Feature Construction Approach for Email Categorization Based on Term Space Partition

Guyue Mi, Pengtao Zhang and Ying Tan

Abstract— **This paper proposes a novel feature construction approach based on term space partition (TSP) aiming to establish a mechanism to make terms play more sufficient and rational roles in email categorization. Dominant terms and general terms are separated by performing a vertical partition of the original term space with respect to feature selection metrics, while spam terms and ham terms are separated by a transverse partition with respect to class tendency. Strategies for constructing discriminative features, named term ratio and term density, are designed on corresponding subspaces. Motivation and principle of the TSP approach is presented in detail, as well as the implementation. Experiments are conducted on five benchmark corpora using cross-validation to evaluate the proposed TSP approach. Comprehensive experimental results suggest that the TSP approach far outperforms the traditional and most widely used feature construction approach in spam filtering, which is named bag-of-words, in both performance and efficiency. In comparison with the heuristic and state-of-the-art approaches, namely CFC and LC, the proposed TSP approach shows obvious advantage in terms of accuracy and F_1 measure, as well as high precision, which is warmly welcomed in real spam filtering. Furthermore, the TSP approach performs quite similar with CFC in efficiency of processing incoming emails, while much faster than LC. In addition, it is shown that the TSP approach cooperates well with both unsupervised and supervised feature selection metrics, which endows it with flexible capability in the real world.**

I. Introduction

SPAM, generally defined as unsolicited bulk email (UBE) or unsolicited commercial email (UCE) [1] and always aiming at commercial promotion or marketing, cause many problems to our daily-communication life. Ferris Research [2] revealed that large amount of spam not only occupied network bandwidth and server storage, but also wasted users' time on reading and deleting them, which resulted in loss of productivity. Moreover, the spam with malware threatens internet safety and personal privacy. According to Commtouch Internet Threats Trend Report [3], though the spam-sending Grum botnet was taken offline near the end of July, 2012, the effect on global spam levels was very short-lived and still 87 billion spam in average were sent every day in the third quarter of 2012. The statistics from Symantec Intelligence Report [4] demonstrate that spam made up 70.6% of the total email traffic in December, 2012, an increase of 1.8 percentage points since November. Meanwhile, one in 277.8 emails contained malware and one in 377.4 emails was identified as phishing.

In solving the spam problem, machine learning based automatic spam filtering gives promising performance. In spam filtering, the pivotal problem to be addressed turns into email categorization, i.e. classifying emails as spam or ham (non-spam), which can be seen as a special binary text categorization task. There are three main related research fields for text categorization, namely feature selection, feature construction and classifier design. Machine learning methods are widely applied for classifier design in spam filtering, such as Naive Bayes (NB) [5], [6], Support Vector Machine (SVM) [7], Artificial Neural Network (ANN) [8], [9], k-Nearest Neighbor (k-NN) [10], [11], Boosting [12] and so on. In text categorization, features are the terms (words) found in texts. The purpose of feature selection lies in reducing the number of terms to be further processed and the affect from possible noisy terms, so as to reduce the computational complexity and enhance the categorization accuracy respectively. Feature construction approaches transform the set of features available into a new set of features by finding relationships between existing features and constructing feature vectors to represent texts. Since performance of machine learning methods seriously depends on the feature vectors constructed, research on feature construction approaches has been focused in recent years. Some prevalent feature selection metrics and feature construction approaches in spam filtering are introduced in Section II.

In this paper, we propose a term space partition (TSP) based feature construction approach for email categorization by taking inspiration from the distribution of terms with respect to feature selection metrics. Class tendency of terms is also involved to perform the term space partition. Term ratio and term density are defined as features on corresponding subspaces. The TSP approach commits to make the terms in different subspaces play more sufficient and rational roles in email categorization, so as to improve the performance and efficiency of spam filtering. We conducted experiments on five benchmark corpora PU1, PU2, PU3, PUA and Enron-Spam to investigate the performance of the TSP approach. Accuracy and F_1 measure are selected as the main criteria in analyzing and discussing the results.

The rest of this paper is organized as follows: Section II introduces and analyzes the prevalent feature selection metrics and feature construction approaches applied in spam filtering. The proposed TSP approach is presented in Section III. Section IV gives the experimental results. Finally, we

Guyue Mi, Pengtao Zhang and Ying Tan are with the Key Laboratory of Machine Perception (Ministry of Education) and Department of Machine Intelligence, School of Electronics Engineering and Computer Science, Peking University, Beijing, 100871, China (email: gymi@pku.edu.cn, pengtaozhang@gmail.com, ytan@pku.edu.cn).

This work is supported by the National Natural Science Foundation of China under grants No. 61170057 and 60875080.

Prof. Ying Tan is the corresponding author.

conclude the paper in Section V.

