TREC-11 Experiments at NII: The Effect of Virtual Relevant Documents in Batch Filtering

Kyung-Soon Lee, Kyo Kageura, Akiko Aizawa

NII (National Institute of Informatics)
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan
{kslee, kyo, akiko}@nii.ac.jp
http://research.nii.ac.jp/~{kslee, kyo, akiko}

1 Introduction

In machine learning techniques, many researches have shown the effectiveness according to training examples by sampling from training set and incorporating prior knowledge into training set.

Researches on document retrieval, text categorization and routing have shown the effects of learning by sampling relevant documents or non-relevant document from training set. Allan et al. (1995) considered only the top K non-relevant documents, which is the same number of all known relevant documents in the training set to learn a routing query. This is motivated by the need to have a balance between the number of the relevant and the negative documents in Rocchio's learning. Singhal et al. (1997) selectively used the non-relevant documents that belong to a query's domain to learn the feedback query. Kwok and Grunfeld (1997) selected the best training subset of the relevant documents for creation of a feedback query based on genetic algorithm. Most sampling techniques in machine learning aim at the reducing the size of training set.

On the other hand, many machine learning applications on image recognition, image classification and character recognition have incorporated prior knowledge about the desired behavior of the system into training data. Prior knowledge is information for the learning which is available in addition to the training examples and makes it possible to generalize from the training examples to novel test examples (DeCoste and Schölkopf, 2002). For example, image recognition system uses new examples by small distortions of the input image such as translations, rotations, scaling; speech recognition system produces those by time distortions or pitch shifts. In 3D object recognition problem, Poggio and Vetter (1992) exploited appropriate transformations to generate new views from a single 2D view. In handwritten digit recognition, DeCoste and Schölkopf

(2002) added virtual examples generated by simply shifting the images by one pixel in the four principal directions to the training examples. In incorporating prior knowledge, the open issue is to what extent transformations can be safely applied during training since some distortions can lead to significantly worse errors (DeCoste and Schölkopf, 2002).

In TREC-11 batch filtering, we have incorporated prior knowledge called *virtual relevant documents* to training documents by combining each two relevant documents pair and giving distinct weight for co-occurring terms on assumption that they might be related to the topic. Support vector machine (SVM) was used to learn decision boundary for the artificially enlarged training documents.

2. Virtual Relevant Documents at Batch Filtering

Intuitively, a document produced by concatenating two relevant documents will be relevant to the topic since one large size of document can be divided into two documents while preserving the topic. And relevant documents will share terms which describe the topic. This characteristic has been used in feature selection. Therefore, prior knowledge generated by multiplying weights of a term which is co-occurring in each two relevant documents pair will provide new information about the decision boundary for classification.

2.1 Virtual Relevant Documents

A document is represented as a weight vector, $di = <w_1, w_2, \ldots, w_k, \ldots, w_n>$. The weight is calculated by LogTF, IDF and cosine normalization.

A virtual relevant document (VRD) is generated by combining two relevant documents in training documents. For *n* relevant documents, n*(n-1)/2 documents are produced:

$$ {}_nC_2 = \frac{n \cdot (n-1)}{2 \cdot 1} \quad \textbf{(1)} $$

The weight of term which occurs in two relevant documents is calculated by multiplying two weights of a term of each vector. The weight of a VRD is calculated as follows:

$$ w_{vij_k} = w_{di_k} \cdot w_{dj_k} \quad \textbf{(2)} $$

where vij_k is the term k of a VRD, di_k and dj_k is the term k of relevant document di and dj, respectively. If the term k does not occur in one document of two relevant documents, the weight is assigned as minimum value instead of zero value and is multiplied to keep the term's existence. Finally, the weight vector of terms is normalized by cosine normalization.

The effect of VRD is that if two relevant documents do not have any sharing term, the resulting VRD become generalized vector of two documents. If two relevant documents share common terms, the resulting VRD would represent strong indicator of relevance to the topic for co-occurring terms. In case of general terms which are not related to the topic, they will have low value by idf in basic vector representation. Therefore, their effects would not be strong.

2.2 Support Vectors

Given training documents which include not only training documents but also VRDs, we have used support vector machine (Vapnik, 1995). Support vectors (SVs) are essential subset of relevant and non-relevant examples in training set. They represent the whole training examples. In test phase, SVs are used for determining on which side of the decision boundary.

Scholkopf et al. (1995) and Vapnik (1995) observed that the SV set contains all information necessary to solve a given classification task. In handwritten digit recognition task, DeCoste and Schölkopf (2002) showed that it is sufficient to generate virtual examples only from the support vectors.

3. Experiments

3.1 Experimental Procedure

In the runs (kNII11bf1 and kNII11bf2) submitted to TREC-11 batch filtering, VRDs are generated from whole relevant documents in training set (VRDs_TRs). In additional experiments, VRDs are generated from support vector set obtained after training SVM for training set (VRDs_SVs). In this paper, we compare two incorporating methods for batch filtering.

