Semi-automated Feature Selection for Web Text Filtering

Ying Chen
Beijing Electronic Science and Technology Institute
Beijing, China
ychen@besti.edu.cn

Ou Wu
NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China.
wuou@nlpr.ia.ac.cn

***Abstract*—The explosive growth of the Internet inevitably leads to the proliferation of harmful information such as pornography, drug and violence. A great deal of filtering techniques based on image and text categorization is proposed in the literature. Among them, text-based filtering plays a leading role for its good performance. Existing text filtering algorithms can be seen as a classical text categorization approach of discerning two topics, i.e. harmful and benign. In this paper, motivated by the linguistic character of text features and other related text classification tasks such as genre detection, a new feature selection framework for text filtering is proposed. It combines linguistics and domain knowledge in an effective way. Experimental results have demonstrated that our method is more adapt to special domain text filtering tasks.**

Keywords-Web filtering; feature selection; semi-automated;

I. INTRODUCTION

Wolak et al. [1] made a survey of a nationally representative sample of 1500 young Internet users in USA. The results show that forty-two percents of them had been exposed to online pornography in the past year. So web information filtering is of great importance. Commercial web-filtering systems mainly use mechanical techniques like Blacklists. Evaluation studies [2] showed that they could not provide satisfactory results in real applications. Researchers have focused on the intelligent filtering techniques in recent years. Because intelligent methods are usually based on large amount of collected harmful/ benign samples, they are more robust than mechanical techniques. From the information mode emphasized, intelligent filtering techniques existing can be roughly divided into text filtering, image filtering, and multi-mode filtering. Among the existing methods, text filtering plays the leading role for two reasons: 1) text is predominant on the web and 2) object recognition is still an open problem in pattern recognition. It is difficult to extract the underlying essential features of harmful images currently. Most existing pornographic image recognition algorithms can only adapt to a certain style of pornographic images and thus lead to high false recognition rate on the whole. Then how to increase the performance of text filtering is essential to decrease the gap between academic research and actual use.

Web text filtering can be viewed as a classical text categorization (TC) task in special domains. There are three key steps in TC: training sample compiling, feature selection and classifier choosing. We first review the related work in text filtering combined with our analyses. We investigate the unbalanced distributions of part of speech of feature words selected by classical methods. Motivated by this phenomenon and recent development in TC such as genre detection and sentiment analysis, we divide the whole feature selection approach into several feature subset selection tasks. In each task, the selected subset can reflect a certain aspect of underlying differences between harmful texts and benign texts. At last, all the subsets are fused to the whole set of features.

The remainder of this paper is organized as follows. Section 2 briefly reviews the feature selection in existing text filtering algorithms and explains the primitive motivation of this study. Section 3 gives our analyses of features for text filtering with our motivation. Section 4 introduces the proposed feature selection approach. Empirical evaluation is given and discussed in Section 5, prior to conclusions in Section 6.

II. WHY SEMI-AUTOMATED?

Because our work is mainly about feature selection, only the feature selection parts of previous work are reviewed. Table 1 summarizes the feature selection parts of several methods proposed in previous literature [3, 4, 5, 6, 14]. It shows that features are usually manually constructed and the dimension is usual no more than the five hundreds while it is over thousands in general TC tasks. Because samples in text filtering are not so diversity as samples in general TC tasks, it is feasible to manually construct the features although it is a tiring job. But why some well known automatically selected features such as IG are not used? The main reason is that automatically selected features overall depend on the training set (it is usually unbalanced and small), so collection a training set reflecting the underlying distribution of harmful/benign texts is the main

978-1-4244-5934-6/10/$26.00 ©2010 IEEE

TABLE 1. FEATURE SELECTION IN EXISTING METHODS

METHODS	*DIMENSION*	*MANUALLY/AUTOMATED*
Lee's [3]	61	Manually
Du's [4]	Not referred	Automated
Hammami's [5]	300	Manually
Hu's [6]	109	Semi-automated

challenge and usually very difficult. For example, if only football documents are collected as positive set and pornography as negative set, many football-related words will be selected as feature words. Manually construction relies on human's inference and usually need a small number of samples. As a result, manually construction seems more reliable than automated feature selection. However, manually compiling is a tiring job and lack of extendibility to other types of harmful content. In conclusion, manually construction has better generalization ability with in the domain but worse extendibility to other domains; automated construction has better extendibility to other domains but worse generalization ability with in the domain (this conclusion is declared in the context of special kinds of text filtering). As a result, combining parts of human analysis with some well-known automated feature selection methods i.e. semi-automated selection may be a more practical way.

