CAS-ICT at TREC 2005 SPAM Track: Using Non-Textual Information to Improve Spam Filtering Performance

Shen Wang, Bin Wang and Hao Lang, Xueqi Cheng

Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
{wangshen,langhao}@software.ict.ac.cn,
{wangbin,cxq}@ict.ac.cn

Abstract: This paper introduces our work in the TREC2005 SPAM track. Naïve Bayes and Littlestone's Winnow are chosen as our basic classifiers. In our investigation, we found that when the structures of Ham and Spam are very different, the feature distributions of them vary a lot. Thus the factor of structure is introduced into our filter. Besides textual word feature, some kind of other features are also considered in our filter. Our experimental results show that Winnow outperforms Naïve Bayes and the multi-feature model outperforms structure based model.

1 Introduction

This is the first year that TREC introduces the SPAM track. And the task is to develop an automatic spam filter to classify a chronological sequence of email messages as SPAM or HAM (non-spam). The subject filter is run on several email collections, some of them are public and some are private. The performance of the filter is measured and compared to *go*ld standard judgments made by human assessors.

Our experiments (four runs) can be divided into two distinct parts, two of them are based on Naïve Bayes classifiers, and the other two based on Winnow algorithm. For each part, we applied two strategies separately -- SBF model(structure based 2-layers filtering model),and Multi-Feature model.

The goals of our investigation in SPAM track 2005 include:

- To compare the performance of winnow with Naïve Bayes
- To evaluate the performance of SBF model(structure based 2-layers filtering model)
- To check the effect of Multi-Feature

The rest of this paper is organized as follows. Section 2 briefly describes the main and common techniques we used in all the four runs we submitted; Section 3 describes the multi-feature filtering model, which we used to supplement the information contained

in email text. Section 4 introduces the SBF model(structure based 2-layers filtering model), which we used to improve the robustness of the system among different corpus with different structures. Finally, section 5 lists and then discusses the results of our system on the public trec05p-1 corpus.

2 General Filter Description

In preprocessing phase, we adopted the IG(Information Gain) method to select the features and reduce the feature dimension. And we used Naïve Bayes and Winnow [1] algorithm as our basic classification methods.

2.1 Feature Dimension Reduction

The original feature space transformed with the vector space model may contain tens of thousands of different features, and not all classifiers can handle such a high dimension gracefully. Dimension reduction (also called feature pruning or feature selection) is usually employed to reduce the size of the feature space to an acceptable level, typically several orders of magnitude smaller than the original one. The benefit of dimension reduction also includes a small improvement in prediction accuracy in some cases. In SPAM track, we used the Information Gain method, an outstanding feature selection algorithm, which is defined as:

$$IG(x) = \sum_{t=\{0,1\},\ c=\{c_1,c_2\}} \sum P(x=t,c) \log \frac{P(x=t,c)}{P(x=t)P(c)} \qquad (1)$$

2.2 Classification Method

2.2.1 Naïve Bayes

Naive Bayes (NB) is a widely used classifier in text categorization task. It also enjoys a blaze of popularity in anti-spam researches [3][4][5], and often serves as baseline method for comparison with other approaches.

2.2.2 Winnow

Winnow is a fast linear classifier. The training of Winnow is online and mistake driven. Furthermore, Winnow is suitable for feedback. The Winnow algorithm was proved to be effective to filter spam in PU1 and Ling-Spam e-mail collections [6].

3 Multi-Feature Spam Filtering

Compared to the common objects in text-mining problems, e-mail has its special

information such as sender’s address, sending time, etc. Thus, spam filtering is not only a text categorization task. We wonder whether these non-textual features can improve the whole filtering performance and the model that contains textual and non-textual features is called multi-feature model. First, we categorized all the features that can be used in content-based spam filtering into textual and non-textual features. And then, besides textual features such as words, which have been widely used in text categorization problem or spam filtering problem, some non-textual features are also introduced in our spam filter. Most of these non-textual features are attributes of the whole e-mail message, e.g., the number of “$” in the email body text. Some non-textual features we used in our experiments are listed in table 1.