II. Related Works

A. Feature Selection Metrics

1) Document Frequency Thresholding: Document Frequency (DF) Thresholding [13] is the simplest method for feature selection with the lowest cost in computation, by computing the number of documents which a certain term occurs in. It assumes that rare terms are not informative for class prediction and only terms which occurs frequently are selected and retained. When applied in spam filtering, DF of term t_i is calculated as

$$DF(t_i) = |\{m_j | m_j \in M, t_i \in m_j\}| \tag{1}$$

where M denotes the training set, and m_j is an email in M.

2) Term Strength: Term Strength (TS) [13] estimates term importance based on how commonly a term is likely to appear in "closely-related" documents, by computing conditional probability that a certain term occurs in the second half of a pair of related documents given that it occurs in the first half. It assumes that documents with many shared words are related, and terms in the heavily overlapping area of related documents are relatively informative. When applied in spam filtering, TS of term t_i is calculated as

$$TS(t_i) = P(t_i \in y | t_i \in x) \tag{2}$$

where x and y are an arbitrary pair of distinct but related emails in the training set M.

3) Information Gain: Information Gain (IG) [14] is the most widely employed feature goodness criterion in machine learning area, and the most commonly used method for feature selection in spam filtering. It measures the number of bits of information obtained for class prediction by knowing the presence or absence of a certain term in a text. When applied in spam filtering, IG of term t_i is calculated as

$$IG(t_i) = \sum_{c \in (s,h)} \sum_{t \in (t_i, \bar{t_i})} P(t,c) \log \frac{P(t,c)}{P(t)P(c)} \tag{3}$$

where c denotes the class of an email, s stands for spam, and h stands for ham, t_i and $\bar{t_i}$ denotes the presence and absence of term t_i respectively.

4) Term Frequency Variance: Based on DF, Koprinska *et al.* [15] proposed a Term Frequency Variance (TFV) method, by not selecting terms that are with high DF but occur frequently in each class. It can be obtained by computing the term frequency (i.e. document frequency of a certain term) in each class and then calculating the variance. When employed in spam filtering, TFV of term t_i is calculated as

$$TFV(t_i) = \sum_{c \in (s,h)} (T_f(t_i, c) - T_f^{\mu}(t_i))^2 \tag{4}$$

where $T_f(t_i, c)$ denotes term frequency of t_i in class c, $T_f^{\mu}(t_i)$ is the mean frequency of t_i in two classes.

5) Chi Square: Chi Square (χ^2) [13] measures the lack of independence between term t_i and class c. It has a natural value of 0 if term t_i and class c are completely independent, i.e. term t_i is useless for categorization. When applied in spam filtering, χ^2 of term t_i is calculated as

$$\chi^2(t_i) = \sum_{c \in (s,h)} P(c) \chi^2(t_i, c) \tag{5}$$

$$\chi^2(t_i, c) = \frac{|M|(P(t_i, c)P(\bar{t_i}, \bar{c}) - P(\bar{t_i}, c)P(t_i, \bar{c}))^2}{P(t_i)P(\bar{t_i})P(c)P(\bar{c})} \tag{6}$$

6) Odds Ratio: Odds Ratio (OR) [16]compares the odds of a term occurring in one class with the odds for it occurring in another class. It gives a positive score to terms that occur more often in one class than in the other, and a negative score otherwise. A score of 0 means the odds for a term to occur in one class is exactly the same as that in the other. When employed in spam filtering, OR of term t_i with class c is calculated as

$$OR(t_i, c) = \frac{P(t_i|c)}{1 - P(t_i|c)} \frac{1 - P(t_i|\bar{c})}{P(t_i|\bar{c})} \tag{7}$$

Usually, logarithms of OR values of a certain term with all different classes are calculated and added together as measure of the term.

Among the feature selection metrics above, DF and TS are class independent, while the others all use information about term-class associations. Comparative study on feature selection [13] demonstrates that IG and χ^2 perform the most effectively on aggressive dimensionality reduction in text categorization. DF performs similarly with IG and χ^2 which suggests that DF can be reliably used instead of IG or χ^2 when the computation of these measures are too expensive.

B. Feature Construction Approaches

1) Bag-of-Words: Bag-of-Words (BoW), also known as Space Vector Model, is one of the most widely used feature construction approaches in spam filtering [17]. It transforms an email m to a n-dimensional feature vector $\vec{x} = [x_1, x_2, ..., x_n]$ by utilizing a preselected term set $T = [t_1, t_2, ..., t_n]$, where the value x_i is given as a function of the occurrence of t_i in m, depending on the representation of the features adopted. In a binary representation, x_i is equal to 1 when t_i occurs in m, and 0 otherwise. In a frequency representation, x_i is assigned as the number of occurrences of t_i in m. Experimental results in [18] show binary representation and frequency representation have similar performance.

