A Vector Space Model based spam SMS filter

Wei Li
Shanghai Dahan Tricom Data Co. Ltd.
Shanghai, China
8002@dahantc.com

Sisheng Zeng
Shanghai Dahan Tricom Data Co. Ltd.
Shanghai, China
8601@dahantc.com

Abstract—**Along with the popularity of telecommunication and mobile phone, short message (SMS) enters almost every human life. Meanwhile, each mobile phone client suffers from the harass of spam SMS.**

As the SMS service provider who is in charge of all industry SMS in east China, Dahan Tricom Corporation always invest much in anti spam SMS research. Recent years, we upgrade our anti-spam filter to semantic level. The core technology is described in this paper.

Unlike other anti spam filter, such as anti spam emails, the anti spam SMS filter must face many difficulties oriented by SMS itself. Since SMS contains only 70 Chinese characters or 140 English letters at the most, it is always lack of semantic information. Also, vocal expressions often appear in SMS. In addition, the industry SMS often contains proper terms related to specific industry field.

The anti spam SMS filter in this paper first leverages Vector Space Model (VSM) as its foundation technologies. Then, many modifications are made in the process in VSM method to address the difficulties of spam SMS filtering issue.

Finally, the experiment result turns out to be acceptable in our commercial production environment.

Keywords- Anti spam, short message, SMS, Vector Space Model, Natural Language Processing

I. INTRODUCTION

According to the China National Ministry of industry and information, mobile phone users has beyond 1,300 million in China. However, reports of Spam SMSs indicates Spam SMSs increases greatly and become a big Telecom issue in China.

Along with the rapid improvements of telecommunication technology and its application in China, Dahan Tricom Corporation, in charge of main short message(SMS) service traffic and its value added services of East China, has made great progress in the past 13 years. Dahan Tricom Corporation qualified for High-tech Enterprises, High-tech Service Providers, Enterprise R & D institutions acquitted by China National Ministry of industry and information and Shanghai Official Government and has branch companies in Peking, Chongqing, Jinan, Shenzhen, Suzhou, etc. Currently, more than 300 million SMSs are sent by Dahan per month.

However, we encounter Spam SMS problem which also threatens other SMS services providers and terminal mobile users in China.

According to a statistics' report, in China, reported spam SMSs are more than 440 million calculated from January to September in year 2015. Spam SMSs have been a big problem in China Telecom because it contains advertisement information (e.g. insurance, real estate, education, Telecom service marketing, etc.), illegal information, fraud information. This information not only wastes traffic, spread unintended information, and also causes serious illegal consequences.

By statistics, the spam SMSs are sent from 3 ways: p2p mobile phone, SMS sending port and fake base station. Dahan Tricom corporation are mainly concerning SMS sending port services. That is, SMS senders submit SMS and its recipients to Dahan SMS platform and the SMS will be sent to the recipients by the platform. Since the amount of simultaneously sending SMSs are overwhelming, it is impossible to judge the SMSs manually.

The main SMS service providers, including China Telecom, China Mobile and China Unicom, are using keyword to filter spam SMS. Some of them, using logical combinations of keywords to filter spam SMS. However, the mechanical matching cannot meet anti-spam SMS requirements as it often judges incorrectly.

Therefore, it is necessary to leverage Natural Language Processing technologies to develop a semantic based anti-spam SMS filter to greatly increase the accuracy and recall value.

Vector Space Model (VSM) is a worldwide applied machine learning technology which has been found in many applications including classification, estimation and tracking as in [1], [2], [3] and [4]. In VSM algorithm, it calculates the closest data vectors according to training set called center vector (CV) and it process a given new test vector to corresponding categories according to the boundary determined by using only these center vectors [5][6].

In this paper, a Vector Space Model (VSM)[1] based spam SMS filter is put forward to address this issue in SMSs Service industry. This method can be applied to other plain text data which contains sematic context expressing the main content or purpose of the data.

II. CONCERNS OF SMS FILTER

Although Vector Space Model (VSM) is able to handle various scenarios, in order to enhance the accuracy of the classifier, it is better to considerate the concerning of SMS and make proper optimization of VSM algorithm.

They can be summarized as follows:

A. SMSs are always short.

SMSs contains 70 Chinese characters or 140 English letters at the most. Actually, SMS is always short and short of semantic information inside.

Therefore, it is difficult to gather the limited information to determine the actual meaning of SMS.

B. It is difficult to calculate word weights for SMS.

Since SMS is always short, the important terms may appear infrequently. In consequence, many term weight algorithms, such as term frequency, term position, HTML tags[2] and term length, is not able to take effects in this scenario. Also, for feature selection, many techniques, such as Mutual information[3], the weight of evidence for text, Information Gain[4], Expected cross entropy[5], etc., does not works fine.

