

Negative Pseudo-Relevance Feedback in Content-based Video Retrieval

Rong Yan
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, 15213
yanrong@cs.cmu.edu

Alexander G. Hauptmann
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, 15213
alex+@cs.cmu.edu

Rong Jin
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, 15213
rong@cs.cmu.edu

ABSTRACT

Video information retrieval requires a system to find information relevant to a query which may be represented simultaneously in different ways through a text description, audio, still images and/or video sequences. We present a novel approach that uses pseudo-relevance feedback from retrieved items that are NOT similar to the query items without further inquiring user feedback. We provide insight into this approach using a statistical model and suggest a score combination scheme via posterior probability estimation. An evaluation on the 2002 TREC Video Track queries shows that this technique can improve video retrieval performance on a real collection. We believe that negative pseudo-relevance feedback shows great promise for very difficult multimedia retrieval tasks, especially when combined with other different retrieval algorithms.

1. INTRODUCTION

Pattern recognition techniques have been widely applied in video information retrieval systems. In these systems, there are two fundamental problems to be addressed. One is the representation of multi-modal features and the other the design of a similarity metric which determines the “distance” between two examples. However, CBVR systems that simply rely on a pre-defined generic similarity metric cannot achieve good performance. Therefore, we would like to make the similarity metric adaptive with respect to different queries. This requires approaches that are able to automatically discover the discriminating feature subspace once the queries are provided. A possible solution is to cast this formulation of retrieval as a classification problem, where relevant examples are the positive instances and non-relevant examples are the negative instances of a class. Recent work [9, 8] has suggested that margin-based classifiers such as support vector machines (SVMs) and Adaboosting can yield high generalization performance and automatically emphasize the useful features by learning the maximal mar-

gin hyperplane in the embedding space.

However, it is generally a characteristic of information retrieval that the user’s query only provides a small amount of positive data and no explicit negative training data at all. Thus, if we want to make use of margin-based learning algorithms for multimedia information retrieval, methods have to be devised which can provide more training data, especially some negative examples. Standard relevance feedback addresses this issue in an interactive fashion, in which the system iteratively asks users to label more training examples as relevant/non-relevant for the learning algorithms [4, 9]. However, it is tedious to hand pick negative examples and subjectively quite difficult to provide a good negative sample [14] that clearly shows the distinction to the positive examples, since negative instances are less well-defined as a coherent subset.

Instead of relying on the relevance feedback judgments of real users, it is worthwhile to consider the idea of obtaining additional relevant/non-relevant training examples via automatic relevance feedback based on a generic similarity metric, which is not tailored to the specific queries. However, it quickly becomes apparent that it is inappropriate to consider the top-ranked examples from the generic similarity metric for positive feedback due to the poor performance of current video retrieval algorithms in general applications. We found that it is more reasonable to sample the bottom-ranked examples for negative feedback. From the viewpoint of machine learning, our approach is closely related to a learning framework called positive example based learning or partially supervised learning [14].

In this paper, we present a novel automatic retrieval technique for multimedia data called negative pseudo-relevance feedback (NPRF). It attempts to learn an adaptive similarity space by automatically feeding back the training data which are identified based on a generic similarity metric. An early version of this technique is given in [12]. Based on this work, we provide an in-depth discussion based on a statistical model of average precision. We also discuss the combination strategy for different retrieval algorithms, which is essential in the face of unreliability in the NPRF approach. The experimental results discussed later in the chapter confirm the effectiveness of the proposed approach.

2. NEGATIVE PSEUDO-RELEVANCE FEEDBACK

In the task of content-based video retrieval, a query typ-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MM’03, November 2–8, 2003, Berkeley, California, USA.
Copyright 2003 ACM 1-58113-722-2/03/0011 ...\$5.00.

Input:

- Query examples q_1, \dots, q_n
- Video shots in the video collection v_1, \dots, v_m

Output:

- Final retrieval score s_i^f for each shot v_i

Algorithm:

1. For every i , $1 \leq i \leq m$, compute base similarity score $s_i^0 = f_b(v_i, q_1, \dots, q_n)$
2. Iteratively, k from 0 to max
 - (a) Given the retrieval scores s_i^k , sample the positive examples pos_i^k and negative examples neg_i^k using some sampling strategy p
 - (b) Compute the NPRF score $s_i^{k+1} = f_l(v_i)$ where the learning algorithm f_l is trained by pos_i^k and neg_i^k .
3. Combine all the retrieval scores into a combination score $s_i^f = g(s_i^0, \dots, s_i^{max+1})$

Figure 1: The algorithm for negative pseudo-relevance feedback

ically consists of a text description plus audio, images or video. This query is posed against a video collection. The job of the video retrieval algorithm is to retrieve a set of relevant video shots from a given data collection. The retrieval algorithm should provide a permutation of the video shots v_i in target video collection V , which is sorted by their similarity to the user queries q_i in queries Q . When retrieval is thought of as a classification problem, video data collection can be separated into two parts for each query, where positive data V^+ are the relevant shots and negative data V^- are non-relevant ones.

