Question Categorization Using SVM Based on Different Term Weighting Methods

Priyanka G Pillai
Department of Computer Science and Engineering
Amrita Vishwa Vidyapeetham
Amritapuri
priyanka5488@gmail.com

Jayasree Narayanan
Department of Computer Science and Engineering
Amrita Vishwa Vidyapeetham
Amritapuri
jayasreen@am.amrita.edu

Abstract— **This paper deals with the performance of Question Categorization based on four different term weighting methods. Term weighting methods such as tf*idf, qf*icf, iqf*qf*icf and vrf together with SVM classifier were used for categorization. From the experiments conducted using both linear and nonlinear SVM, term weighting method iqf*qf*icf showed better performance in question categorization than other methods.**

Keywords- term weighting;question categorization;SVM;

I. INTRODUCTION

Question Categorization is the task of automatically organizing questions into appropriate categories based on machine learning techniques. Questions are assigned to categories such as education, sports, music etc. Term weighting methods helps to convert questions into a form readable by classifiers. Weights will be assigned to terms in each question so as to represent it as a vector. Processes such as tokenization, stop-words removal, stemming and vector creation are done to create vector space models. Terms which indicate the type of questions such as what, when, where, why, which etc are considered as stop-words and ignored in question categorization. Each of the term weighting methods result in a unique vector space model. One of the most commonly used term weighting method tf*idf which showed excellent results in document categorization is used so as to check its performance in question categorization. Together with this, other term weighting methods qf*icf, iqf*qf*icf and vrf which were specially designed for question representation are investigated. The aim is to find which term weighting method gives effective results for question categorization. In question categorization both the training and testing data are questions. Linear and nonlinear type of SVM is considered for classification.

II. RELATED WORKS

Term weighting has proven to an effective way to improve the performance of text categorization. However not much work have been done to investigate whether the existing term weighting methods perform consistently in question categorization as they do in text categorization [2] [3] [4] [5] [6]. There are much well known term weighting methods in information retrieval and they can be categorized into two: unsupervised and supervised. In unsupervised term weighting method membership of training documents in categories is not taken into account when weighting the terms [3]. Whereas in supervised term weighting methods membership of training documents in categories is taken into account when weighting the terms [9] [10]. One of the most popular unsupervised term weighting scheme, tf*idf, first proposed in information retrieval has been successfully utilized in text categorization [7] [8]. Debole and Sebastiani proposed the concept of supervised term weighting scheme [2]. Later on several researchers, such as Soucy and Mineau [11] and Lan et al. [12], proposed new supervised term weighting schemes. Most supervised term weighting schemes are borrowed from feature selection methods since feature selection methods includes assigning different scores to the terms to measure

their contribution to categorization [13]. Quan, Liu and Qiu focused their studies on question categorization and proposed three new supervised term weighting schemes [1]. They also conducted experiments on existing unsupervised and supervised term weighting schemes to check which one will give better performance for question categorization. Quan, Liu and Qiu concluded that from the existing term weighting schemes tf*OR is the most significant one for question categorization [1]. Also from the newly proposed term weighting methods, iqf*qf*icf and vrf were proven effective for long document categorization.

III. TERM WEIGHTING METHODS FOR QUESTION CATEGORIZATION

Two steps are needed for the construction of a classifier: question representation and classifier learning. Term weighting is an important component for question representation which assigns different weights to the terms in a question.

