipedia.org/wiki/E-mail_spam.
[19] [VAP01] Vladimir Vapnik, Asa Ben-Hur, David Horn, Hava T, Sieglmann, Support Vector Clustering, Journal of Machine Learning Research 2, pages125-137,2001.