2) Sparse Binary Polynomial Hashing: Yerazunis utilized Sparse Binary Polynomial Hashing (SBPH) to extract a large amount of different features from email [19]. This method applies an n-term-length sliding window shifting over an email with a step of one term. At each movement, the newest term in the window is retained and the others are removed or retained so that the whole window is mapped to different features. Hence, 2^{n-1} features are extracted at each movement. SBPH performed quite promisingly in terms of classification accuracy in experiments as it could extract enough discriminative features. However, so many features lead to a heavy computational burden and limit its usability.

3) Orthogonal Sparse Bigrams: Siefkes *et al.* proposed Orthogonal Sparse Bigrams (OSB) based on analyzing the relevance of features to reduce the redundancy and complexity of SBPH [20]. OSB also utilizes an n-term-length sliding window shifting over an email with a step of one term, while only term-pairs with a common term in the window are considered. At each movement, the newest term is retained, then one of the other terms is selected to be retained while the others are removed. The remaining term-pair is mapped to a feature. Hence, $n - 1$ features are extracted at each

movement, greatly reducing the number of features compared with SBPH. Experiments show OSB slightly outperforms SBPH in terms of error rate.

4) Concentration Based Feature Construction Approach: Inspired by human immune system, Tan *et al.* proposed a concentration based feature construction (CFC) approach for spam filtering by computing "self" and "non-self" concentrations with respect to "self" and "non-self" gene libraries constructed on training set [21], [22]. The CFC approach transforms emails into 2-dimensional feature vectors, making great reduction on dimension of feature vectors. Experiments demonstrate that CFC outperforms BoW in both classification accuracy and efficiency.

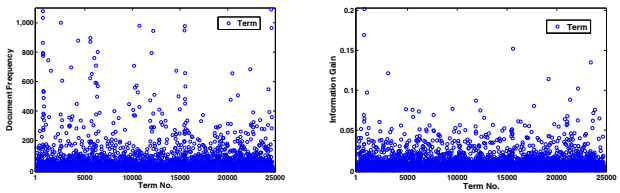
5) Local-Concentration Based Feature Extraction Approach: Zhu *et al.* proposed a local-concentration (LC) based feature extraction approach for spam filtering by taking inspiration from the immune mechanism that local concentrations of antibodies determine whether the corresponding pathogens can be culled from the body [23]. The LC approach is considered to be able to extract position-correlated information by computing "Self" and "non-self" concentrations on local areas. Experimental results demonstrate that the LC approach achieves higher performance than the current approaches as well as high processing speed.

III. Term Space Partition Based Feature Construction Approach

A. Motivation

Feature selection plays quite important roles in spam filtering and other text categorization problems, as the removal of less informative terms can reduce not only computational complexity but also affect from possible noisy terms. In decades, several feature selection metrics have been proposed and applied in text categorization as well as other pattern recognition issues. There are mainly two kinds of feature selection metrics, namely class independent ones and class dependent ones, also referred to as unsupervised and supervised feature selection metrics. We take DF and IG as representatives of unsupervised and supervised feature selection metrics respectively to investigate the distribution of terms with respect to these feature selection metrics in email categorization. Fig. 1(a) and Fig. 1(b) show the distribution of terms in a commonly used benchmark corpus, which is named PU1 and contains nearly 25,000 distinct terms, with respect to DF and IG respectively, where the terms are numbered continuously.

As we can see, no matter which kind of feature selection metrics is employed, similar distribution reveals that rare terms can get much higher and discriminative scores while the great majority of terms are given relatively low and similar scores, though term distribution with respect to IG in high-score space is sparser than that of DF due to the consideration of term-class associations. Terms are selected according to the scores they get. The distribution of terms suggests that only a few terms with obvious superiority can be selected confidently, while the others are with weak



(a) DF (b) IG

Fig. 1. Distribution of Terms in PU1 with Respect to Feature Selection Metrics

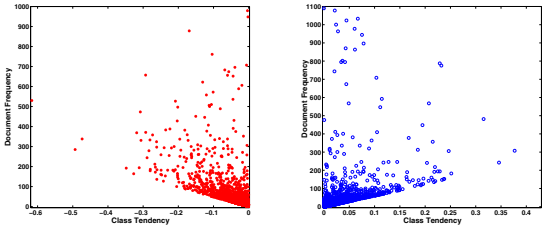
confidence as they are not so superior and discriminative with each other on the evaluation.

Terms reserved by feature selection metrics are further utilized to construct feature vectors of emails. Therefore, the extent of feature selection is determined by the corresponding feature construction approach employed. For the traditional BoW approach, several hundred terms are selected. Each selected term corresponds to an individual dimension of the feature vector and plays a sufficient role in email categorization. In our consideration, BoW has the following restraints: 1) terms reserved far exceed that with obvious superiority in the term space, and it is not reasonable that low score terms are considered equally important with that given much higher scores; 2) only a small part of terms in the term space are utilized, which indicating waste of information; 3) several hundred dimensional feature vectors still lead to a heavy computational burden. While for the heuristic approaches, i.e. CFC and LC by taking inspiration from biological immune system, more terms (empirically more than 50% of the terms in the original term space [21], [22], [23]) are reserved for further constructing feature vectors and feature vector dimension is reduced by computing concentration features. However, there exists a similar but more prominent deficiency in these two approaches as in BoW that terms with obvious superiority are treated equally with terms given much lower scores, which weakens the contributions on categorization from the superior terms.