C. Industry SMSs are always related to specific domains.

Our Corporation, Dahan Tricom, focuses on industry SMSs including E- Commerce, logistics, real estate, etc. For this reason, these SMSs always contains many certain proper terms, which may be unable to be recognized by Chinese segmentation.

D. SMSs express vocally.

Unlike the content in textbooks, SMS always contain vocal expressions because the readers of SMSs are ordinary people. Such expressions are not strict to grammar. Therefore, the method must be flexible so that it can adapt to various way of talking expressions.

E. SMSs often contains links.

The advertisement SMS always contains URL links to encourage reader to click on. Therefore, it is better to use such character to recognize spam SMS.

III. VECTOR SPACE MODEL (VSM)

The classical vector space model (VSM) was put forward by Salton et al. [1] in 1975.

VSM represents the terms in documents as documents vectors, which consists of term weights denoting the semantic importance of certain term in certain document. Then, final decision is made according to similarity measurement of vectors.

By far, as for the term weighting algorithms, TFIDF, Mutual information[3], Information Gain[4], the weight of evidence for text, Expected cross entropy[5], etc. are most frequently used.

Among these, the term frequency/inverse document frequency (TD/IDF) with normalized frequency is the most commonly used term weighting algorithm.

As for the similarity measurement between a test document and a center document vector which is the sematic representation of Spam SMSs and white SMSs, can be calculated by Mahalanobis distance, Euclidean distance, cosine measure or Manhatan distance.

IV. VECTOR SPACE MODEL BASED SPAM SMS FILTER

The Vector Space Model based Spam SMS Filter consists of two parts: trainer and filter. The trainer results in two center document vectors each for Spam SMSs and white SMSs according to training corpus which contains documents of the two categories. The filter analyses the test document, calculate the sematic similarity values with the two center documents and decide the more similar one is the result.

Trainer process documents through the following steps.

A. Pre-process

Training corpus contains 2 categories of documents. One is spam SMSs and the other is white SMSs.

At the very beginning, documents in the training corpus and the one to be filtered, called test document, are represented formulated.

To build a document representation, for Chinese documents, it is necessary to segment these Chinese text, word tagging, etc. to obtain word sequences with part of speech attached.

B. Term Expansion

As mentioned above, SMS is always short and lack of semantic information. The terms in a single SMS is often insufficient to extract semantic features[9]. Therefore, we establish a synonym corpus based on previous Dahan Tricom service data to expand the terms found in SMS and assign a "exp." tag on it which will be taken in consideration in the feature weighting step.

C. Feature Weighting

Feature Weighting, also known as term weighting, will assign each word in each document a weight value to represent its sematic importance in the document. Then, the segmented term sequence is represented by document vectors in which each item indicates the semantic information of each term in the document.

In our system, the following term weight algorithms is included.

Suppose: the term weight $w(t_i, d_j)$ is the term weight value of term i in certain document j.

1) Term Frequency(TF)

As its name says, TF value is the term frequency in a single document. That is, TF is the times of a term's occurrence in a document.

In addition, for other natural language processing systems, since the term's occurrence is not stable, term frequency is taken as an important parameter of term to discriminate itself from other terms.

So, logarithm is fetched on term frequency, which is shown on the TF equation below.

$$TF_i = \log_2(tf_{ij}) . \quad (1)$$

where tf_i represents the frequency of term i in document j.

In SMS scenario, terms will not occur so many times in the short SMS text. Therefore, the original quation is directly used as the value of TF. That is, the TF value of term i is

$$TF_i = tf_{ik} . \quad (2)$$

2) Term Frequency(TF)

As for IDF, many formulas were put foward. The origin formula was given by Robertson[7]. A later discussion between

Spärck Jones[6] and Robertson resulted in the equation of IDF below:

$$IDF_i = \log_2(\frac{N}{n_j}) + 1 = \log_2(N) - \log_2(n_j) + 1 \quad (3)$$

where N is the number of total documents in corpus and n_j is the count of documents which contain at least one occurrence of term i.

Actually, in SMS scenario, based on our research, it is proved that IDF takes much effects to decrease the weight value of terms without much semantic information.

3) Proper terms

For SMS scenario, TF-IDF is not sufficient. Since Dahan Tricom focus on industry SMS, the SMSs sent by our system are mainly concentrated on several industry field, such as E-Commerce, logistics, real estate, etc.