We used SVM^{light} system (Joachims, 1999), and trained classifiers via radial-basis function (RBF) kernels and left all SVM^{light} options that affect learning as their default value.

3.2 Results

3.2.1 Submitted Runs

We have submitted two runs tagged kNII11bf1 and kNII11bf2. In the submitted runs, VRDs were generated from whole training documents. In kNII11bf1, VRDs were generated by multiplying weights from two relevant documents, and subtracting terms in non-relevant documents from in relevant documents. It also included virtual non-relevant documents produced by averaging weights from two non-relevant documents, and subtracting terms in relevant documents from in non-relevant documents. In kNII11bf2, VRDs were generated by multiplying weights.

For evaluation measure T11U and T11F, refer TREC-11 filtering track guideline. For the assessor topic, the performances of kNII11bf1 and kNII11bf2 on MeanT11U are 0.305 and 0.302, respectively. The performances of kNII11bf1 and kNII11bf2 on MeanT11F are 0.190 and 0.188, respectively. The results of two runs are almost similar. It means that virtual non-relevant documents do not affect the performance.

3.2.2 Additional Experimental Results

We have compared the effectiveness for VRDs generated from different sources:

- Org training set: performance of SVM for training set.
- VRDs_SVs: the performance of SVM after incorporating prior knowledge generated from relevant support vector set into support vector set.

Table 1 shows the performance for assessor topics (from R101 to R150).

Table 1. The performances on original training set and incorporating VRDs into SV set.

Evaluation Measure	Org training set	VRDs_SVs	Performance change
Mean T11U	0.359	0.376	4.7%
Mean T11F	0.090	0.190	111.1%
Avg. Precision	0.269	0.400	48.7%
Avg. Recall	0.046	0.101	119.6%
Micro-avg. F1	0.181	0.310	71.3%

Table 2. The statistics of information used in support vector learning (RDs: relevant documents, NRDs: non-relevant documents).

	Org training set	VRDs_SVs
Avg # of training documents	861.48	328.24
Avg # of relevant documents	12.78	216.08
(Avg # of VRDs)	-	(204.92)
Avg # of SV set	123.32	119.44
(Avg # of SVs taken from VRDs)	-	(15.32)
(Avg # of SVs taken from RDs in SV set)	-	(10.64)
(Avg # of SVs taken from NRDs in SV set)	-	(93.48)

Table 2 shows the statistics of information in learning process for assessor topics. A lot of relevant SVs included in the new support vectors are taken from VRDs generated artificially, rather than original relevant documents. And the size of SV set learned from VRDS_SVs are similar with that learned from original training set.

In the experimental results, the proposed method achieved a significant performance improvement on the overall evaluation measures. These results indicate that our VRDs give new information to learn decision

boundary in SVM. It is only 47 topics among the total 50 topics that VRDs_SVs improved performance compared to Org training set. Therefore, VRDs generated by multiplying two relevant documents can be applied to transformation in batch filtering task.

4 Discussion

In TREC-11 batch filtering, we have incorporated virtual relevant documents to training documents by combining each two relevant documents pair on assumption that they might be related to the topic. Support vector machine was used to learn from the artificially enlarged training documents. By adding virtual relevant documents generated by transformation of original documents to training set, we could improve performance significantly. However, the base performance of SVM on the training set is low for TREC-11 test collection. For many topics, SVM system classified test documents as non-relevant to the many topics. For future work, VRDs can be applied in other classification model and adapted new virtual transformation.

References

Allan, J., Ballesteros, L., Callan, J., Croft, W., and Lu, Z. (1996) Recent experiments with INQUERY. In Proc. of the Fourth Text REtrieval Conference (TREC-4).

DeCoste, D. and Schölkopf, B. (2002) Training invariant support vector machines. *Machine Learning 46(1),* pp.161-190.

Joachims, T.(1999) Making large-scale support vector machine learning practical. In *Advances in Kernel Methods: Support Vector Machines* (Schölkopf et al., 1999), MIT Press.

Kwok, K. and Grunfeld, L. (1997) TREC-5 English and Chinese retrieval experiments using PIRCS. In the Proc. of the Fifth Text REtrieval Conference (TREC-5).

Poggio, T. and Vetter, T. (1992) Recognition and structure from one 2D model view: observations on prototypes, object classes and symmetries. A.I. Memo No. 1347, Artificial Intelligence Laboratory, Massachusetts Institute of Technology.

Schölkopf, B., Burges, C., and Vapnik, V. (1995) Extracting support data for a given task. In Proc. of the First International Conference on Knowledge Discovery & Data Mining, Menlo Park. AAAI Press.

Singhal, A., Mitra, M., and Buckley, C. (1997) Learning routing queries in a query zone. In *Proc. of the Twentieth ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 21-29.

Vapnick, V. (1995). *The Nature of Statistical Learning Theory*, Springer-Verlag, New York.