B. Principle

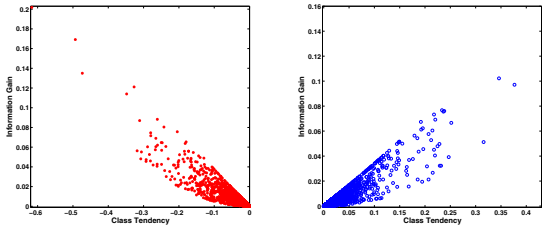
The proposed TSP approach aims to establish a mechanism to make the terms play more sufficient and rational roles in email categorization by dividing the original term space into subspaces and designing corresponding feature construction strategy on each subspace, so as to improve the performance and efficiency of spam filtering.

Feature selection metrics give terms reasonable and effective goodness evaluation. According to the distribution characteristics of terms with respect to feature selection metrics, a vertical partition of the term space is performed to separate the *Dominant Terms* from *General Terms*. By dominant terms, we mean the terms given high and discriminative scores by feature selection metrics and considered to lead the categorization results, and this part of terms has the following features: small amount, sparse distribution, discriminative and informative. While large amount of



(a) Spam Terms (b) Ham Terms

Fig. 2. Distribution of Terms in PU1 with Respect to DF



(a) Spam Terms (b) Ham Terms

Fig. 3. Distribution of Terms in PU1 with Respect to IG

general terms congregate in a narrow range of the term space with similar low scores. Though general terms are less informative than dominant terms and adulterated with redundant and noisy terms, most of them can also contribute to email categorization, which can not be ignored easily. Undoubtedly, dominant terms and general terms should play different roles in email categorization.

To construct discriminative features, we introduce *Class Tendency* to perform a transverse partition of the term space to separate the *Spam Terms* from *Ham Terms*. By class tendency, we mean the tendency of a term occurring in emails of a certain class, defined as

$$tendency(t_i) = P(t_i|c_h) - P(t_i|c_s) \tag{8}$$

where $P(t_i|c_h)$ is the probability of t_i's occurrence, given the email is ham, and $P(t_i|c_s)$ is the probability of t_i's occurrence, given the email is spam. Spam terms are terms that occur more frequently in spam than in ham with negative tendency, and ham terms occur more frequently in ham than in spam with positive tendency.

In this case, each term in the original term space can be represented by a 2-dimensional vector, i.e. $\vec{t} =< tendency, goodness >$. Distribution of terms in PU1 with respect to DF and IG in the newly constructed term space is shown in Fig. 2 and Fig. 3. As we can see, the whole term space is divided into a spam term space and a ham term space. The spam term space and ham term space contain both dominant terms and general terms. Therefore, the original term space is decomposed into four independent and non-overlapping subspaces, namely spam-dominant, ham-dominant, spam-general and ham-general.

For the partition between dominant terms and general terms, we finally employed a thresholding method through analysis and experiments. *Spam Term Ratio* and *Ham Term Ratio* are defined as features on dominant terms, while *Spam Term Density* and *Ham Term Density* are computed on general terms, which will be introduced next.

C. The TSP Approach

We decompose the TSP approach into the following steps:

1) Preprocessing: The purpose of preprocessing is transforming emails into terms by examining the existence of blank spaces and delimiters, also referred to as tokenization. It is a quiet simple but indispensable step. The followings are essential steps of the TSP approach.

2) Term Space Partition: Algorithm 1 gives a detailed description of the term space partition step, which mainly contains term selection and term space partition. Term selection is involved to reduce the computational complexity and affect from possible noisy terms. Parameter p determines the extent of term selection.

Algorithm 1 Term space partition

initialize preselected term set TS_p, spam-dominant term set TS_{sd}, ham-dominant term set TS_{hd}, spam-general term set TS_{sg} and ham-general term set TS_{hg} as empty sets

for each term t_i occurs in the training set **do**
  calculate goodness evaluation $\tau(t_i)$ according to the feature selection metrics employed
end for
sort the terms in descending order of evaluation
add the front $p\%$ terms to TS_p

calculate partition threshold θ_{dg} according to Eq. 9
for each term t_i in TS_p **do**
  calculate $tendency(t_i)$ according to Eq. 8
  if $tendency(t_i) < 0$ **then**
    if $\tau(t_i) \geq \theta_{dg}$ **then**
      add t_i to TS_{sd}
    else
      add t_i to TS_{sg}
    end if
  else
    if $tendency(t_i) > 0$ **then**
      if $\tau(t_i) \geq \theta_{dg}$ **then**
        add t_i to TS_{hd}
      else
        add t_i to TS_{hg}
      end if
    end if
  end if
end for

