Multi-field Learning for Email Spam Filtering

Wuying Liu
College of Computer
National University of Defense Technology
410073 Changsha, Hunan, CHINA
wyliu@nudt.edu.cn

Ting Wang
College of Computer
National University of Defense Technology
410073 Changsha, Hunan, CHINA
tingwang@nudt.edu.cn

ABSTRACT
Through the investigation of email document structure, this paper proposes a multi-field learning (MFL) framework, which breaks the multi-field document Text Classification (TC) problem into several sub-document TC problems, and makes the final category prediction by weighted linear combination of several sub-document TC results. Many previous statistical TC algorithms can be easily rebuilt within the MFL framework via turning binary result to spamminess score, which is a real number and reflects the likelihood that the classified email is spam. The experimental results in the TREC spam track show that the performances of many TC algorithms can be improved within the MFL framework.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *information filtering*

General Terms
Algorithms, Experimentation, Performance

Keywords
Spam Filtering, Multi-field Learning, Text Feature Selection

1. INTRODUCTION
Currently email spam filtering is normally considered as an online binary Text Classification (TC) task, and many robust statistical TC algorithms have been proposed [1]. In these algorithms, email is often treated as a single plain-text document, and text feature is also extracted within this single document. Actually a full email (often including five natural text fields: *Header*, *From*, *ToCcBcc*, *Subject*, and *Body*) is a multi-field text document. Feature extraction from full email document makes many text features disturb each other, and text feature from one field is often noise to other fields.

In statistical TC algorithms, a document is normally represented as a text feature vector. The dimension of feature vector space, the total number of text features, reflects the representational granularity of vector space model. Previous research has shown that overlapping word-level *k*-grams model can achieve promising results [2]. For email document, single plain-text model (SPTM) and multi-field model (MFM) are two representations. The SPTM ignores the field information of text feature, regarding the same string occurrence in different fields as single text feature, while the MFM treats it as distinct text features. The dimension of feature vector space for trec07p email set is showed in Table 1. For the two email representations, four overlapping word-level models are applied respectively. For MFM, the five natural text fields' information is considered.

Table 1. Dimension of Feature Vector Space.

	1-grams	2-grams	3-grams	4-grams
SPTM	1,037,395	4,189,054	9,447,962	13,869,560
MFM	1,258,491	4,906,594	10,390,571	14,880,647

Table 1 shows the dimension of MFM is larger than that of SPTM for each *k*-grams model. For instance, this obvious difference between two representations reaches 1,011,087 for 4-grams model. The result from Table 1 indicates that text feature noises exist indeed in SPTM. Because more finely granular text feature can reduce the noises and increase the TC accuracy, this paper proposes a multi-field learning (MFL) framework, which is an alignment technique of text feature sources. In MFL framework, text features are enhanced by field information, and the disturbances among text features from different fields are expected to be reduced.

2. MULTI-FIELD LEARNING
In order to reduce the text feature noises of multi-field document, the proposed MFL framework makes use of multi-field structural feature by the divide-and-conquer strategy. Figure 1 shows the MFL framework for multi-field document binary TC. The framework includes a *Splitter*, a *Combiner*, and several *Scorer*s. The *Splitter* analyses a multi-field document, and splits it to several sub-documents according to the natural field structure or some explicit rules. The text feature extracting, the scorer training and updating, and the sub-document predicting are only localized in the sub-documents from the same field. Each scorer calculates a spamminess score (SS) for its corresponding sub-document, and sends the SS to the *Combiner*. The *Combiner* combines multi-scorer's *SS*s to form the final *SS*, which is a real number in [0, 1]. If the final *SS* is in [0, 0.5], then the document is predicted as a ham, otherwise, if the final *SS* is in (0.5, 1], it is predicted as a spam.

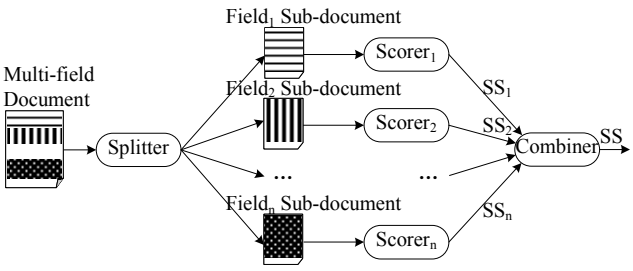

Figure 1. Multi-field Learning Framework.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

In MFL framework, the weighted linear combination method is designed for combining n scorer's output scores to form the final SS. The formula of calculating SS is $SS=\Sigma\alpha_i SS_i$, (i=1, 2, ... , n), where SS_i indicates the ith scorer's output SS, and the weight α_i indicates the historical classification ability of the ith scorer. The normalized TC accuracy rates are used to estimate the weights.

Except five natural sub-documents from the five natural text fields of email, some artificial sub-documents can be extracted by some explicit rules. For instance, the regular expression can be applied to extract all IP addresses in email *Header* to form an artificial sub-document. This artificial method can generate a new field sub-document which does not exist in actual multi-field document, which is equivalent to increasing the statistical weight for some attributed texts, and such texts often have an explicit optimal TC rule.

