PSEUDO RELEVANCE FEEDBACK WITH INCREMENTAL LEARNING FOR HIGH LEVEL FEATURE DETECTION

Shaoxi Xu[1,2]*, Sheng Tang*[1]*, Jintao Li*[1]*, Yongdong Zhang*[1]

[1]Center for Advanced Computing Technology Research
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2]Graduate University of Chinese Academy of Sciences, Beijing, China

ABSTRACT

Pseudo Relevance Feedback (PRF) has shown effective performance in information retrieval, but it has seldom been applied in the area of high level feature detection (HLF). In this paper, we explicitly propose to introduce PRF into HLF. Our contributions mainly lie in two-fold: (1) proposing three novel PRF approaches to extract pseudo positive samples, i.e., Nearest-Neighbor (NN) based PRF, Score-Evaluation (SE) based PRF and Multi-Classifier Decision (MCD) based PRF; (2) utilizing incremental learning to reduce the re-training time. We evaluate our approaches on the benchmark of TRECVID2008. Reported results have shown that MCD based approach outperforms the other two and obtain an excellent gain in average precision with respect to the baseline without PRF.

Index Terms—HLF, PRF, Incremental Learning

1. INTRODUCTION

PRF is one of the most effective techniques for improving the accuracy of ad hoc information retrieval and has been shown to be effective across all retrieval models [14]. This technique was originally applied in the area of text document retrieval. The basic idea of PRF is to assume a set of top-rank documents in the initial retrieval to be relevant to the query and then learn from these pseudo feedback samples to expand or re-weight terms in the original query. Through a query expansion or re-weighting, some relevant documents missed in the initial retrieval can be picked up in the second round and thus the retrieval performance can be improved to a certain extent.

However, due to the poor performance of current information retrieval algorithms, it is a common observation that the top-retrieved documents contain noise and even most of them are non-relevant to the query. Therefore, one of the main tasks of PRF is to separate relevant documents and terms from non-relevant ones in the top-rank lists to emphasize core topics. Many previous researches [4,9,13,16] have been done in this aspect. A cluster-based resampling method proposed by K. S. Lee et al [9] focus on the effects of resampling the top-retrieved documents to extract dominant documents which appear in several different clusters. The essential point of this approach is based on the hypothesis that a good representative document for a query may have several nearest neighbors with high similarity distributed in many clusters. A term classification method proposed by G. Cao et al [4] is used to select good expansion terms directly according to their possible impact on the retrieval effectiveness and to provide a framework for incorporating different source of evidence. Both of work conducted by K. S. Lee et al [9] and by G. Cao et al [4] are based on a language model for retrieval (the former is query-likelihood retrieval [7] while the latter is KL divergence-based retrieval model [1]) with Dirichlet smoothing [2] and assume that expansion terms are sampled from a relevance model [15]. Another sampling method by T. Saikai [13] skips some top-retrieval documents by selective sampling to adjust the number of pseudo relevant documents and/or the number of expansion terms for each topic. The essence of the algorithm is that it tries to avoid collecting too many documents with the same set of query terms based on the assumption that the initial top-ranked documents may be too similar or redundant. On the other hand, X. Huang et al investigate data mining methods for PRF to promote retrieval performance [16]. They use text classification and co-training techniques to identify more relevant passages from PRF of a retrieval system.

The applications of PRF mentioned above are all those in text information retrieval. Research on the application of PRF in image retrieval[6,10,11] has been also carried out. However, the structure of image examples is quite different from that of documents, so some additional techniques should be explored for PRF in image retrieval. Firstly, R. Yan et al suggest in [10, 11] that using lowest ranked image examples for negative pseudo relevance feedback (NPRF) because of their high reliability. In [10], they propose to decompose the bottom negative samples into several partitions and combine all the positive examples in query with each partition as training data to build several classifiers, and finally use a logistic regression to combine the outputs of all the classifiers. Meanwhile, the research proposed in [11] is a more general one and gives an in-depth study to the relationship between retrieval score and their performance criterion. Second, the denotation of bag-of-visual-words (BOW) makes the representation of images similar with that of documents thus some PRF techniques in text retrieval can be reused. J. H. Hsiao et al [6] introduce BOW representation into image retrieval task in which the KL-divergence language modeling-based retrieval with dirichlet smoothing become feasible as it is used in [4]. The PRF in [6] is an unsupervised learning process using a linear interpolation between query and feedback model. The probabilistic model of feedback is a summation over language models of the top-ranked retrieval images.