Table 1 Non-textual features used in our experiments

No.	Feature description	Feature Name
1	The mail sending time (hour)	DateHour
2	MIME content type	ContentType
3	Symbol proportion in Subject	HSymbolProb
4	Symbol proportion in body	BSymbolProb
5	Message length	MailLen
6	The number of '!' in body	HExclamNum
7	The number of '!' in Subject	HDollarNum
8	The number of '$' in body	HUpperWord
9	The number of '$' in Subject	HSingleAlpWord
10	Proportion of upper words in body	BExclamNum
11	Proportion of upper words in Subject	BDollarNum
12	Proportion of single letter words in body	BUpperWord
13	Proportion of single letter words in subject	BSingleAlpWord
14	The number of urls in body	BUrlNum
15	The sender’s domain	FromDomain
16	If the message is a reply one?	Re
17	If the message have an attachment	Attachment
18	The number of relay	ReSceivedNum

Our runs ICTSPAM1WNB and ICTSPAM4NBB used the multi-feature model described above.

4 Bi-Layer Spam Filtering

There are lots of differences between the problem of common data mining and spam filtering. When the structures are very different between two email corpus, the feature distributions vary a lot. And the diversity of the feature distributions has affect on the performance of machine learning algorithm. We analyzed the problem mentioned above, and designed a structure based 2-layers filtering model, which uses different machine learning filter to train and classify mail of different structure, shown in Figure1. Experiments show that machine learning algorithm's performance was improved a lot after using this model.