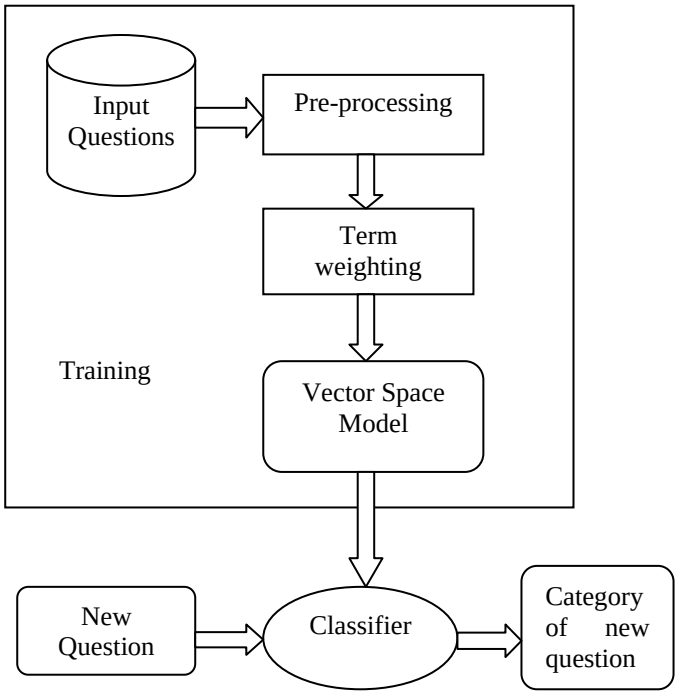

Figure 1: Flowchart of the system

A. UNSUPERVISED TERM WEIGHTING METHOD

*tf*idf*

tf*idf [3] is the most state-of art unsupervised term weighting method. In tf*idf more appearances of a term in a document should be more importance than less appearing terms. Also rare terms are given greater scores since they are more effective for discriminating between documents. tf * idf weight of a term j in a document i can be derived by:

$$w = tf * idf = tf_{ij} * \log\left(\frac{N}{df_j}\right) \quad (1)$$

where, tf is the term frequency, N is the total number of documents in the collection, dfj is the document frequency of term j, i.e., the number of documents in which term j occurs.
In information retrieval tf*idf gives better results when compared to its performance in question categorization. As in question categorization, each category consists of some words which usually occur in many questions of

that category. These words can be good discriminators for the category and should be assigned with larger weights, whereas the idf will treat these terms as insignificant terms and reduce their weight accordingly.

B. SUPERVISED TERM WEIGHTING METHOD

*qf*icf*

qf*icf [1] is one of the supervised term weighting methods where membership of training documents in categories is taken into account. When weighting a term for a question, the current category of the question is assumed as the positive category and all other categories are treated as one negative category. Terms that are present in more number of questions in the positive category but appear in very few other categories will be assigned large weights. If tp is the question frequency of the term in the positive category, cf is the category frequency of the term and C is the number of categories in the corpus, qf * icf weight of a term can be calculated using the equation:

$$qf * icf = log(tp + 1) * log\left[\left(\frac{|C|}{cf}\right) + 1\right] \quad (2)$$

*iqf*qf*icf*

The number of questions in which a term occurs in the negative category is considered in iqf*qf*icf [1] whereas it is ignored in qf*icf. If question frequency of a term in negative category (fn) is more than positive category then it should be given less weights. It can be calculated using the following equation:

$$iqf * qf * icf = log\left(\frac{N}{t_p + f_n}\right) * log(tp + 1) * log\left(\frac{|C|}{cf} + 1\right) \quad (3)$$

vrf

In vrf [1], frequency of a term in the positive category is compared with that in the negative category. It can be calculated using following equation:

$$vrf = \frac{log(tp+1)}{log(fn+1)} \quad (4)$$

IV. EXPERIMENTS

A series of experiments were conducted to study the performance of the above four term weighting methods.

A. *Question Collection*

Questions belonging to four different categories are selected from various sources. Each category will be uniquely labelled. Punctuations, numbers and stop-words are removed from these questions. All letters are then converted to lowercase and stemming is performed. After performing these preprocessing steps, term weighting will be applied and a dataset consisting of question vectors will be generated. Dataset can be divided into two set: training set and testing test. In our experiments we used dataset of 4000 questions out of which 2000 questions were used for training the classifier and 2000 questions for testing. Each term weighting scheme generates a unique dataset.