Both in text retrieval and image retrieval the techniques of PRF have achieved successful performance. However, these feedback methods have hardly been used in the field of HLF. To our best knowledge, this is the first attempt trying to introduce PRF into HLF. The reason why we conduct this introducing mainly focuses on the basic design criterion of HLF task. Firstly, it is widely accepted that relevance feedback has been recognized as an effective and necessary strategy in CBIR due to the fact that the retrieved images can only meet users' needs partly. Nevertheless relevance feedback badly violates the guideline of HLF as completely no manual intervention should be allowed in testing dataset. Thus the alternative of PRF for HLF makes the process of feedback reasonable. Second, it is common known that HLF is a binary classification problem. The difference between retrieval and classification is substantial since a retrieval algorithm might only obtain a small amount of training "data" from query and there is no negative training data at all [10]. Previous work in [10, 11] try to transfer the process of PRF into classification problems by sampling the lowest ranked samples as negative ones. Therefore, the technique of PRF seems to be more appropriate to be applied in HLF task, because abundant training data including both positive and negative samples which describe the detecting "concept" can provide more reliable and truthful relevance information to guide the process of PRF directly. As a consequence, we have proposed three kinds of strategies to validate this notion, i.e., Nearest-Neighbor based PRF, Score-Evaluation based PRF, and Multi-Classifier Decision based PRF. All the strategies are combined with incremental learning [3] to reduce the re-training time. These three strategies are evaluated on the TRECVID2008 dataset, and MCD based strategy outperforms the other two and achieves a significant improvement in average precision compared with the baseline results without PRB.

The remainder of this paper is organized as follows. In section 2, a brief introduction about the overall framework in HLF with PRF is presented. In section 3, our proposed PRF-mechanisms in HLF are described in detail. Section 4 presents the experimental results of HLF average precision by using our PRF approaches.

2. BASELINE FRAMEWORK OF HLF WITH PRF

We use our baseline framework of HLF participating in TRECVID2008 as the platform for exploring PRF approaches proposed in this paper. The detection unit is based on shot-level. Each shot is represented by one or two key-frames. We extract five categories of image features as representation of key-frames, including 166-d ColorCorrelogram (CC), 166-d ColorHistogram (CH), 225-d ColorMoments (CM), 320-d EdgeHistogram (EH) and 96-d TextureCooccurence (TC). Each kind of image feature is corresponding to a margin-based SVM classifier and the procedure of each classification is independent. The overall structure of our detection system with PRF is depicted in Figure 1. The process of PRF is based on the initial output of these five classifiers. Because of plentiful training samples, the re-training process after PRF is time consuming if we add all training data with pseudo feedback samples. Therefore, technique of incremental learning in [3] is adopted to reduce the re-training time. That is only support vectors (SV), samples misclassified by classifiers, are combined with pseudo feedback samples for the next training iteration. From previous experience, the amount of support vectors is much less than the total number of training data and most of support vectors surround the two maximal separating hyperplanes. Due to computational issues, the feedback process iterates only once.

3. THE PROPOSED PRF ALGORITHMS

In this section, we detail the proposed PRF algorithms in our framework. Since the performance of HLF detection algorithms are not as ideal as people expected, it is also not reasonable to assume the top-ranked shots as pseudo positive samples. Consequently, the strategy of how to utilize positive training samples to select more pseudo positive samples as much as correctly and filter out samples in top-ranked list which may lead to destroy the concept's separability become an urgent task in PRF of high level feature detection.

