Mobile SMS Classification

An Application of Text Classification

Deepshikha Patel, Monika Bhatnagar

Abstract—**Text Classification is the process of classifying documents into predefined classes based on its content. Text classification is important in many web applications like document indexing, document organization, spam filtering etc. In this paper we analyze the concept of a new classification model which will classify Mobile SMS into predefined classes such as jokes, shayri, festival etc. All sms are converted into text documents. After preprocessing vector space model is prepared and weight is assigned to each term. In the proposed model we have used entropy term weighting scheme and then PCA is used for reparameterization. Artificial Neural Network is used for classification.**

Index Terms— **Text Classification, Short messaging service (sms), feature selection, Principal Component Analysis, Neural Network.**

I. Introduction

A good text classifier is a classifier that efficiently categorizes large sets of text documents in a reasonable time frame and with an acceptable accuracy, and that provides classification rules that are human readable for possible fine-tuning. If the training of the classifier is also quick, this could become in some application domains a good asset for the classifier. Many techniques and algorithms for automatic text categorization have been devised.

The text classification task can be defined as assigning category labels to new documents based on the knowledge gained in a classification system at the training stage. In the training phase we are given a set of documents with class labels attached, and a classification system is built using a learning method. Classification is an important task in both data mining and machine learning communities, however, most of the learning approaches in text categorization are coming from machine learning research. A number of text classification techniques have been applied including, Naive Bayes [3,4] k-NN[5], Neural Network [6], centroid-based approaches [9,10,19,21], Decision Tree [12,20,7], SVM [13,22], Rocchio Classifier [8], Regression Models [11,20], Bayesian probabilistic approaches [17], inductive rule learning [18], and Online learning [14].Although a lot of approaches have been proposed, automated text categorization is still a major area of research primarily because the effectiveness of current automated text classifiers are not faultless and still needs improvement and the time to train a classifier is still very significant.

In this paper, we focus on some well known applications of text categorization and also propose a new model for the classification of text sms into some predefined categories.

II. Related work

Many text classifiers have been proposed in the literature using machine learning techniques, probabilistic models, etc. They often differ in the approach adopted: decision trees, naıve-Bayes, rule induction, neural networks, nearest neighbors, and lately, support vector machines. Although many approaches have been proposed, automated text categorization is still a major area of research primarily because the effectiveness of current automated text classifiers is not faultless and still needs improvement. A classifier is built by applying a learning method to a training set of objects. This model is further used to predict the labels to new incoming objects. With all the effort in this domain there is still place for improvement and a great deal of attention is paid to developing highly accurate classifiers. Refrence[15] classify web news stories based on memory based reasoning. Refrence [16] uses neural network with PCA to classify web news.

III. Applications

There are many potential applications of text classification. A good survey on the methods of text categorization and applications of text categorization can be found in [26]. In this section we examine some of the text classification applications.

A. Document Organization

A news or media company will typically see hundreds and thousands of submissions every day. In order to efficiently handle such vast flow of information, there is a need of an automatic text classification system, which would categorize each document by topics so that they could be sent to the relevant recipient.Maintaining the Integrity of the Specifications.

B. Spam Filtering

Receiving of vast quantities unsolicited junk e-mail, i.e, spam is a big problem. A text classification system could, in the ideal case, categorize incoming messages into genuine and spam categories, rejecting these that it found to be spam.

C. Filtering Pornography Content

As the Internet has rapidly been expanded, we can find information quickly and easily. The exponential increase of information in internet has raised the issue of information security. Pornography web content is one of the biggest harmful resources that pollute the mind of children and teenagers. Several web content

Manuscript received March 20, 2011.

Deepshikha Patel, Information Technology, Rajeev Gandhi Technical University, Bhopal, India, (23.deepshikha@gmail.com).

Monika Bhatnagar, Information Technology, Rajeev Gandhi Technical University, Bhopal, India, (monasikkim@yahoo.com).

classification approaches have been proposed to avoiding these illicit web contents accessing by the children. Text classification controls search results from google, yahoo and other search engines. When used, web sites containing pornography and explicit sexual content can be blocked from google, yahoo and other search engines.

D. Automatic Summary Evaluation

An interesting application of text classification is the ability to evaluate automatically produced summaries of text [28]. It works on the assumption that a summary should compute the most significant features in a document. This idea may be automatically generated from feature vectors.

E. Web Page Prediction

Text classification can be used to predict, which hyperlink on a given web page the user is likely to click on [29]. Each hyperlink text description is treated as a miniature document. Also a text categorization system could be used to naively predict the next page for a fast look-ahead caching system.

F. Identity Based access & reporting

Filtering can be configured to create access policies based on groups, departments, levels in hierarchy or even the individual user. This allow enterprises to create different policies based on work profile to finance, marketing, HR, department or for educational institutions based on academic requirements for students, staff, administrator. In short we can categorize access on the policy “Who is doing what?”

G. Proposed Application: Mobile SMS Classification

Apart from above mentioned applications we can also use text classification for classifying text sms into categories.