B. *Support Vector Machine*

To the aim of evaluating term weighting methods for question categorization, support vector machine (SVM) is selected. SVM has the ability to efficiently handle high dimensional and large scale datasets without decreasing classification accuracy. The effectiveness of SVM depends on the selection of kernel and kernel's parameters .Both linear and nonlinear SVM are used in experiments. In our paper the kernel used for nonlinear classification is Gaussian radial basis function (RBF). The software used here for SVM classification is LIBSVM [15].

C. *Performance and Results*

Performance of each of the term weighting schemes in question categorization is evaluated based on F1 measure. To calculate F1, precision and recall is to be calculated [13]. Precision and Recall of categorization to a particular category is calculated as follows:

$$Precision = \frac{Questions\ correctly\ classified\ into\ this\ category}{Total\ questions\ classified\ into\ this\ category}$$

$$Recall = \frac{Questions\ correctly\ classified\ into\ this\ category}{Total\ correct\ questions\ in\ this\ category}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

Table 1 and Table 2 shows the performance of term weighting methods based on precision and recall using different kernels of SVM: LINEAR and RBF respectively. The F1 measure of term weighting scheme iqf*qf*icf is larger for question categorization using both linear and RBF SVM. However when comparison is done between linear and RBF SVM, linear SVM gives better results than RBF SVM.

TABLE 1
Performance of different term weighting schemes on linear SVM

MEASURE	**tf*idf**	**qf*icf**	**iqf*qf*icf**	**vrf**
PRECISION	0.939	0.950	0.963	0.95
RECALL	0.9375	0.950	0.950	0.925
F1	0.938	0.950	0.960	0.937

TABLE 2
Performance of different term weighting schemes on nonlinear SVM

MEASURE	**tf* idf**	**qf*icf**	**iqf*qf*icf**	**vrf**
PRECISION	0.897	0.555	0.906	0.859
RECALL	0.825	0.888	0.850	0.675
F1	0.859	0.842	0.877	0.755

From figure 2 it is evident that the term weighting method iqf*qf*icf shows more categorization accuracy than other three methods in both linear and nonlinear SVM. Linear kernel outperforms RBF kernel in the

performance of all the four term weighting methods. By using linear SVM much faster training and testing speed can be achieved.

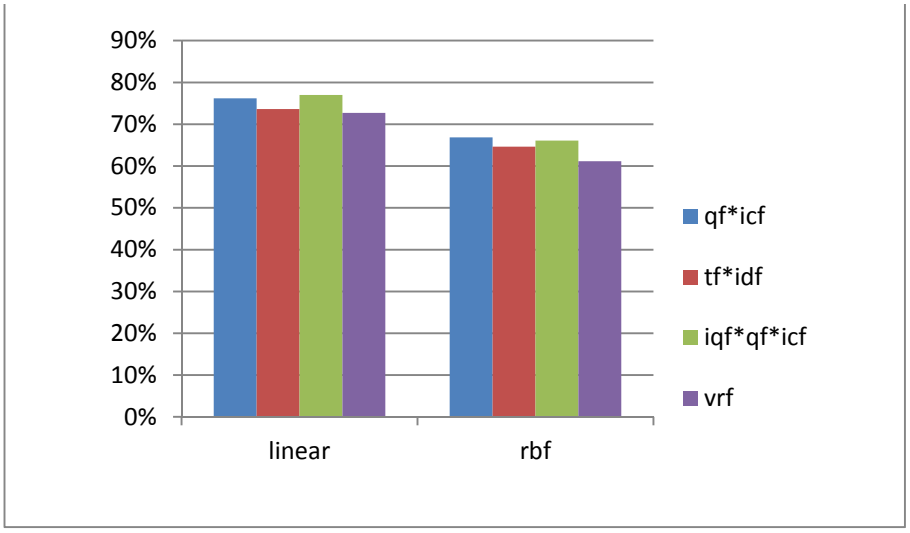

Figure 2: Categorization accuracy of term weighting methods using Linear and RBF SVM