III. TEXT FEATURES ANALYSIS

A. Distributions of part of speech

Two automated feature selection methods analyzed and used in our study: Bi-Normal Separation (BNS) [7] and Weighted log likelihood ratio (WLLR) [8]. Both of them show well performances in previous work. They both need to calculate the score of each word appear in the document firstly except the stop words, and then rank the words according to their scores. Top N (N is predetermined manually) words are selected as the feature words. The formula for score calculating of BNS for a word w is as follows:

$$F^{-1}(tpr) - F^{-1}(fpr) \qquad (1)$$

where F is the Normal cumulative distribution function; *tpr* is the ratio of the number of positive samples containing word w to the number of positive samples; *fpr* is the ratio of the number of negative samples containing word w to the number of negative samples.

The formula for score calculating of WLLR for w is as follows:

$$p(w \mid c_i) \log \frac{p(w \mid c_i)}{p(w \mid \neg c_i)} \qquad (2)$$

where c_i denotes the i-th category.

According to PKU standard for Chinese word segmentation and tagging, there are 21 kinds of POS in Chinese texts. We used BNS to select 2000 features from a set of 6444 documents. The distributions of different POS of the features are shown in Figure 1. Top 5 kinds of POS occupy 86% while other 16 kinds occupy 14% totally. Chen [9] made a statistic of POS distributions on large-scale data and showed that the proportion of nouns over the whole document is averagely 32.62% and verbs is averagely 20.90% while each of other POS occupies less than 9% on average. It illuminates us that imbalanced distributes of POS in training set may lead to the imbalanced distributes of POS in features. As a result, some useful feature words can not be selected for their minority in the training set. To alleviate this problem, we resample each document according POS tokens in this study.

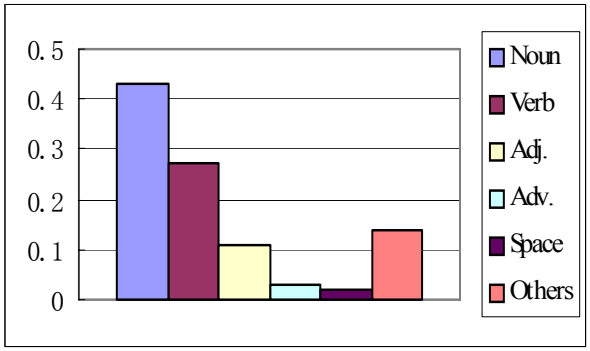

Figure 1. Distributions of part of speech on features

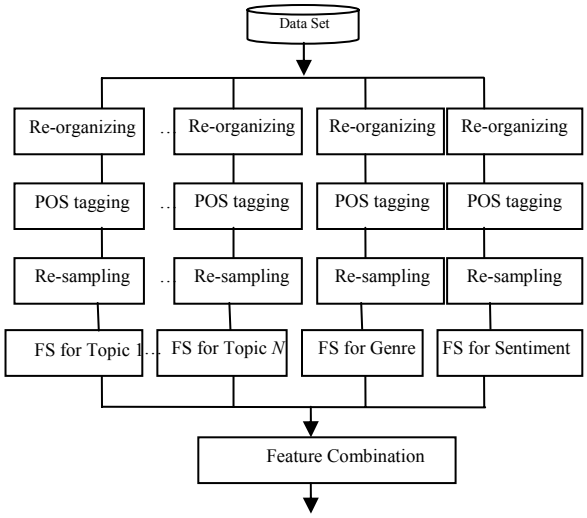

Figure 2. Semi-automated feature selection

B. Non-topical features

In harmful content filtering, the labels are usually viewed as two semantic labels: harmful and benign. Consequently, existing work mainly attempt to select the most effective features such that they can reflect the semantic difference. We call such selected features as topical features because they

directly reflect the topical differences. However, topical features can not discriminate the subtle differences between such as porno documents and sexual health documents, or drug and anti-drug documents. Recent studies such as genre detection, sentiment analysis show that two semantic-similar documents can be different from each other in genre or sentiment. For example, the genres of healthy text and pornographic text may differ a lot. That is to say, although the difference in topics between two documents is small, there may be obvious difference in genre or other non-topic aspects. We define the basic unit of text document as terms which are words, punctuations and POS tokens (actually, they are unseen before POS tagging). We denote the middle-level unit of text document as property such as topic, genre and sentiment. If we have known on which properties the two-class texts differ from each other, we can first extract different feature subset according each property and then combine them as the whole features. Note that the properties of harmful texts are not as diverse as benign ones. They can be easily chosen according to domain knowledge achieved.