The vertical partition is performed to separate dominant terms and general terms by defining a threshold θ_{dg} with

respect to the corresponding feature selection metrics employed, as shown in Eq. 9.

$$\theta_{dg} = \frac{1}{r}(\tau_{max} - \tau_{min}) \tag{9}$$

where τ_{max} and τ_{min} depict the highest and lowest evaluation of terms in the training set respectively, and variable r controls the restriction level of dominant terms. Term t_i with $\tau(t_i) \geq \theta_{dg}$ is considered as dominant term, and general term otherwise. When performing the transverse partition to separate spam terms and ham terms, terms with $tendency(t_i) = 0$ are considered useless and discarded.

3) Feature Construction: To construct discriminative and effective feature vectors of emails, we define *Term Ratio* and *Term Density* on dominant terms and general terms respectively as features to make the terms play sufficient and rational roles in email categorization. Algorithm 2 shows the process of feature construction.

Algorithm 2 Feature Construction

1: calculate spam term ratio TR_s according to Eq. 10
2: calculate ham term ratio TR_h according to Eq. 11
3: calculate spam term density TD_s according to Eq. 12
4: calculate ham term density TD_h according to Eq. 13
5:
6: combine TR_s, TR_h, TD_s and TD_h together to form the feature vector

For the very small amount of dominant terms, which are considered to lead the categorization results, each individual term should be given more weights to play sufficient roles in email categorization. Spam term ratio and ham term ratio are calculated on spam-dominant terms and ham-dominant terms respectively. Spam term ratio is defined as

$$TR_s = \frac{n_{sd}}{N_{sd}} \tag{10}$$

where n_{sd} is the number of distinct terms in the current email which are also contained in spam-dominant term space TS_{sd}, and N_{sd} is the total number of distinct terms in TS_{sd}. Similarly, ham term ratio is defined as

$$TR_h = \frac{n_{hd}}{N_{hd}} \tag{11}$$

where n_{hd} is the number of distinct terms in the current email which are also contained in ham-dominant term space TS_{hd}, and N_{hd} is the total number of distinct terms in TS_{hd}.

While for the large amount of general terms, which are less informative and may be adulterated with redundant and noisy terms, the affect of individual term should be weakened. Spam term density and ham term density are calculated on spam-general terms and ham-general terms respectively. Spam term density is defined as

$$TD_s = \frac{n_{sg}}{N_e} \tag{12}$$

where n_{sg} is the number of distinct terms in the current email which are also contained in spam-general term space TS_{sg}, and N_e is the total number of distinct terms in the current email. And ham term density is defined as

$$TD_h = \frac{n_{hg}}{N_e} \tag{13}$$

where n_{hg} is the number of distinct terms in the current email which are also contained in ham-general term space TS_{hg}.

In this step, term ratio and term density are two essential but completely different concepts. Term ratio indicates the percentage of dominant terms that occur in the current email, emphasizing the absolute ratio of dominant terms. In this way, the contributions to categorization from dominant terms are strengthened and not influenced by other terms. While term density represents the percentage of terms in the current email that are general terms, focusing on the relative proportion of terms in the current email that are general terms. The effect on categorization from general terms is weakened and so is the affect from possible noisy terms.

Finally, the achieved features are combined together to form the feature vector, i.e. $\vec{v} =< TR_s, TR_h, TD_s, TD_h >$.

IV. Experiments

A. Corpora

Experiments were conducted on PU1, PU2, PU3, PUA [24] and Enron-Spam [25], which are all benchmark corpora widely used for effectiveness evaluation in spam filtering. PU1 contains 1099 emails, 481 of which are spam. PU2 contains 721 emails, and 142 of them are spam. 4139 emails are included in PU3 and 1826 of them are spam. 1142 emails are included in PUA and 572 of them are spam. Enron-Spam contains 33716 emails, 17171 of which are spam. Emails in the five corpora all have been preprocessed by removing header fields, attachment and HTML tags, leaving subject and body text only. For privacy protection, emails in PU corpora have been encrypted by replacing meaningful terms with specific numbers.

B. Experimental Setup

All the experiments were conducted on a PC with E4500 CPU and 2G RAM. SVM was employed as classifier and LIBSVM [26] was applied for implementation of SVM. 10-fold cross validation was utilized on PU corpora and 6-fold cross validation on Enron-Spam according to the number of parts each of the corpora has been already divided into. Accuracy and F_1 measure are the main evaluation criteria, as they can reflect the overall performance of spam filtering.