IV. PROPOSED MODEL

Figure 1 shows our classification process for Mobile SMS. This process is made up of six main parts; SMS collection, preprocessing, feature selection, term weighting, reparameterization using PCA and Neural Network classification. At first, the collected sms documents are preprocessed as text documents. Text documents would go through feature selection process. Term weights are assigned to each term. Then PCA is applied to reduce dimensions. Then we use these selected features as input to the artificial neural network, which in turn classify the sms in to some well known categories namely jokes, shayri, festival, funny etc. The details of each process of classification system will be discussed in the following section.

A. SMS Documents Retrival

This step gathers various sms documents. We have collected various sms from internet belonging to different categories. Those retrieved sms are stored in the local database for further processing.

B. Preprocessing

At this stage, terms that do not provide any information about class or category selection are to be removed. Two concepts should be introduced here:

1) Stop-word Removal

Stopping is a process of removing most frequent words that exist in a web document by using a stop words dictionary.

2) Word Stemming

Stemming is used for the morphological analysis of words. However stemming reduces the occurrence of term frequency, which has similar meaning in the same document. Porter Stemming is widely used stemming algorithm.

C. Feature Selection

Many feature selection techniques are used in the area of text classification such as mutual information, CHI statistics, Information Gain, Term strength, document frequency, etc. In our classification document frequency thresholding is used for feature selection.

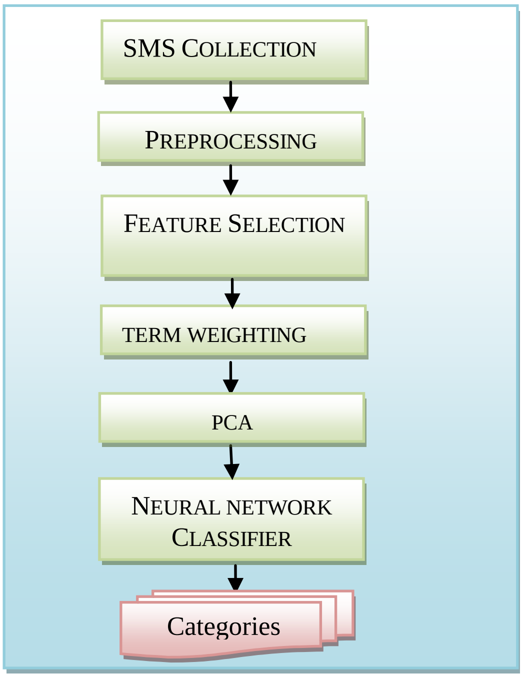

Figure 1: SMS Classification Model

D. Feature Selection

Many feature selection techniques are used in the area of text classification such as mutual information, CHI statistics, Information Gain, Term strength, document frequency, etc. In our classification document frequency thresholding is used for feature selection.

E. Term Weighting

Entropy method is based on a probabilistic analysis of the texts. Entropy believes that significance of term is proportional to the frequency of a term in most documents. Term can be assigned weights according to local term weighting and global term weighting.

F. PCA (Principal Component Analysis)

Using PCA, the dimension reduction process will reduce the original data vector into small number of relevant features [1, 2].

Let M to be the matrix of document terms weights as follows.

$$M = \begin{pmatrix} a_{11}a_{12} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1}a_{n2} & \cdots & a_{nm} \end{pmatrix}$$

Where, a_{ij} refers to the terms in the collection of documents. n is the number of terms and m is the number of documents.

Then we calculate the mean a and subtract it from each data points $a - \bar{a}$. After variance-covariance matrix M can be calculated, where the new value of a_{ij} = ($a_j - \bar{a}$)($a_i - \bar{a}$).Then we determine eigenvalues and eigenvectors of the matrix M which *C* is a real symmetric matrix so a positive real number λ and a nonzero vector α can be found such that, $C\alpha = \lambda\alpha$ where λ is called an eigenvalue and α is an eigenvector of *C* .In order to find a nonzero vector α the characteristic equation $| C - \lambda I |$ must be solved. If *C* is an $n \times n$ matrix of full rank, *n* eigenvalues can be found such that ($\lambda 1, \lambda 2, .., \lambda n$). By using ($C - \lambda I$) $\alpha = 0$, all corresponding eigenvectors can be found. The eigenvalues and corresponding eigenvectors will be sorted so that $\lambda 1 \geq \lambda 2 \geq ... \geq \lambda n$. Then we select the first $d = n$ eigenvectors where *d* is the desired value.

V. NEURAL NETWORK AS CLASSIFIER

For many years, there was no theoretically sound algorithm for training multilayer artificial neural network. Since single layer network proved severely limited in what they could represent, the entire field went into virtual eclipse. The resurgence of interest in artificial neural network began after the invention of back propagation algorithm. In OWPCM, most popular artificial neural network (ANN) architecture, the Multilayer Feedforward (MLFF) network with back propagation (BP) learning is used. This type of network is sometimes called multilayer perceptron because of its similarity to perceptron network with more than one layer.