It is nearly a supervised online binary TC process that the scorer receives a sub-document and calculates a SS according to its TC model. Previous supervised online binary TC algorithms can be rebuilt into these scoring algorithms by changing a binary output to a continuous SS output. So, MFL framework is a general frame for ensemble previous TC algorithms.

3. EXPERIMENT

Email is a typical multi-field document, so this paper verifies the validity of MFL framework through the email spam filtering experiment of Immediate Full Feedback defined in the TREC2007 spam track [3]. The TREC spam filter evaluation toolkit and the associated evaluation methodology are applied. Experiment corpus is trec07p email set. The running hardware environment is a PC with 1GB memory and 2.80GHz Pentium D CPU.

A MFL framework of seven sub-documents for email document is implemented, in which the *Splitter* extracts five sub-documents (*Header*, *From*, *ToCcBcc*, *Subject*, and *Body*) by natural field structure and extracts two sub-documents (*H.IP*, *H.EmailBox*) by regular expressions. The *H.IP* contains IP address text and *H.EmailBox* contains Emailbox address text within email *Header*. Each scorer's historical SS outputs can be drawn to a receiver operating characteristic (ROC) curve. The percentage of the area below the ROC curve (ROCA%) indicates the historical classification ability, and the ROCA% is reasonable to estimate the classification accuracy rate of a scorer. So, before an email classified, the MFL framework normalize current seven ROCA% values to estimate the weights of scorers.

To verify that MFL framework's effect on improving the performance of previous TC algorithms, two typical online TC algorithms are run in MFL framework. The bogo filter (bogo-0.93.4) is a classical implementation of online Bayesian statistical algorithm [4], while the tftS3F filter is based on relaxed online SVMs algorithm and has gained several best results in the TREC2007 spam track [5]. We report (1-ROCA)% overall performance, where 0 is optimal. Table 2 shows the overall performance of filters affected by this paper proposed approaches in the rank reference of top three filters in the TREC2007 spam track whose font is italic. In Table 2, the (.mfl) postfix indicates running in MFL framework. The experimental results show that the bogo filter's (1-ROCA)% is optimized from original 0.1558 to mfl's 0.0103, and the tftS3F filter's (1-ROCA)% is also optimized from original 0.0093 to mfl's 0.0083.

Table 2. Overall Performance of Email Spam Filtering.

	wat3	tftS3F.mfl	*tftS3F*	bogo.mfl	*fdw4*	bogo
(1-ROCA)%	*0.0055*	0.0083	*0.0093*	0.0103	*0.0109*	0.1558
TREC Rank	*1*		*2*		*3*	

Table 2 shows that the performance of the online Bayesian and relaxed online SVMs algorithms can be improved within the MFL framework, which demonstrates the advantage of MFL framework. The improvement of MFL framework can be explained in two main reasons: (1) The MFL framework can reduce the disturbances among text features from different fields; (2) Multi-field ensemble learning has statistical, computational and representational advantages [6].

4. CONCLUSION

This paper elucidates that the structural feature of multi-field document is very useful for statistical TC algorithm. The proposed MFL framework represents more finely granular text feature with field information, and takes advantage of the structural feature. The experiment shows that MFL framework can improve the performance of many TC algorithms. Moreover, MFL framework is suitable to parallel running environment, if it is applied on the reduplicate hardware for multiple scorers, the theoretical computational time of MFL framework to classify a document is nearly equal to the lowest scorer's running time.

Further research will concern semi-supervised learning, active learning, and personal learning for spam filtering within MFL framework. We will apply large-scale unlabeled emails, select effective samples for training by mining differences among multiple scorers of MFL framework, and improve the TC model for both global and personal filtering.

5. ACKNOWLEDGMENTS

The research is supported by the National Natural Science Foundation of China (No.60873097, No.60933005) and Program for New Century Excellent Talents in University (No.NCET-06-0926). Many thanks to Dr. D. Sculley for his tftS3F filter code.

6. REFERENCES

[1] Gordon V. Cormack. Email spam filtering: a systematic review. *Foundations and Trends in Information Retrieval*, 1(4):335-455, 2008.

[2] H. Drucker, D. Wu, V. N. Vapnik. Support vector machines for spam categorization. *IEEE Transactions on Neural Networks*, 10(5):1048-1054, 1999.

[3] Gordon V. Cormack. TREC 2007 spam track overview. In *TREC2007: Proceedings of the 16th Text REtrieval Conference*, National Institute of Standards and Technology, Special Publication 500-274, 2007.

[4] Paul Graham. Better bayesian filtering. In *Proceedings of the 2003 Spam Conference*, January 2003.

[5] D. Sculley, Gabriel M. Wachman. Relaxed online SVMs for spam filtering. In *SIGIR'07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 415-422, 2007.

[6] Thomas G. Dietterich. Ensemble methods in machine learning. In *MCS2000: Proceedings of the Multiple Classifier Systems*, pages 1-15, 2000.