Kernel gamma parameter controls the shape of the separating hyperplane. Increasing gamma usually increases number of support vectors. For small values of gamma the decision boundary is nearly linear. As gamma increases flexibility of the decision boundary increases. Large values of gamma lead to overfitting. Figure 3 shows the performance of term weighting methods on different values of the gamma parameter of RBF SVM and from this we can observe that performance increases gradually as gamma value moves from 0 to 30 after that a constant performance is shown.

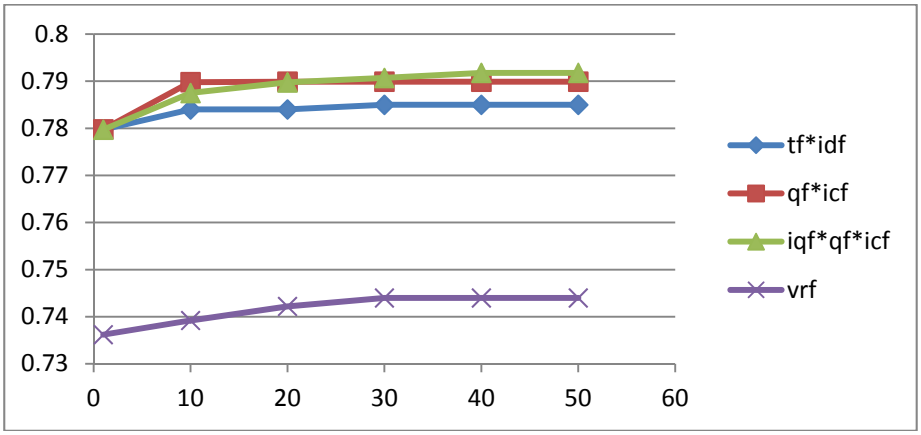

Figure 3. Performance of term weighting methods on different values of gamma parameter using RBF kernal SVM

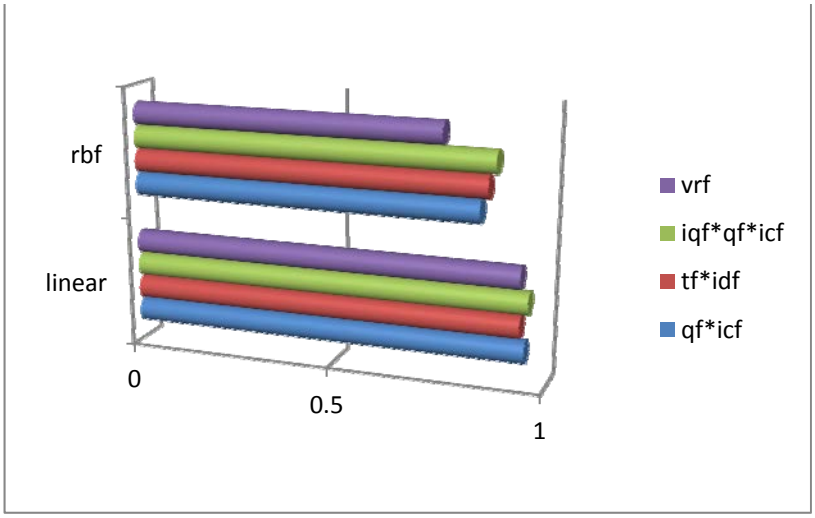

Figure 4. F1 measure of SVM with different kernals

The results of SVM with different kernels, LINEAR and RBF are depicted in figure 4. LINEAR kernel outperforms RBF kernel in the performance of all the four term weighting methods. Term weighting method iqf*qf*icf shows the best results in question categorization using SVM. qf*icf and tf*idf also gives equally good results. When compared with term weighting methods tf*idf, qf*icf and iqf*qf*icf, least performance in question categorization is given by vrf.