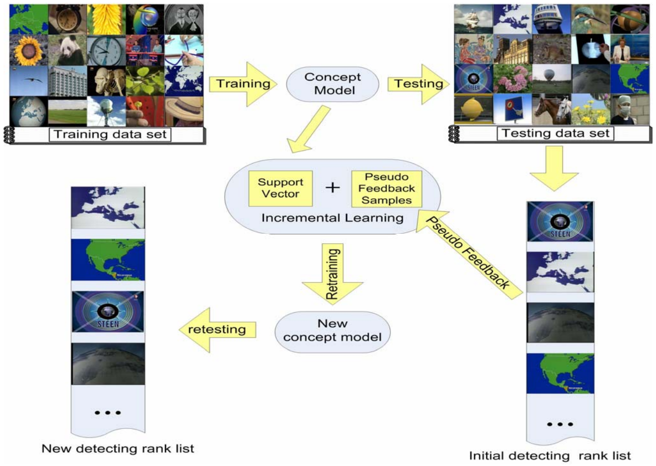

Figure 1: Overall structure of HLF detection system with PRF

3.1 Nearest-Neighbor based approach

The most straightforward method is to use Nearest-Neighbor strategy to solve this problem, because in our intuitive notion that image representation of positive sample in testing dataset should be similar to that in training dataset. Thus for a number of top-ranked shots called pre-processing sample set (in our experiments, the number of this set is set to be 1000), we calculate the distance between each sample in this set and each positive training sample. For each pre-processing sample, we examine its nearest distance with positive training sample. If the distance value is below a pre-defined threshold (in our experiment, five features distance possess different thresholds as the values of features distance fall in different intervals), this pre-processing sample will be added into pseudo positive sample set which is combined with SV set for the next detecting iteration. The distance metric used in this paradigm is Euclidean distance. The formulation can be expressed as:

$$\text{for each } u \in PT \quad if \quad \min\{\, dis(u,v) \quad v \in T\,\} \begin{cases} \leq \quad \theta \quad u \in PP \\ > \quad \theta \quad u \notin PP \end{cases}$$

PT : the pre-processing sample set

T : the set of positive samples in training dataset

PP : the set of pseudo positive samples for feedback

After getting the pseudo positive sample sets, five classifiers conduct their process of incremental learning independently as demonstrated in Figure 2(a).The *PP* sets corresponding to different feature representations will be different and the SV sets are also quite different from classifier to classifier.

3.2 Score-Evaluation based approach

The work of utilizing Score-Evaluation to do PRF is enlightened by the methods of score distribution evaluation and feature score aggregation proposed in [8]. They aggregate the outputs of multiple classifiers through Baye's rule to assume "pseudo" labels. As we know that each testing sample is predicted with a score indicating how likely it contains the detecting concept. Given the scores on instance $x_i : z_i^1, ..., z_i^M$ produced by a set of M independent classifiers(in our algorithm they are corresponding to CC, CH, CM, EH and TC), the way to aggregate scores to predict "pseudo" label is to compute the posterior distribution $P(y_i = 1 | z_i^1, ..., z_i^M)$ as follow:

$$P(y_i = 1 | z_i^1, ..., z_i^M) = \frac{P(y_i = 1) \prod_{k=1}^{M} p(z_i^k | y_i = 1)}{\sum_{y_i = -1,1} P(y_i) \prod_{k=1}^{M} p(z_i^k | y_i)}$$

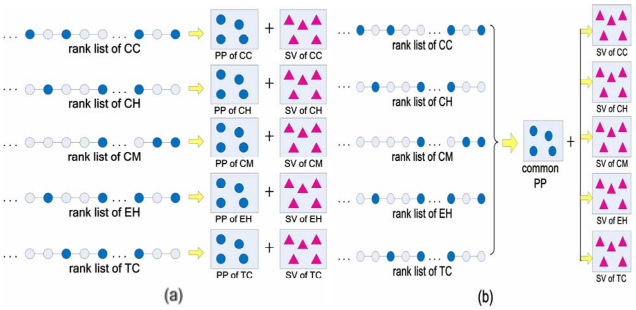

Figure 2: the process of incremental learning

The prior $P(y_i=1)$ is set to the ratio of positive samples in training dataset while $P(y_i=-1)$ is the ratio of negatives. Two Gaussian distributions are used to fit the score distribution of positive and negative samples respectively, i.e., the distribution of $p(z_i^k \mid y_i=1)=N(u_p,\sigma_p)$ and $p(z_i^k \mid y_i=-1)=N(u_n,\sigma_n)$. The overall score distribution of testing dataset predicted by classifier k is modeled through a Gaussian mixture model with two components:

$$p(z^k) = \pi N(u_p,\sigma_p) + (1-\pi) N(u_n,\sigma_n)$$

The parameters $(\pi, u_p, \sigma_p, u_n, \sigma_n)$ are estimated through EM algorithm, which only depends on the scores $\{z_i^k\}$.