Figure 2 shows the main idea of relationships among categories, texts, terms and properties. In our view, categories depend on both topic properties and non- topic properties. Although in the study of ATC topics are regarded as orthogonal to non-topic properties, it is believable that in many text-filtering tasks they are dependent and should be useful for categorization. In the following section, we propose a semi-automated feature selection method based on our view.

IV. SEMI-AUTOMATED FEATURE SELECTION

Once the properties are determined, the feature selection (FS) can be performed as shown in Figure 2, i.e. the whole feature selection task is decomposed into several feature subset selection tasks. The re-organizing step and re-sampling step are specific to the following feature subset selection task. In the FS step for each property, traditional method such as WLLR and BNS are used. Two non-topical properties are taken into consideration: genre and sentiment. Topical properties depend on the filtering task and are manually determined, which can utilize the domain knowledge.

Semi-automated mainly lies in three points: 1) the topical properties in Figure 3 are manually analyzed and 2) samples for each property feature selection are manually constructed by reorganization of the original data set and 3) texts are re-sampled according to POS, which relays on our analysis and appointment. Although they seems subjective, we believe that in real applications it is worthwhile for us to achieve sufficient manually analysis to alleviate the insufficiency of automatically methods. Furthermore, the workload is much smaller than selection features completely manually.

We take the porno text filtering as an example to illuminate our feature selection approach. 'Sex' is certainly the major topic of porno text. Based on our observation and inference, we obtain one topics: 'people'. Then the selected properties are 'sex', 'people', genre, and sentiment.

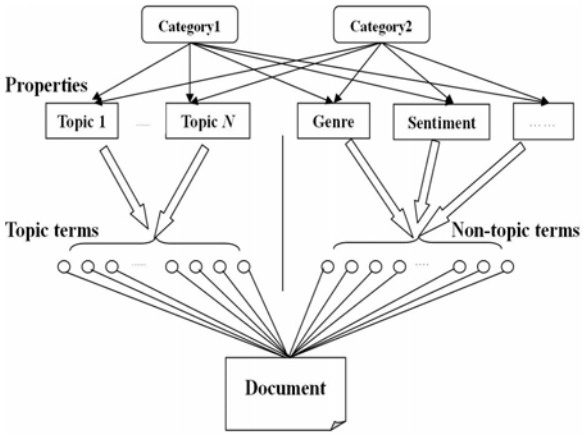

Figure 3. Our view: Terms are determined by properties of document. Categories depend on these properties

- ***Features for 'Sex'***. Note that sex is not equal to porn. Consequently, the training set should be constructed according to sex/non-sex instead of porno/benign. Based on our observation, we found that almost all the sex-related words are just nouns and verbs. Our feature selection for 'sex' is also based on the classical method but with the two distinct points as following: 1) the negative set is composed of both porno ones and legitimate sex-related ones; 2) in the preprocessing step, each document is re-sampled so that only nouns and verbs are remain and other POS are cleaned. Then the selection steps (for Chinese texts) are listed as follows:
  - *Step1*.Prepare for the training set;
  - *Step2*.Segment words for Chinese documents;
  - *Step3*. Resample each document;
  - *Step4*. Use a classical method such as WLLR to select features in the new training set.

- ***Features for 'People'.*** We choose porno documents and some story ones about people as the negative training set and legitimate sexual-related documents and other ones in different topics as the positive set. Only nouns, adjectives and adverbs are taken into account. The steps are as the same as 'sex' feature selection.

- ***Features for Genre.*** It is commonly regarded that the style of a document is orthogonal to topic. Genre reflects a certain style rather than being related to the content [10]. In our study, conjunctions, prepositional, auxiliary words are the candidates. The punctuations considered are: exclamatory mark, interrogation mark, suspension points, comma and full stop. The negative training set is composed by porno documents and the positive set is composed by normal documents. The steps are as the same as feature selection for 'sex'.