C. Investigation of Parameters

Experiments have been conducted on PU1 to investigate the parameters of the TSP approach. 10-fold cross validation was utilized. There are two important parameters. Parameter p in the term space partition step determines the percentage of terms reserved for feature construction. As mentioned before, removal of less informative terms can reduce not only computational complexity but also affect from possible noisy terms, so as to improve the efficiency and performance

of spam filtering. While parameter r, which is much more essential and controls the restriction level of dominant terms in the term space partition step, depicts the core idea of this TSP approach. With small r, the restriction level of dominant terms is high and thus the number of dominant terms as defined is small, and vice versa.

Since the distribution of dominant terms with respect to supervised feature selection metrics is sparser than that of unsupervised ones, we first investigate the parameters in TSP with respect to unsupervised feature selection metrics and DF is selected as the representative. Fig. 4 shows the performance of TSP with respect to DF under varied r. As expected, the performance of TSP shows improvements with r getting larger in the first half. Thus, $r = 7$ is considered a suitable selection of parameter r, where the TSP approach performs best and relatively high precision and recall are achieved.

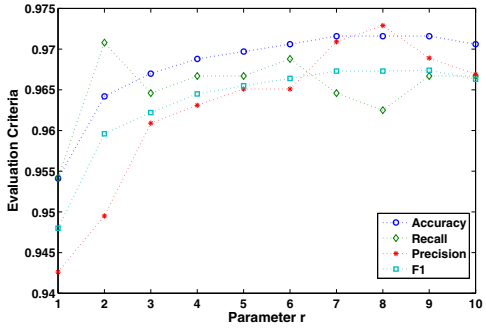


Fig. 4. Performance of TSP with respect to DF under varied r

The performance of TSP with respect to DF under varied p is shown in Fig. 5. As we see, the TSP approach always performs quite well though p varies, and better performance can be achieved with larger ps. For efficiency consideration, $p = 30$ is selected, which means the front 30% terms with respect to DF are reserved for feature construction. On the other hand, better performance on larger ps indicates that the term density strategy is effective to make general terms paly roles in email categorization.

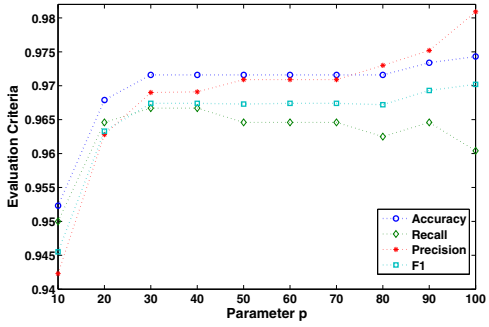


Fig. 5. Performance of TSP with respect to DF under varied p

Similar experiments were conducted to tune the parameters of TSP when supervised feature selection metrics are employed, and IG is selected as the representative of this kind of metrics. From the experimental results, $r = 3$ and $p = 30$ are selected.

D. Performance with Different Feature Selection Metrics

In the proposed TSP approach, the vertical partition of the term space is performed according to term evaluation given by feature selection metrics. Therefore, it is quite necessary to verify whether the TSP approach can cooperate well with both unsupervised and supervised feature selection metrics.

We selected DF and IG as representatives of unsupervised and supervised feature selection metrics respectively to conduct comparison experiments, for these two metrics are widely applied in spam filtering and other text categorization issues. Performance of TSP with respect to DF and IG on five benchmark corpora PU1, PU2, PU3, PUA and Enron-Spam is shown in Table I. As the experimental results reveal, the TSP approach performs quite well with both DF and IG, which represent unsupervised and supervised feature selection metrics correspondingly. It is worth noting that DF outperforms IG with TSP as feature construction approach in most cases of the experiments, indicating that the transverse partition of term space is effective to make use of the information of term-class associations.

E. Comparison with Current Approaches

Experiments were conducted on PU1, PU2, PU3, PUA and Enron-Spam to compare the performance of the proposed TSP approach with that of current approaches. The selected current approaches are BoW, CFC and LC. Table II shows the performance of each feature construction approach in spam filtering when incorporated with SVM, and the corresponding dimensionality of feature vectors constructed. As mentioned before, we take accuracy and F_1 measure as comparison criteria without focusing on precision and recall, which are incorporated into the calculation of F_1 measure and can be reflected by F_1 measure.

BoW is a traditional and one of the most widely used feature construction approach in spam filtering. As we can see, the proposed TSP approach not only make significant reduction on the feature vector dimension so as to improve efficiency but also achieve much better performance in terms of both accuracy and F_1 measure when compared with BoW, indicating that the TSP approach is effective for email categorization.