After aggregation, the way to get the "pseudo label" based on the posterior distribution is that $y_i = \text{sgn}(P(y_i=1 \mid z_i^1,...,z_i^M)-0.5)$. Samples with positive pseudo labels are added into the pseudo positive sample set. The set for five classifiers is the same and is combined with different support vector sets produced by different models as depicted in Figure 2 (b).

3.3 Multi-Classifier Decision based approach

The score prediction by individual classifier is not accurate and might probably cause some "false positive" samples embedded in the top-ranked shots. Therefore, it is unreliable to determine the pseudo positive samples for feedback by individual classifier's decision. We propose this approach to identify the "true" pseudo positive samples through the joint decision of multiple classifiers. That is for a sample in testing dataset, if the five classifiers we have mentioned are all considered it as a "positive" one, then it will be assumed as pseudo positive sample. However, each classifier predicts an instance with a score and doesn't assume it positive or not. The way $y_i = \text{sgn}(P(y_i=1 \mid z_i^k)-0.5)$ to convert score into label can also be used here, but the probability scores predicted by different classifiers might converge in different intervals in [0, 1] which results in unequal processing for the five classifiers. The strategy we adopted in our algorithm is to examine the rank position of testing samples in the rank list. As we know that a low rank position stands for a high probability of being positive. If a testing sample possesses a comparatively low rank position through all rank lists produced by the five classifiers, then it will be considered as pseudo positive sample. For conventional expression, we transfer the position to values of probability between [0, 1] by:

$$p(x) = \frac{TN - Po(x)}{TN}$$

TN : the number of samples in testing dataset

$Po(x)$: the rank position of sample x in the rank list

$P(x) = \sum_{i \in M} p_i(x)$ M = {ranklists of CC, CH, CM, EH, and TC}

The samples in the testing dataset are re-ranked by the value of *P(x)* in a descending order. The top-ranked samples in the re-rank list are assumed as pseudo positive samples for the next training round. The quantity of pseudo positive samples is proportional to the ratio of positive samples in training dataset. The constraint condition of assuming a sample "positive" which is jointly identified by five classifiers is a little strong since only few testing samples are actually top-ranked at the lists of all five classifiers simultaneously in our experimental results. We loosen this condition by requiring only three of five classifiers common considering one sample as positive is qualified. The formulation can be expressed as:

$$P(x) = \max(\sum_{i=1}^{3} p_i(x))$$

$(p_i(x)$: probability for instance x in ranklist CC,CH,CM,EH, and TC)

The incremental process of Multi-Classifier Decision based approach is depicted in Figure 2(b), the same as in section 3.2.

4. EXPERIMENTS AND OBSERVATIONS

In this section we start to describe the experimental setup and implementation process in details. Several experiments are conducted to compare the effectiveness of each approach.

4.1 Experimental Setup

The video data come from the video collection provided by TREC-VID 2008 and 12 concepts are detected to validate our assumption. We sample training data in TRECVID2008 workshop as our base dataset, because their shot labels are available for the public. We extract 39674 keyframes from the shots of 220 training videos which are decomposed according to TRECVID benchmark of shot boundary detection. Then we sample 18767 keyframes from the overall training set as our experimental training set and 8093 keyframes as the testing set. The detecting performance is evaluated with the average precision (AP) [12]. The Mean Average Precision (MAP) is also calculated. We setup six experimental settings to compare the effectiveness of our approaches. They are SVM Baseline without PRF, NN based PRF with negative samples (NN<with neg>**)**, NN based PRF without negative samples (NN<no neg>**)**, SE based PRF without negative samples (SE<no neg>), MCD based PRF with negative samples (MCD <with neg>) and MCD based PRF without negative samples (MCD<no neg>). Similar to [10, 11], we conduct experiments with negative samples sampled from the bottom of initial rank list. However, the results of these negative settings are not as good as those without negative as shown in Table 1. The reason may be that the negative samples make the two maximum-margin hyperplanes drifting away from their original position for initial training dataset because the bottom sampled negative samples are far from the hyperplanes while most support vectors are around the planes.

