Filtering Template Driven Spam Mails using Vector Space Models

Liny Varghese
Cochin University of Science and Technology, Cochin, India

Supriya M.H
Cochin University of Science and Technology, Cochin, India

K. Poulose Jacob
Cochin University of Science and Technology, Cochin, India

ABSTRACT

Spam became a big problem to the society. Some spammers are using templates for sending spam. To send a particular promotion they create some template and merge the details of receivers with the template. Similarities can find among these mails and easily ignore the forthcoming spam. Most high-volume spam is sent using tools those randomizes parts of the message - subject, body, sender address etc. The general form of the template that the spammer is using can often guess by inspecting the features of messages. Most of the spam filters are either rule based models or Bayesian models. The main objective in this paper is to find out semantic distance and evaluate the applicability of the two information retrieval techniques, Simple Vector Space Models (VSM) and VSM using Rocchio Classification in the spam context. Both methods are using cosine similarities to identify the spam

Keywords

Spam, vector space models, Rocchio classification, cosine similarity

1. INTRODUCTION

Unsolicited bulk or commercial e-mail ('spam') has become a severe problem on the Internet over the last years. Although many strategies for addressing this problem have been proposed, we are still away from a satisfactory and lasting solution. This is due to the fact that many of the methods proposed and developed have some heuristics applied and pertain to those corpora only. So those methods are not useful after a period. And spammers who make profit based on the spam easily find ways to overcome these solutions or finding new methods to send of spams.

Spam e-mail messages tend to have several elements or features in common, which are usually not present in legitimate e-mail. However, these features are hard to locate comprehensively, because they are not fully known, not always explicit and it is always dynamic. There is no exact method to unambiguously charecterise spam.

Some spammers are using templates for sending spam. To send a particular promotion they create some template and merge the details of receivers with the template. Similarities can find among these mails and easily ignore the forthcoming spam. Most high-volume spam is sent using tools that randomizes parts of the message - subject, body, sender address etc. The general form of the template that the spammer is using can often be guessed by inspecting the features of messages.

2. APPROACH

The problem of filtering spam is a binary classification problem in the sense that every incoming e-mail has to be classified as either 'spam' or 'not spam'. The main objective of this paper is to investigate and evaluate the applicability of the two information retrieval techniques, Simple Vector Space Models (VSM) using cosine similarities and VSM using Rocchio Classification in the context of spam classification.

Classification Methodology

The standard VSM using cosine similarity and Euclidian distance are used in the study. Simple VSM and VSM using Rocchio Classification methods and their adaptation and application to the task of spam filtering are discussed.

The Simple Vector Space Model

Vector Space Model is an algebraic model for representing text documents as vectors of terms. In information retrieval, a vector space model (VSM) [1] is a widely used model for representing information. Documents and queries are represented as points in a potentially very high dimensional, metric vector space. The distance (or similarity) between a query vector and the document vectors is the basis for the information retrieval process.

Documents and queries are represented as vectors.

$$Dj = \{w_{1,j}, w_{2,j}, w_{3,j} ..., w_{t,j}\}$$

$$q = \{w_{1,q}, w_{2,q}, w_{3,q} ..., w_{t,q}\}$$

Each dimension corresponds to a separate term. If a term occurs in the document, its value in the vector is non-zero. We can use term frequency or tf-idf weight as the value in the vector. The terms are distinct words in the vocabulary/corpus and the dimensionality of the vector is the number of distinct words in the vocabulary/corpus.

Vector operations can be used to compute the distance between documents and queries by comparing the deviation of angles between each document vector and the query vector where the query is represented as same kind of vector as the documents. In practice, we calculate the cosine of the angle between the vectors; instead of the angle itself. The documents are similar if the angle has small value.

$$\cos\phi = \frac{D_2 \bullet q}{\|D_2\| \times \|q\|}$$

Where $D_2 \bullet q$ is the intersection (i..e dot product) of the document and the query vectors, $\|D_2\|$ is the norm of vector D_2, and $\|q\|$ is the norm of vector q. The norm of a vector is calculated as such:

$$\|v\| = \sqrt{\sum_{i=1}^{n} v_i^2}$$

$$similarity = \cos(\phi) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

A cosine value of zero means that the query and document vector are orthogonal and have no match and one means they are identical.

Rocchio Classification

In basic vector space model discussed above, we compute the cosine similarity of new incoming mail with each training mails and assign the class of mail with maximum cos(θ). This is the decision boundary, which is chosen to separate the two classes. Another way to determine the decision boundary is Rocchio Classification [2, 3]. This method uses the centroids of each class to determine the boundaries. The centroid of a class is computed as the vector average or center of mass of its members.

$$\vec{\mu}(c) = \frac{1}{|D_c|} \sum_{d \in D_c} \vec{v}(d)$$

where $|D_c|$ is the set of documents in D whose class is c: D_c = {d: (d, c) ε D}. The normalized vector of d is denoted by $\vec{v}(d)$.

The boundary between two classes in Rocchio Classification is the set of points with equal distance from the two centroids and the new mail is classified into class with closest centroid μ(c) from the new mail.

3. METHODOLOGY

The working of the method is: Consider emails as documents and words as terms, tokenize the mails and store all the tokens in the feature vector , for finding the exact template , stemming and stop word removals are not done , assign each training mail into two given classes (spam or ham), find Document Frequency, df, find term frequency-inverse document frequency, tf-idf.

$$\omega_{t,d} = tf_{t,d} \cdot \log \frac{|D|}{\left|\left\{d^1 \in D \middle| t \in d^1\right\}\right|}$$

Where | D | is the total number of documents in the document set $\left|\left\{d^1 \in D \middle| t \in d^1\right\}\right|$ is the number of documents containing the term t. $tf_{t,d}$ is term frequency of term t in document d (a local parameter), normalize feature vector by dividing with Euclidian length to make it unit vector and convert each email in the test corpora into unit vector and store into query vector

For Simple VSM

Find the cosine similarities between each training mail vectors and the query vector. Select the training email vector with maximum cosine value; assign the spam class of the selected training email vector to the query vector.