CFC and LC are heuristic and state-of-the-art approaches in spam filtering by taking inspiration from biological immune system. The CFC approach transforms emails into two-dimensional feature vectors by calculating "self" and "non-self" concentrations, while the LC approach extracting position-correlated information from messages additionally to CFC by constructing concentration features on local areas. LC-FL and LC-VL utilize different strategies of defining local areas respectively. The CFC and LC approaches achieve not only good performance but also high efficiency. The experimental results show that the TSP approach far outperforms CFC and LC in terms of both accuracy and

TABLE I

PERFORMANCE OF TSP WITH RESPECT TO DIFFERENT FEATURE SELECTION METRICS

Corpus	Feature sel.	Precision(%)	Recall(%)	Accuracy(%)	F_1(%)
PU1	DF	96.90	96.67	**97.16**	**96.74**
	IG	96.07	96.46	96.70	96.21
PU2	DF	94.09	83.57	**95.63**	**88.12**
	IG	96.32	80.00	95.35	87.09
PU3	DF	95.69	95.88	96.20	95.73
	IG	96.37	97.09	**97.05**	**96.69**
PUA	DF	95.91	96.49	**96.05**	**96.11**
	IG	95.62	94.74	95.00	95.06
Enron-Spam	DF	94.29	98.21	**97.02**	**96.14**
	IG	94.18	98.23	96.90	96.12

TABLE II

PERFORMANCE COMPARISON OF TSP WITH CURRENT APPROACHES

Corpus	Approach.	Precision(%)	Recall(%)	Accuracy(%)	F_1(%)	Feature dim.
PU1	BoW	93.96	95.63	95.32	94.79	600
	CFC	94.97	95.00	95.60	94.99	2
	LC-FL	95.12	96.88	96.42	95.99	20
	LC-VL	95.48	96.04	96.24	95.72	6
	TSP	96.90	96.67	**97.16**	**96.74**	4
PU2	BoW	88.71	79.29	93.66	83.74	600
	CFC	95.12	76.43	94.37	84.76	2
	LC-FL	90.86	82.86	94.79	86.67	20
	LC-VL	92.06	86.43	**95.63**	**88.65**	6
	TSP	94.09	83.57	**95.63**	88.12	4
PU3	BoW	96.48	94.67	96.08	95.57	600
	CFC	96.24	94.95	96.05	95.59	2
	LC-FL	95.99	95.33	96.13	95.66	20
	LC-VL	95.64	95.77	96.15	95.67	6
	TSP	96.37	97.09	**97.05**	**96.69**	4
PUA	BoW	92.83	93.33	92.89	93.08	600
	CFC	96.03	93.86	94.82	94.93	2
	LC-FL	96.01	94.74	95.26	95.37	20
	LC-VL	95.60	94.56	94.91	94.94	6
	TSP	95.91	96.49	**96.05**	**96.11**	4
Enron-Spam	BoW	90.88	98.87	95.13	94.62	600
	CFC	91.48	97.81	95.62	94.39	2
	LC-FL	94.07	98.00	96.79	95.94	20
	LC-VL	92.44	97.81	96.02	94.94	6
	TSP	94.29	98.21	**97.02**	**96.14**	4

F_1 measure, which verified that the proposed term space partition strategy and newly defined features, namely term ratio and term density, are successful to make terms play more sufficient and rational roles in email categorization. Meanwhile, the TSP approach reduces the feature vector dimension with fixed 4-dimensional feature vectors, compared with both LC-FL and LC-VL. It is worth noting that the TSP approach achieves much higher and stabler precision in spam filtering, which is warmly welcomed in spam filtering as email users would rather accept more spam than discard useful emails.

We conducted experiments on PU1 to compare the efficiency of TSP with that of the selected current feature construction approaches. 10-fold cross validation was utilized. Table III shows the average time spent of each approach on processing one incoming email. As we can see, all of the CFC, LC, and TSP approaches perform far more efficient than BoW, due to significant reduction on feature vector dimension. Since the LC approach need to calculate concentrations on each local area and finally construct feature vectors with more dimension, the TSP approach can process incoming emails faster than both LC-FL and LC-VL. Al-

TABLE III
EFFICIENCY COMPARISON OF TSP WITH CURRENT APPROACHES

Approach	BoW	CFC	LC-FL	LC-VL	TSP
Seconds/email	$9.57e^{-3}$	$3.75e^{-4}$	$5.52e^{-4}$	$4.50e^{-4}$	$3.91e^{-4}$

though the feature vectors constructed by TSP has additional two dimension compared with CFC, TSP can achieve similar efficiency with CFC, as the term space partition strategy dividing the original term space into four non-overlapping subspaces and less time is spent on computing term ratio and term density compared with the computing of concentrations in CFC. On the other hand, less terms are reserved in TSP than CFC and LC for feature construction to achieve better performance as the dominant terms can play sufficient roles in TSP.

V. CONCLUSIONS

In this paper, we proposed a TSP based feature construction approach for email categorization by dividing the original term space into subspaces and constructing features on each subspace independently. The vertical and transverse partitions are performed with respect to feature selection metrics and class tendency respectively. Discriminative features are constructed by computing term ratio and term density on corresponding subspaces. The TSP approach is proved effective in establishing a mechanism to make terms in the original term space play more sufficient and rational roles in email categorization by comprehensive experiments. Furthermore, it achieves high efficiency by transforming emails into 4-dimensional feature vectors.