4.2 Experimental Results and Conclusions

The results of six experimental settings are presented in Table 1. The percentages in parentheses denote the percent by which these settings outperform the SVM baseline without PRF. We can see from Table 1 that the setting MCD<with neg> and MCD<no neg> have achieved good performance. Figure 3 presents average precision of each concept detector for MCD <with neg> and <no neg> compared with the baseline paradigm without PRF. However, the setting NN-based PRF and SE-based PRF are not as good as we expected. The NN-based approach may be due to the semantic

	SVM Baseline	NN <with neg>	NN <no neg>	SE <no neg>	MCD <with neg>	MCD <no neg>
CC	0.072854	0.047627	0.051056	0.012234	0.077683(6.6%)	0.093577(28.4%)
CH	0.062159	0.053041	0.053896	0.012635	0.070431(13.3%)	0.079865(28.5%)
CM	0.092330	0.072204	0.072844	0.029651	0.096858(4.9%)	0.106128(14.9%)
EH	0.094482	0.069499	0.059431	0.013882	0.093452(-1.1%)	0.101762(7.7%)
TC	0.044424	0.034204	0.030113	0.023349	0.057972(30.5%)	0.070244(58.1%)

Table 1: Mean Average Precision of 12 concept detectors for five experimental settings.

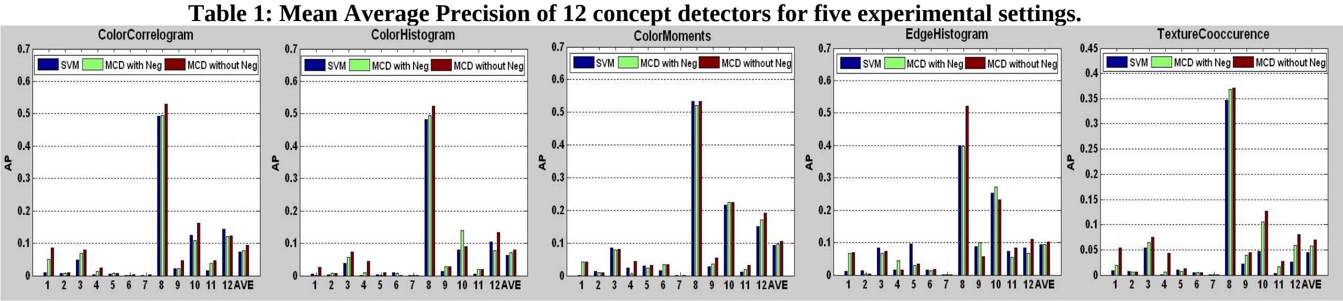

Figure 3: Average Precision of 12 concept detectors for SVM Baseline, MCD <with neg> and MCD <no neg> settings.

gap between the similarity computed in feature space for image representation and the similarity of user's perception [5]. The process of selecting pseudo positive samples in this method only utilizes the low level features for the distance comparison while ignores those semantic information embedding in positive training samples. Digging out semantic information from positive training samples as much as possible is a critical issue for PRF in HLF. The SE-based PRF reduces the HLF performance dramatically. Two possible reasons can be concluded for this problem. First, the process of score evaluation and aggregation perhaps destroys semantic information judged by each classifier and in the whole selecting process there are no explicit and implicit clues drawn from positive samples in training dataset at all. Second, there perhaps exists mismatching between the actual score distribution and the modeling distribution for scores. Some more suitable probabilistic models should be adopted for score distribution evaluation. The success of MCD-based PRF validates the effectiveness of applying PRF to HLF by utilizing semantic information in training dataset, because the initial detection by each classifier is also a process to map these low level features to the semantic level and then the joint identifying by several classifiers for selection is a simple strategy to utilizing these semantic information. Thus the MCD-based approach achieves excellent performance in PRF for HLF. Future work should focus on exploring some more sophisticated strategies to obtain useful and truthful semantic information to guide the process of PRF.