Back propagation learning

Back propagation neural network employ one of the most popular neural network learning algorithms, the *Back propagation (BP) algorithm.*

The basic algorithm loop structure is given as:

```
Initialize the weights

Repeat
  For each training pattern
    Train on that pattern
  End

Until the error is acceptably low
```

VI. CONCLUSION

In this paper we analyze the concept of a new classification model to classify sms and also applications of text classification. Current classification system has achieved a greate success and it is clear that more can and should be done in the area of text classification. All approaches to the text classification have their own importance and new approaches are also developing as per the requirements. We expect that this model will be successful in efficiently classifying sms text documents. Implementation of this model will be further in future.

REFRENCES

[1] A. Selamat, 2003. Studies on Mobile Agents for Query Retrieval and Web Page Categorization Using Neural Networks, in Division of Computer and Systems Sciences, Gradute School of Engineering, vol. Doctoral. Osaka: Osaka Prefecture University , pp. 94.

[2] R. A. Calvo, M. Partridge, and M. A. Jabri, 1998. A Comparative Study of Principal Component Analysis Techniques, presented at In Proc. Ninth Australian Conf. on Neural Networks, Brisbane

[1] P. Frasconi, G. Soda and A. Vullo. "Text categorization for multi page document: a hybrid naive Bayes HMM approach*", In proceeding of 1st ACM/IEEE-CS joint conference on Digital libraries; ACM Press New York, NY, USA*, pages 11-20. 2001

[2] A.M. Kibriya, E. Frank, B. Pfahringer and G. Holmes. "Multinomial naive bayes for Text categorization" revisited. *AI 2004: Advances in Artificial* Intelligence, 3339, pp. 488–499, 2004.

[3] G. D. Guo, H. Wang, D. Bell, Y. X. Bi, and K. Greer."Using kNN model for automatic text categorization". *Soft Computing, 10(5), pp.* 423–430, 2006.

[4] R. N. Chau, C. S. Yeh, and K. A. Smith. :"A neural network model for hierarchical multilingual text categorization". *Advances in Neural Networks, LNCS, 3497, pp.* 238–245, 2005.

[5] S. Gao, W. Wu, C. H. Lee, and T. S. Chua. "A maximal .gure-of-merit (MFoM)-learning approach robust classifier design for text categorization*". ACM Transactions on Information Systems, 24(2*), pp. 190–218, 2006.

[6] R. Schapire, Y. Singer, and A. Singhal. "Boosting and Rocchio applied to text clustering*". In Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia*, pp. 215–223, 1998.

[7] A. Broder, M. Fontoura, E. Gabrilovich, A. Joshi,V. Josifovski, and T. Zhang. "Robust classi.cation ofrare queries using web knowledge". *In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands*, pp. 231–238, 2007.

[8] Z. Cataltepe and E. Aygun. "An improvement of centroid-based classi.cation algorithm for text classification*". IEEE 23rd International Conference on Data Engineering Workshop*, 1-2 pp. 952–956, 2007.

[9] D. Lewis and J. Catlett. "Heterogeneous uncertainty sampling for supervised learning". *In Proceedings of the Eleventh International Conference on Machine Learning*, pp. 148–156, 1994.

[10] S. Dumais and H. Chen. "Hierarchical classification of Web content". *In Proceedings of the 23rd Annual International ACM SIGIR conference on Research and Development in Information Retrieval, Athens, Greece,* pp. 256–263, 2000.

[11] David D.Lewis, Robert E. Schapire, James P. Callan, nad Ron Papka. "Training algorithms for linear text classifiers". *IN SIGIR'96: Proceeding of the 19th Annual International AGM SIGIR Conference on Research and Development in Information Retrieval*, pp.298-306,1996.

[12] Makato Iwayama and Takenobu Tokunaga. "Cluster based text categorization: a comparison of category search strategies". *In proceedings of the 18th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'95),* pp. 273-281,1995.

[13] D.D. Lewis, "Naive (bayes) at forty: The independence assumption in information retrieval". *In 10th European Conference on Machine Learning (ECML-98)*, pp. 4-15, 1998.

[14] A. McCallum and K. Nigam. "A comparison of event models for naive bayes text classification". *In AAAI-98 Workshop on Learning for Text Categorization*, 1998.

[15] B .Mansand, G. Linoff, and D. Waltz. "Classifying news stories using memory based reasoning*". In 15th ANN Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'92*), pp. 59-64, 1992.

[16] Ali Selamat, Hidekazu Yanagimoto and Sigeru Omatu," Web News Classification Using Neural Networks Based on PCA,"*SICE02-0163.*

[17] S. Weiss, C. Apte, F. Damerau, D. Johnson, F. Oles, T. Goetz, and T. Hampp."Maximizing text-mining performance". IEEE *Intelligent Systems*, pp. 63–69, 1999.

[18] S. Tan. "An improved centroid classifier for text categorization. "Expert Systems with Applications", 35(1-2): pp. 279–285, 2008.

[19] R. Klinkenberg and T. Joachims. "Detecting Concept Drift with Support Vector Machines". *In Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 487–494, 2000.

[20] Y. Yang. Expert network:" Effective and efficient learning from human decisions in text categorization and retrieval". *In 17th ANN Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94)*, pp. 13-22, 1994.