- ***Features for Sentiment.*** Adjectives, adverbs, exclamations, onomatopoeic words in the training set are candidates. We choose pornographic documents and some story ones as the negative training set and many other normal ones as the positive training set. The steps are the same as 'sex' property feature selection.

- ***Feature combination.*** Once all the four property features are obtained, they can be integrated as the whole features. This study undertakes a linear integration strategy, i.e. we choose four subsets of features according to the score ranks from the four property features introduced above and combine them as the whole feature set.

V. EMPIRICAL EVALUATION

A. Experiment Setup

To evaluate our approach, 2758 Chinese porno texts, 3042 sex-related benign Chinese texts, and 1073 story benign texts have been collected from the Web. They are equally divided into training part and test part. In order to keep training part as diverse as possible, two public sets are used [11,12]. ICT-CLAS [14] is applied to word segmentation and POS tagging. We used the correct recognition rate (CRR) to evaluate the performance which is defined as the fraction of the false classified samples. SVM (kernel is radial basic function) is applied.

Because in harmful text filtering, the dimension of features are usually small as shown in Table 1, we just extract no more than four hundreds of features for each property. Four feature subsets can be obtained for the four properties using method described in Section 4. For genre, its dimension is 76. For each other, the dimension is set as 500.

B. Results

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

TABLE 2. RESULTS OF TRADITIONAL METHODS

	100	*200*	*300*	*400*
WLLR	.8801	.8865	.8913	.8832
BNS	.8743	.8618	.8852	.8729

Table 2 shows results achieved by WLLR and BNS on the test set when top 100, 200, 300, 400 features are used. When using our framework, we first need to combine the features of four properties. For the 'sex' property, we choose top 20, 100, 200 and 300 features, and for genre property, we choose all of them, and for the other three properties we choose top 0, 10, 20, 50, 100, 200 and 300 from each. The whole features are combined by these three parts. Table 3 shows results achieved by each combination. The leftmost value of each row is the number of 'sex' features and the up most value of each column denotes the combination of genre feature and other three properties. For example, 20+*g* means the entire genre features with 20-dimensional each other three property except 'sex'.

The highest CRR is 0.9487. If the dimensions of features except 'sex' and genre continue to increase, the CRR is decreased. The reason may be that when the dimension of whole feature increases, more training samples are needed. In the combination of (300, *g*+50), the whole feature dimension is 448 (300+76+50*3-68, 68 is the number of redundant feature words) while in the combination of (300, *g*+100) the whole feature dimension is 595. Results achieved by our feature selection framework (BNS-based) is shown in Table 4. The combination (600, g+50) achieves the best result.

C. Discussions

The above test results show our method is encouraging: the best results achieved by WLLR and BNS are 0.8913 and 0.8852 respectively while ours are 0.9487 and 0.9445, respectively. However, there is still something unclearly. For example, how to get the topical properties for new task and why only nouns and verbs are used for 'sex' features selection. How to make them more clearly is our future work. One main reason which hinders achieving better results is that the word segment system can not recognize some sexual slang correctly. We can manually select them or perform a new feature identification approach. The latter choice may be our future work since with the slang and 'sex' feature words, blacklist words which are the keys to commercial systems can be easily obtained.

Feature selection plays an important role in text classification. In Hammami's paper [5], they mentioned that finding keywords or sentence automatically to overcome the drawbacks of laborious building will be their future work. A good feature selection method can make even the simplest classifier model get a satisfied performance through training. However, feature selection is also an open problem. In general ATC such as news categorization, the focus is the evaluation function for candidate feature words. That is partly because the documents to be classified contain so many categories for which it is hard to do some empirical analysis. Consequently, poor prior-knowledge may be obtained. In text filtering tasks, there are usually only two categories: negative (harmful) and positive (normal). And then we can at least make empirical analysis on one of them to obtain some useful prior-knowledge. Take the pornography as an example, we are sure that sexual words are useful and should be the feature words. The prior-knowledge used in this study help us with the construction of the training set and feature selection. The experimental results show the effectiveness of our method. However, there are still something unknown and some empirical assumptions. For example, how to get the useful properties for new filtering tasks and why only nouns and verbs are used for 'sex' feature word selection. How to make

them more clearly to us may be our future work. This study does focus on the key points which obtain less attention in text filtering to the best of our knowledge. And it may be useful for general ATC study. For example, we can first choose feature terms in each POS dimension and combine them as the whole feature, which can lead to better or at least no worse results.