For VSM using Rocchio Classification

Find the centroid (μ(c), c={spam, legitimate}) of each class by applying Rocchio Classification . Then calculate the cosine similarity of test mail from the centroids cos (q, μ(c)) where c= {spam, legitimate} and then assign test mail to class with Max (cos(q, μ(c))).

4. COMPOSITION OF TRAINING SET AND TEST SET

Table 1: Data set Composition

Dataset	No. of spam Mails	No. of Legitimate mails
Training set	42	59
Test set	42	59

The above dataset is prepared using the mails received in 2 days for the testing. Large numbers of spams are received every day, if the server is not capable to handle spam. The testing is done using only this miniature dataset. Large datasets are available online, but when go for large datasets, the computational time increases and this will delay the mail delivery. Also template based spams are time dependent, earlier templates may not helpful to detect spam. Only unique templates are included in the training set. For finding the uniqueness cosine similarity is used.

5. CLASSIFICATION RESULTS

The following Performance and correctness measures are considered while evaluating the experiment results.

- Sensitivity or true positive rate TPR (recall) = TP / P = TP / (TP + FN)
- Specificity= (SPC) or True Negative Rate = TN / (TN+FP) = 1 − FPR
- Positive predictive value (PPV) = precision = TP / (TP + FP)
- F-measure F = 2 x (Precision x Recall)/ (Precision + Recall)

The experiment results on the given dataset are given below:

Table 2: Experiment results

Simple VSM			
Confusion Matrix		Predicted	
		Spam	Legitimate
Actual	Spam	28	14
	Legitimate	8	51
TP=28		FN=14	
FP=8		TN=51	

VSM using Rocchio Classification			
Confusion Matrix		Predicted	
		Spam	Legitimate
Actual	Spam	37	5
	Legitimate	14	45
TP=37		FN=5	
FP=14		TN=45	

Performance and correctness measures are given in the following table

Table 3: Performance and correctness measures

Performance and correctness measures	Simple VSM	VSM using Rocchio Classification
Sensitivity(Recall)	66.66%	88%
Specificity	86.44%	76.27%
Positive predictive value (precision)	77.78%	72.54%
F-measure(F)	71.78%	79.38%

6. CONCLUSION

In this paper we considered the problem of spam filtering. In literature most of the spam filters are either rule based models or Bayesian models. This paper considered another idea focused on two schemes based on vector space models followed in classic Information Retrieval. To find semantic distance, cosine similarity is used in both methods. This study has been carried out on 101 real datasets with attributes of td-idf values. First method used all the mails in the training set to test against the spam, while in the second method, only the centroids of each class (only two vectors) are used to find the similarity. VSM using Rocchio Classification is much faster than simple VSM because the number of iterations required is less. The results showing that VSM using Rocchio Classification scheme performs better than Simple VSM scheme. Since templates are changing with time and promotional activities, the training data need to be changed periodically in order to incorporate new templates. The simple VSM model is efficient to find out the exact spam template. But when the test training set becomes large, time to find similarity is also increasing (O (n)). Hence we have to update the training corpus by deleting the templates that are not used by spammers and by adding new mail templates. The training data size can be further reduced by storing only unique mail templates. In that way simple VSM can performs better than Rocchio Classification. The optimum size of the training set has to be studied. The method presented here can be enhanced to find semantic distance between mails.

7. REFERENCES

[1] G. Salton, A. Wong, and C. S. Yang, "A Vector Space Model for Automatic Indexing," Communications of the ACM, vol. 18, nr. 11, pages 613–620(1975).

[2] J. J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, The SMART Retrieval System: Experiments in Automatic Document Processing, Prentice-Hall Series in Automatic Computation, chapter 14, pages 313â€“323. Prentice-Hall, Englewood Cliffs NJ, 1971.

[3] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze: An Introduction to Information Retrieval, page 181. Cambridge University Press, 2009.

[4] Wilfried N. Gansterer_ Andreas G. K. Janecek Robert Neumayer, Spam Filtering Based on Latent Semantic Indexing

[5] http://en.wikipedia.org/wiki/Vector_space_model viewed on January 2012

[6] Tuomo Korenius, Jorma Laurikkala, Martti Juhola, On principal component analysis, cosine and Euclidean measures in information retrieval, Information Sciences, Volume 177, Issue 22, 15 November 2007, Pages 4893-4905, ISSN 0020-0255

[7] Congnan Luo, Yanjun Li, Soon M. Chung, Text document clustering based on neighbors, Data & Knowledge Engineering, Volume 68, Issue 11, November 2009, Pages 1271-1288, ISSN 0169-023X, 10.1016/j.datak.2009.06.007.

[8] Angel R. Martinez, Data Mining of Text Files, In: C.R. Rao, E.J. Wegman and J.L. Solka, Editor(s), Handbook of Statistics, Elsevier, 2005, Volume 24, Pages 109-131, ISSN 0169-7161, ISBN 9780444511416, 10.1016/S0169-7161(04)24004-4.

[9] Thamarai Subramaniam, Hamid A. Jalab and Alaa Y. Taqa, Overview of textual anti-spam filtering techniques, International Journal of the Physical Sciences Vol. 5(12), pp. 1869-1882, 4 October, 2010