V. CONCLUSION

Term weighting schemes improves the performance of question categorization. Through a series of evaluations in question categorization we find that performance of term weighting methods varies significantly. From our study it can be concluded that linear SVM outperforms non-linear SVM on question categorization. Term weighting method iqf*qf*icf exhibit consistently good performance over other methods in question categorization.

ACKNOWLEDGMENT

We would like to acknowledge and extend our heartfelt gratitude to Dr. Ramachandra Kaimal of Computer Science Department, Amrita School of Engineering for his motivation and direction towards the preparation of this paper. We would like to express our gratitude to Amrita School of Engineering, Computer Science Department for providing us with facilities to complete our project.

REFERENCES

[1] X.Quan, W.Liu, and B. Qiu, “Term Weighting Schemes for Question Categorization,”IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 33, no. 5,pp. 1009-1021, May. 2011, doi:10.1109/TPAMI.2010.154.

[2] F. Debole and F. Sebastiani, “Supervised Term Weighting for Automated Text Categorization,” Proc. ACM Symp. Applied Computing, pp. 784-788, 2003.

[3] M. Lan, C.L. Tan, J. Su, and Y. Lu, “Supervised and Traditional Term Weighting Methods for Automatic Text Categorization,”IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 31, no. 4,pp. 721-735, Apr. 2009, doi:10.1109/TPAMI.2008.110.

[4] Z.-H. Deng, S.-W. Tang, D.-Q. Yang, M. Zhang, L.-Y. Li, and K.Q.Xie, “A Comparative Study on Feature Weight in Text Categorization,”Proc. Asia Pacific Web Conf. ’04, pp. 588-597, 2004.

[5] F. Sebastiani, “Machine Learning in Automated Text Categorization,”ACM Computing Surveys, vol. 34, no. 1, pp. 1-47, 2002.

[6] T. Joachims, “Text Categorization with Support Vector Machines:Learning with Many Relevant Features,” Proc. 10th European Conf.Machine Learning, pp. 137-142, 1998.

[7] H. Wu and G. Salton, “A Comparison of Search Term Weighting:Term Relevance versus Inverse Document Frequency,” Proc. ACM SIGIR ’81, pp. 30-39, 1981.

[8] Y. Yang and C.G. Chute, “An Example-Based Mapping Method for Text Categorization and Retrieval,” ACM Trans. Information Systems, vol. 12, no. 3, pp. 252-277, 1994.

[9] Y. Yang, “An Evaluation of Statistical Approaches to Text Categorization,” J. Information Retrieval, vol 1, nos. 1/2, pp. 67-88, 1999.

[10] Y. Yang and X. Liu, “A Re-Examination of Text Categorization Methods,” Proc. ACM SIGIR ’99, pp 42-49, 1999.

[11] P. Soucy and G.W. Mineau, “Beyond tfidf Weighting for Text Categorization in the Vector Space Model,” Proc. Int’l Joint Conf.Artificial Intelligence, pp. 1130-1135, 2005.

[12] M. Lan, C.L. Tan, and H. Low, “Proposing a New Term Weighting Scheme for Text Categorization,” Proc. 21st AAAI Nat’l Conf. Artificial Intelligence, pp. 763-768, 2006.

[13] Y. Yang and J.O. Pedersen, “A Comparative Study on FeatureSelection in Text Categorization,” Proc. 14th Int’l Conf. Machine Learning, pp. 412-420, 1997.

[14] T. Joachims, “Text Categorization with Support Vector Machines:Learning with Many Relevant Features,” Proc. 10th European Conf.Machine Learning, pp. 137-142, 1998.

[15] C. Chang and C. Lin, “LIBSVM: A Library for Support Vector Machines,” http://www.csie.ntu.edu.tw/cjlin/libsvm/, 2001.