In order to emphasize proper terms[10], we categorized these terms into 3 levels according to their semantic importance in certain industry fields manually and assign a proper term coefficient to the TF-IDF algorithm.

4) Expanded terms

Also, for the expanded terms mentioned above, we will assign another coefficient for such terms reduce its importance because they do not appear in the document actually.

D. Feature Selection

Including the terms mentioned above, features for the text documents are words or phrases appearing in the documents. For text representation, each word is considered as a feature.

However, the processing data always contains redundant or irrelevant features. Redundant features provide no more information than selected features, and irrelevant features provide no useful information in any context. The redundant or irrelevant features will waste much computation time and memory requirement. Furthermore, irrelevant features also bring down the filter accuracy.

Feature selection is the process of selecting a subset of relevant features for use in document vector and excluding the redundant or irrelevant features.

A careful selection of words is desired instead of all words [8]. A simple unordered list of selected words based on its importance and associated proper weights are usually sufficient to represent a document.

To build a document representation, a collection of documents is involved rather than individual documents. The main purpose is to make it easy to classify documents. The size of an index can be reduced when the stems of words are used instead of all word forms.

Currently, many methods, such as Inverse Document Frequency, Entropy, Information Gain, Mutual Information, CHI Square, etc, have been applied to feature extraction in text classification systems.

Based on our experiment, Information Gain and CHI Square method usually show better performance and results.

E. Stop words

Much redundant and irrelevant information can be removed by the feature selection process.

However, some of the noise information may not be able to be removed by feature selection process. As defined, noise, generally defined in Natural Language Processing as the insignificant, irrelevant words or stop words, are normally present in any plain text.

Therefore, as a Stop words have an average distribution in any standard language corpus and do not normally contribute any information to classification tasks. But, since these stop words have high frequencies of occurrences and the average distribution, they will bring IDF value to a higher level.

F. Calculate center document vector for each category: spam SMSs and white SMSs

After being processed by trainer, the documents are segmented, weighted, feature selected, etc. and result in documents vectors of every document in training corpus.

The center document vector of spam SMSs and white SMSs can be calculated via the following method:

Suppose $\overrightarrow{d_1}, \overrightarrow{d_2}, ..., \overrightarrow{d_m}$ are all document vectors of a certain category, spam SMSs or white SMSs. Document i is represented as document vector $\overrightarrow{d_i}$ $(w_{i1}, ..., w_{it})$ in which w_{ij} is the term weight of feature-selected term i in document j. The center document vector is calculated based on the following formula:

$$d_{center} = (\frac{\sum_{i=1}^{m} w_{i1}}{m}, \frac{\sum_{i=1}^{m} w_{i2}}{m}, ..., \frac{\sum_{i=1}^{m} w_{in}}{m}) \quad (5)$$

m is the total number of the documents in the category and the n is the total number of the features selected from whole training corpus.

G. Similarity Measurement of documents

Before calculating the similarity value of the test document with white center document vector and spam center document vector, the test document will be processed through segment, term expansion, feature weighting, feature selection, and stop word movement.

Then, any one of the following distance measures can be applied. such as Euclidean distance, Manhatan distance Mahalanobis distance, or cosine measure to find the similarity of documents.

The Euclidean distance of test document vector and center document vectors is:

$$|d_{test}, d_{center}| = \sqrt{\sum_{j=1}^{m} (w_{test,j} - w_{center,j})^2} \quad (6)$$

Cosine measure or normalized correlation coefficient is:

$$\cos(d_{test}, d_{center}) = \frac{\sum_{j=1}^{m} w_{test,j} w_{center,j}}{\sqrt{\sum_{j=1}^{m} (w_{test,j})^2 (w_{center,j})^2}} \quad (7)$$

In our anti spam SMS filter, the cosine value between the vector of test document and the center document vector of spam SMS or white SMS category is used. The cosine similarity equation is as follows.

$$\cos(d_q, d_j) = \frac{d_q \cdot d_j}{\|d_q\| \|d_j\|} \quad (4)$$

where d_q and d_j denotes the vector of test document and center vector. Both $\|d_q\|$ and $\|d_j\|$ are the magnitude of the vector d_q and d_j respectively.

H. Gudgement of Spam SMS

To make final decision, the similarity value of test document and spam SMS center document and the one of test document and white SMS center document is compared. The bigger one indicates the category that the test document belongs to.

V. Experimental Results and Discussion

The Vector Space Model based Spam SMS filter has been deployed in production environment processing 10 million SMS per day.

Before deployment, we perform the following experiment on the filter.