5. ACKNOWLEDGMENTS

This work was supported by National Basic Research Program of China (2007CB311100), National High Technology and Research Development Program of China (2007AA01Z416), National Nature Science Foundation of China (60873165) and Co-building Program of Beijing Municipal Education Commission.

6. REFERENCES

[1] C. X. Zhai, and J. Lafferty, "Model-based feedback in the KL-divergence retrieval model", *In Proc. of the 10th ACM Intl. Conf. on Information and Knowledge Management*, pp. 403-410, 2001.

[2] C. Zhai, and J. Lafferty, "A study of smoothing methods for language models applied to information retrieval", *ACM Transaction on Information System*, Vol. 22, pp. 179-214, June 2004.

[3] F. Wu, Y. Zhuang, and Y. Pan "Audio Clip Recognition and Retrieval Based on Incremental Learning with Support Vector Machine", *Journal of Computer Research and Development,* Vol. 40, pp. July 2003.

[4] G. Cao, J. Y. Nie, J. Gao, and S. Robertson, "Selecting Good Expansion Terms for Pseudo-Relevance Feedback", *in Proc. of the 31st ACM Intl. Conf. SIGIR* , pp. 243-250, 2008.

[5] G. Giacinto, "A Nearest-Neighbor Approach to Relevance Feedback in Content Based Image Retrieval", *in Proc. of the 6th ACM Intl. Conf. on Image and Video Retrieval*, pp. 456-463, 2007.

[6] J. H. Hsiao, C. S. Chen, and M. S. Chen, "Visual-Word-Based Duplicate Image Search with Pseudo-Relevance Feedback", *in IEEE Intl. Conf. on Multimedia & Expo,* pp. 699-672, 2008.

[7] J. M. Ponte, and W. B. Croft, "A Language Modeling Approach to Information Retrieval", *in Proc. of the 21st Annual ACM Intl. Conf. SIGIR* , pp. 275-281, 1998.

[8] J. Yang, R. Yang, and A. G. Hauptmann, "Cross-Domain Video Concept Detection Using Adaptive SVMs", *in Proc. of ACM intl. conf. on Multimedia*, pp. 188-197, 2007.

[9] K. S. Lee, W. B. Croft, and J. Allan, "A Cluster-Based Resampling Method for Pseudo-Relevance Feedback", *in Proc. of the 31st ACM Int.l Conf. SIGIR*, pp. 235-242, 2008.

[10] R. Yan, A. G.. Hauptmann, and R. Jin, "Multimedia Search with Pseudo-relevance Feedback", *in Proc. of Intl. Conf. on Image and Video Retrieval*, pp. 238-247, 2003.

[11] R. Yan, A. G. Hauptamn, and R. Jin, "Negative Pseudo-Relevance Feedback in Content-based Video Retrieval", *in Proc. of ACM intl. conf. on Multimedia*, pp. 343-346, 2003.

[12] http://www-nlpir.nist.gov/projects/tv2008/tv2008.html

[13] T. Sakai, T. Manabe, and M. Koyama, "Flexible Pseudo-Relevance Feedback via Selective Sampling", *ACM Transaction on Asian Language Information Procession,* Vol. 4, pp. 111-135, June 2005.

[14] T. Tao, and C. X. Zhai, "Regularized Estimation of Mixture Models for Robust Pseudo-Relevance Feedback", *in Proc. of the 29st ACM Intl. Conf. SIGIR*, pp. 162-169, 2006.

[15] V. Lavenko, and W. B. Croft, "Relevance-Based Language Models", *in the 24th ACM Intl. Conf. SIGIR*, pp. 120-127, 2001.

[16] X. Huang, Y. R. Huang, M. Wen, A. An, Y. Liu, and J. Poon, "Applying Data Mining to Pseudo-Relevance Feedback for High Performance Text Retrieval", in Proc. of the Sixth Intl. Conf. on Data Mining, pp. 295-306, 2006.