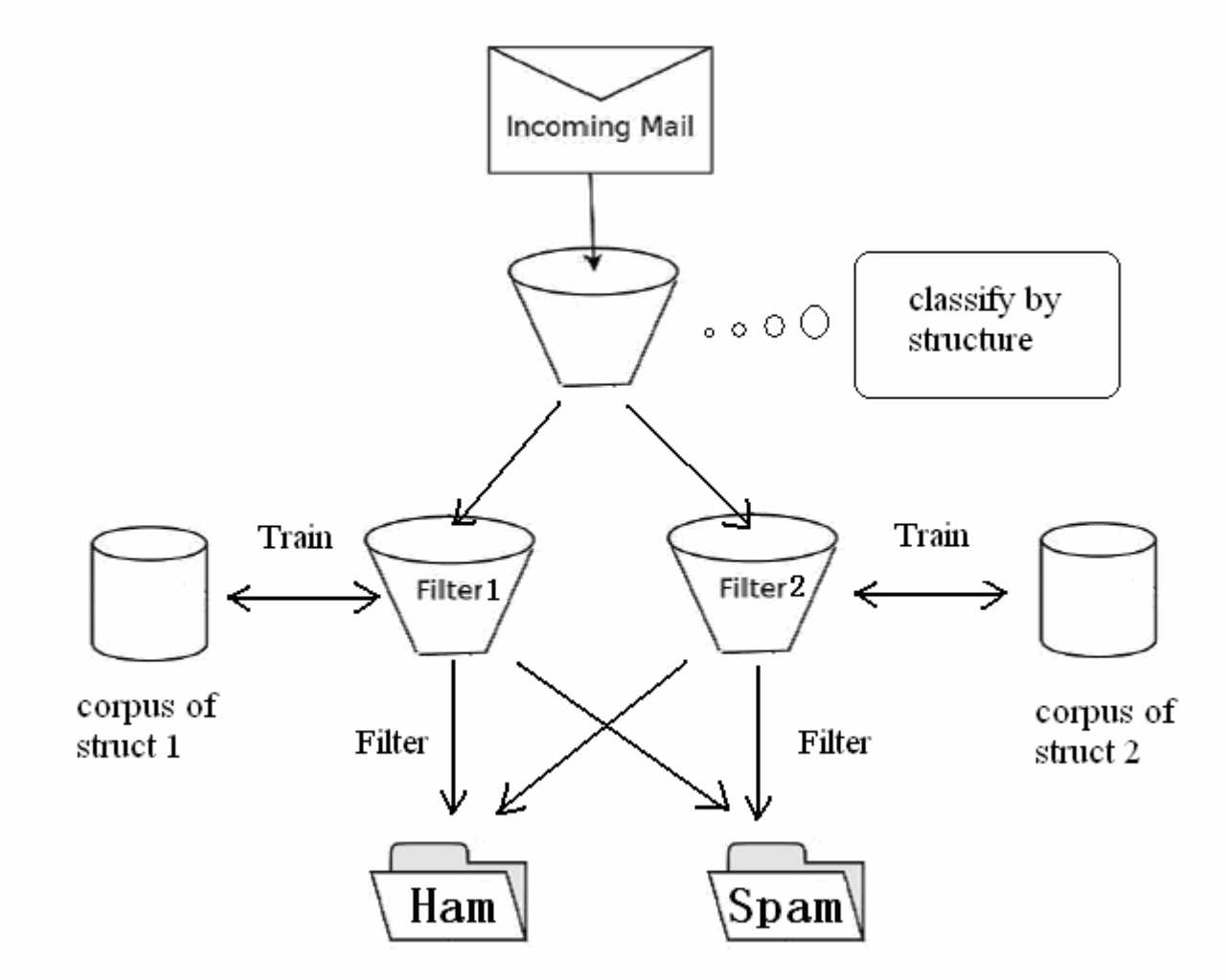

Figure 1 Architecture of Structure-based Bi-layer spam Filtering model(SBF)

Our runs ICTSPAM2WNH and ICTSPAM3NBH use the Bi-layer Spam Filtering Model describes above.

4 Performance in the SPAM Track

We have submit four runs, describe by table 2.

Table 2 Submitted runs

Runs	**Algorithm**	**Model**
ICTSPAM1WNB	Winnow	Multi-feature model
ICTSPAM2WNH	Winnow	SBF
ICTSPAM3NBH	Naïve bayes	SBF
ICTSPAM4NBB	Naïve bayes	Multi-feature model

The public corpus trec05p-1 have five indexes: ham25, ham 50, spam 25, spam 50, full. Our system gets similar results on these indexes. And we only display the results on the ham25 index in table 3.

Table 3 Performance on trec05p-1/ham25

	Ham miss%	Spam miss%	Misc%	1-ROCA%
ICTSPAM1WNB	15 (14.44-15.87)	3.74 (3.58-3.90)	5.52 (5.34-5.70)	4.01213 (3.80041 - 4.23514)
ICTSPAM2WNH	11.32 (10.70-11.97)	14.29 (13.99-14.59)	13.83 (13.56-14.10)	6.1402 (5.86185 - 6.43087)
ICTSPAM3NBH	13.68 (13.00-14.38)	26.99 (26.61-27.37)	24.91 (24.57-25.25)	19.9473 (19.5376 - 20.3635)
ICTSPAM4NBB	19.51 (18.72-20.31)	9.65 (9.40-9.91)	11.19 (10.94-11.44)	10.821 (10.3873 - 11.2705)

From the results, we can see, the winnow algorithm is superior to Naïve Bayes. And the multi-feature model performed much better than SBF.

Refences

[1] N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold

algorithm. Machine Learning, 2(4):285-318, 1988.

[2] YANG, Y. AND PEDERSEN, J. O. 1997. A comparative study on feature selection in text categorization. In *Proceedings of ICML-97, 14th International Conference on Machine Learning* (Nashville, TN),D. H. Fisher, Ed. Morgan Kaufmann San Francisco, CA, 412–420.

[3] M. Sahami, S. Dumais, D. Heckerman and E. Horvitz, “A Bayesian approach to filtering junk e-mail”, in *Proc. of AAAI Workshop on Learning for Text Categorization*, pp. 55-62, 1998

[4] I. Androutsopoulos, J. Koutsias, K.V. Chandrinos and C.D. Spyropoulos, “An Experimental Comparison of Naive Bayesian and Keyword-Based Anti-Spam Filtering with Encrypted Personal E-mail Messages”, in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000)*, Athens, Greece, pp. 160-167, 2000

[5] K. Schneider, "A Comparison of Event Models for Naive Bayes Anti-Spam E-Mail Filtering", in *Proc. 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)*, Budapest, Hungary, pp. 307-314, Apr. 2003

[6] Pan Wenfeng. Master’s dissertation ,“Research on Content-Based Spam Filtering”.2004