A. Training Corpus

The training corpus is collected from Dahan Tricom database. The corpus in the primary test experiment includes more than 12 thousand spam SMSs and 8 thousand white SMSs. The spam SMSs includes fraud SMS, insulted SMS, malicious advertisements, Rumors, Criminal SMS and other illegal SMS. The white SMS includes advertisement SMS, logistics SMS, verification SMS, financial notification SMS.

TABLE I. The Total Number Of Documents Used For Training and Testing

Data Sets		*Training Documents*	*Testing Documents*
Spam SMSs	fraud SMS	2000	2000
	insulted SMS	2000	2000
	malicious advertisements	2000	2000
	Rumors	2000	2000
	Criminal SMS	2000	2000
	other illegal SMS	2000	2000
	Total	12,000	12,000
White SMSs	advertisement SMS	2000	2000
	logistics SMS	2000	2000
	verification SMS	2000	2000
	financial notification SMS	2000	2000
	Total	8,000	8,000

B. Experiment Design

In the experiment, the test SMSs are the same with the training one in order to exclude unrelated factors.

That is, the filter use training corpus to process and result in center document. Then, the filter judges every same SMS in training corpus to see the precise, recall and F1 result.

C. Experiment result

The results turn out to be the following:

TABLE II. The Number of Correct Incorrect Judge Instance

Data Sets		*Correct*	*Incorrect*	*Total*
Spam SMSs	fraud SMS	1893	107	2000
	insulted SMS	1877	123	2000
	malicious advertisements	1452	548	2000
	Rumors	1887	113	2000
	Criminal SMS	1857	143	2000
	other illegal SMS	1789	211	2000
	Total	10,755	1245	12,000
White SMSs	advertisement SMS	1622	378	2000
	logistics SMS	1902	98	2000
	verification SMS	1739	261	2000
	financial notification SMS	1753	247	2000
	Total	7,016	984	8,000

Apparently, the result for malicious advertisements and normal advertisement does not work well because these SMSs do not differ much. The spam senders disguise these SMS by replacing a small part of the white SMS, such as URL of payment webpage or a bank account. Although the disguise URLs or accounts are being kept collecting, this solution only resolve current issue because the URLs and accounts are kept changing.

Based on the data above, the overall precise, recall and F1 result is:

TABLE III. The Precise, Recall and F1 Result

Data Sets	*Precise*	*Recall*	*F1*
Mechanical & Electronic Engineering	88.89%	88.89%	79.01%

VI. Conclusions

In this paper, a Vector Space Model based Spam SMS Filter is discussed. It addresses the particularity of SMS, such as short, vocal, domain related, etc. This technology considers much about the particularity and apply much modification on the traditional VSM model, such as proper term corpus, expanded corpus, etc.

This technology has been deployed in the production environment of Dahan Tricom Corporation and the results in production environment turns out to be applicable in SMS commercial companies.

References

[1] G. Salton, Automatic Information Organization and Retrieval, McGraw-Hill Book Corporation, New York, 1968.

[2] Zhang, Dongli, et al., Chinese text classification system based on VSM, Journal of Tsinghua University, v43, n9, p1288-1291, September 2003.

[3] Wang, Luda, et al., WordNet-based hybrid VSM for document classification, International Journal of Database Theory and Application, v9, n1, p185-200, 2016.

[4] Abilhoa, Willyan D., De Castro, Leandro N. A keyword extraction method from twitter messages represented as graphs, Applied Mathematics and Computation, v 240, p 308-325, August 1, 2014.

[5] LIU Yun-feng , QI Huan, Xiang'en Hu, Zhiqiang Cai, A Modified Weight Function in Latent Semantic Analysis. Journal of Chinese Information Processiong. vol. 19(6), pp. 64-69, 2005.

[6] S. Robertson, Understanding inverse document frequency: on theoretical arguments for IDF. Journal of Documentation 60, 503-520, 2004.

[7] K. Spärck Jones, IDF term weighting and IR research lessons. Journal of Documentation. 60, 521-523, 2004.

[8] Marvin, S., & Scott, S.. Feature engineering for text classification. In Proceedings of international conference on machine learning, 1999.

[9] Yuan, Man. Ouyang, Yuanxin. Xiong, Zhang. Luo, Jianhui. Short text feature extension method based on frequent term sets. Journal of Southeast University (Natural Science Edition), v 44, n 2, p 256-260, 2014

[10] Wei, Chao. Luo, Sen-Lin. Zhang, Jing. Pan, Li-Min. Short text manifold representation based on AutoEncoder network. Journal of Zhejiang University (Engineering Science), v 49, n 8, p 1591-1599, August 1, 2015