VI. CONCLUSIONS

The goal of feature selection is to find the feature set which can represent the underlying differences between different categories as much as possible. Different from previous work which just extracts topic features, this study has proposed a semi-automated feature selection framework to explore both topical and non-topical features. In additional, the POS are taken into consideration to select more diverse features. We take porno text feature selection as an illustrative example. Four properties are chosen. Empirical evaluation shows that the proposed feature selection method is promising. We believe that it is worthwhile for us to perform a semi-automated approach for the difficulty in high quality data collection.

ACKOWLEDGMENT

This work is partly supported by Foundation of Beijing Electronic Science and Technology Institute Key laboratory of Information Security and Privacy(Grant No. YZDJ0808) and NSFC (Grant No.60903147).

REFERENCES

[1] J. Wolak， K. Mitchell, and D. Finkelhor, "Unwanted and wanted exposure to online pornography in a national sample of youth internet users", PEDIATRICS, Feb. 2007, Vol. 119, No.2.

[2] NetProtect Research Group, "Report in filtering techniques and approaches", NETPROTECT: WP2: 2.3 V1.0 23, Oct. 2001, *Technical Report*.

[3] P.Y. Lee , S. C. Hui, and A. M. Fong, "Neural networks for web content filtering". *IEEE Intelligent Systems*, 2002, 17(5): 48-57.

[4] R. Du et al, "Web filtering using text classification", *Proc. of International Conference on Network*, 2003, pp.325-330.

[5] M. Hammami , C. Youssef, and L. Chen, "A web filtering engine combining textual, structural, and visual content-based Analysis", *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 2, 2006, pp. 272-284.

[6] W. Hu, Ou Wu, Z. Chen, Z. Fu and S. Maybank, "Recognition of pornographic web pages by classifying texts and images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, 29(6):1019-1034.

[7] G. Forman, "An extensive empirical study of feature selection metrics for text classification", *Journal of Machine Learning Research,*, 3 (2003) pp.1289-1305.

[8] V. Ng , S. Dasgupta, and S. M. Arifin, "Examining the role of linguistic knowledge sources in the automatic identification and classification of review", in *Proceedings of the joint conference of COLING/ACL*, 2006*, Poster*.

[9] K. Chen, "Analysis on balance-corpus and text categorization based on large-scale realistic corpora", *Advances in Computation of Oriental Languages*, 2003.

[10] A. Finn and N. Kushmerick, "Learning to classify documents according to genre", *Journal of American Society of Information Science and Technology,*, 2006, Vol.57 (11), pp. 1506-1518.

[11] http://www.sogou.com/labs/dl/t.html.

[12] S. Tan, X. Cheng, M. M. Ghanem, B. Wang, and H. Xu, "A Novel Refinement Approach for Text Categorization", in *Proceedings of ACM international Conference on Information and Knowledge Management*, 2005.

[13] Zhaohui Zheng, Rohini Srihari, Sargur Srihari, "A Feature Selection Framework for Text Filtering," icdm, pp.705, Third IEEE International Conference on Data Mining (ICDM'03), 2003

[14] http://www.i3s.ac.cn

TABLE 3. CLASSIFICATION RESULTS USING THE PROPOSED METHOD (WLLR-BASED)

	0	*10*+G	*20*+G	*50*+G	*100*+G	*200*+G	*300*+G
20	.8367	.8731	.9107	/	/	/	/
100	.9154	.9225	.9321	.9159	.9252	/	/
200	.9221	.9298	.9422	.9414	.9404	.9243	/
300	.9102	.9325	.9446	**.9487**	.9334	.9275	.9354

TABLE 4. CLASSIFICATION RESULTS USING THE PROPOSED METHOD (BNS-BASED)

	Story benign text	Sex-related benign text	Erotic text	All
300	.8866	.9137	.8929	.8962
300+g	.8866	.9413	.9008	.9048
300 +g+10	.9314	.9390	.9170	.9270
450 +g+ 20	.9372	.9415	.9278	.9339
600 +g +50	.9380	.9444	.9482	**.9445**
750 +g+100	.9330	.9537	.9361	.9363