In future work, we intend to incorporate the TSP approach with other classification methods. In addition, we hope to apply the TSP approach on multi-class tasks, where the strategy of transverse partition should be redesigned.

REFERENCES

[1] L. Cranor and B. LaMacchia, "Spam!" *Communications of the ACM*, vol. 41, no. 8, pp. 74–83, 1998.

[2] F. Research, "Spam, spammers, and spam control: A white paper by ferris research," *Tech. rep.*, 2009.

[3] Commtouch, "Internet threats trend report - october 2012," *Tech. rep.*, 2012.

[4] Symantec, "Symantec intelligence report: December 2012," *Tech. rep.*, 2013.

[5] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A bayesian approach to filtering junk e-mail," in *Learning for Text Categorization: Papers from the 1998 workshop*, vol. 62. Madison, Wisconsin: AAAI Technical Report WS-98-05, 1998, pp. 98–105.

[6] A. Ciltik and T. Gungor, "Time-efficient spam e-mail filtering using n-gram models," *Pattern Recognition Letters*, vol. 29, no. 1, pp. 19–33, 2008.

[7] H. Drucker, D. Wu, and V. Vapnik, "Support vector machines for spam categorization," *Neural Networks, IEEE Transactions on*, vol. 10, no. 5, pp. 1048–1054, 1999.

[8] J. Clark, I. Koprinska, and J. Poon, "A neural network based approach to automated e-mail classification," in *Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on*. IEEE, 2003, pp. 702–705.

[9] C. Wu, "Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks," *Expert Systems with Applications*, vol. 36, no. 3, pp. 4321–4330, 2009.

[10] I. Androutsopoulos, G. Paliouras, V. Karkaletsis, G. Sakkis, C. Spyropoulos, and P. Stamatopoulos, "Learning to filter spam e-mail: A comparison of a naive bayesian and a memory-based approach," *Arxiv preprint cs/0009009*, 2000.

[11] G. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, C. Spyropoulos, and P. Stamatopoulos, "A memory-based approach to anti-spam filtering for mailing lists," *Information Retrieval*, vol. 6, no. 1, pp. 49–73, 2003.

[12] X. Carreras and L. Marquez, "Boosting trees for anti-spam email filtering," *Arxiv preprint cs/0109015*, 2001.

[13] Y. Yang and J. Pedersen, "A comparative study on feature selection in text categorization," in *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*. MORGAN KAUFMANN PUBLISHERS, INC., 1997, pp. 412–420.

[14] Y. Yang, "Noise reduction in a statistical approach to text categorization," in *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1995, pp. 256–263.

[15] I. Koprinska, J. Poon, J. Clark, and J. Chan, "Learning to classify e-mail," *Information Sciences*, vol. 177, no. 10, pp. 2167–2187, 2007.

[16] W. Shaw, "Term-relevance computations and perfect retrieval performance," *Information Processing & Management*, vol. 31, no. 4, pp. 491–498, 1995.

[17] T. Guzella and W. Caminhas, "A review of machine learning approaches to spam filtering," *Expert Systems with Applications*, vol. 36, no. 7, pp. 10 206–10 222, 2009.

[18] K. Schneider, "A comparison of event models for naive bayes anti-spam e-mail filtering," in *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2003, pp. 307–314.

[19] W. Yerazunis, "Sparse binary polynomial hashing and the crm114 discriminator," *the Web.[Online]. Available: http://crm114. sourceforge. net/CRM114 paper. html*, 2003.

[20] C. Siefkes, F. Assis, S. Chhabra, and W. Yerazunis, "Combining winnow and orthogonal sparse bigrams for incremental spam filtering," *Knowledge Discovery in Databases: PKDD 2004*, pp. 410–421, 2004.

[21] Y. Tan, C. Deng, and G. Ruan, "Concentration based feature construction approach for spam detection," in *Neural Networks, 2009. IJCNN 2009. International Joint Conference on*. IEEE, 2009, pp. 3088–3093.

[22] G. Ruan and Y. Tan, "A three-layer back-propagation neural network for spam detection using artificial immune concentration," *Soft Computing-A Fusion of Foundations, Methodologies and Applications*, vol. 14, no. 2, pp. 139–150, 2010.

[23] Y. Zhu and Y. Tan, "A local-concentration-based feature extraction approach for spam filtering," *Information Forensics and Security, IEEE Transactions on*, vol. 6, no. 2, pp. 486–497, 2011.

[24] I. Androutsopoulos, G. Paliouras, and E. Michelakis, *Learning to filter unsolicited commercial e-mail.* " DEMOKRITOS", National Center for Scientific Research, 2004.

[25] V. Metsis, I. Androutsopoulos, and G. Paliouras, "Spam filtering with naive bayes-which naive bayes," in *Third conference on email and anti-spam (CEAS)*, vol. 17, 2006, pp. 28–69.

[26] C. Chang and C. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.