Figure 1 summarizes the algorithm, which is similar to a relevance feedback process except that users' judgement is replaced by the output of a generic similarity metric. This algorithm consists of four major components, that is, a generic similarity metric f_b , a sampling strategy p , a learning algorithm f_l and a combination strategy g . It starts by computing the retrieval scores using f_b for every video shot v_i . Next, it iteratively identifies new training examples and computes an updated retrieval score. For each run k , the sampling strategy p is used to extract positive and negative examples from the video collection. These positive and negative data are combined to train a learning algorithm f_l . The output s_i^{k+1} of f_l can be interpreted as an updated retrieval score for each shot v_i . Finally, the retrieval scores are fused into a final result via some combination strategy g . In our implementation, the positive examples are the query examples and the negative examples are sampled from the strongest negative examples. Due to the computational issues, the feedback process repeat for only one iteration. For the sake of simplicity, we call s_i^0 the base similarity score, s_i^1 the NPRF score and s_i^f the combination score.

Although the NPRF approach can be applied to various retrieval tasks such as text retrieval and audio retrieval, in this initial work we have mainly applied it in image retrieval.

The base similarity score is chosen as the aggregate dissimilarity model proposed by Wu et al[11], which is able to learn disjunctive models within any metric space. In their model, the aggregate dissimilarity for example x to the query q_1, \dots, q_n is expressed by

$$d(x)^\alpha = \begin{cases} 0 & \text{if } (\alpha < 0) \wedge d(x, q_i) = 0 \\ \frac{1}{n} \sum_{i=1}^n d(x, q_i)^\alpha & \text{otherwise} \end{cases} \quad (1)$$

where α is set to -1. The distance $d(x, q_i)$ is computed using the Euclidean distance function.

In the feedback step, the query images are considered the only positive examples and a subset of the images that are most dissimilar to the queries will be considered as the negative examples. As a basic setting, the number of negative examples is equivalent to the number of positive examples. The training examples provided by sampling strategies are fed back to train a margin-based classifier. In our experiments, support vector machines (SVMs)[1] were used since SVMs are known to yield good generalization performance especially in high dimensional data. Finally, after the retrieval scores are calibrated to posterior probabilities, they can be linearly combined into the final score $s^f = \sum_j \lambda_j s^j$

3. ANALYSIS

In this section, we would like to provide in-depth discussion on how the NPRF approach works when applied to an actual video collection. Our analysis is based on a statistical model of average precision. We also present several score fusion paradigms by transforming different types of similarity metrics into probabilistic outputs. Note that we adopted average precision(AP) [10] as the performance measure, which corresponds to the area under an ideal recall / precision curve. Mean average precision(MAP) is the average of these average precision over all topics.

Without loss of generality, we can define the positive distance d^+ as the distance between a relevant shot $v \in V^+$ and the queries. The negative distance d^- is similarly defined. As suggested in previous studies[7], it is reasonable to assume that d^+ and d^- are both approximately Gaussian distributed, although gaussian distribution might not be the best model. Tarel et al [7] show that if a similarity metric can be represented by the sum of the similarity metrics of its components, the positive distance d^+ and negative distance d^- will converge towards a Gaussian distribution when the number of features goes to infinity and each dimension is independent identical distributed(i.i.d).

3.1 A statistical model for average precision

To provide more insights for NPRF approach, it will be useful to study the relationship between retrieval score distribution and our performance criterion, i.e. average precision. In the following discussion, let the mean and variance for the d^+ be μ^+, σ^+ , while the mean and variance for d^- be μ^-, σ^- . The video collection has a total of N^+ relevant shots and N^- irrelevant shots. Under the assumption of Gaussian distribution, the MAP can be simplified to

$$AP = \frac{N^+}{N^-} \left[\frac{1}{2} + e^{-C \frac{\mu^- - \mu^+}{\sigma^-}} \cdot \text{Beta}(2 - \frac{\sigma^+}{\sigma^-}, 1 + \frac{\sigma^+}{\sigma^-}) \right] \quad (2)$$

where more details can be seen in [13].

Figure 2 plots the distributions of base similarity score and NPRF score for two actual queries in TREC02 search

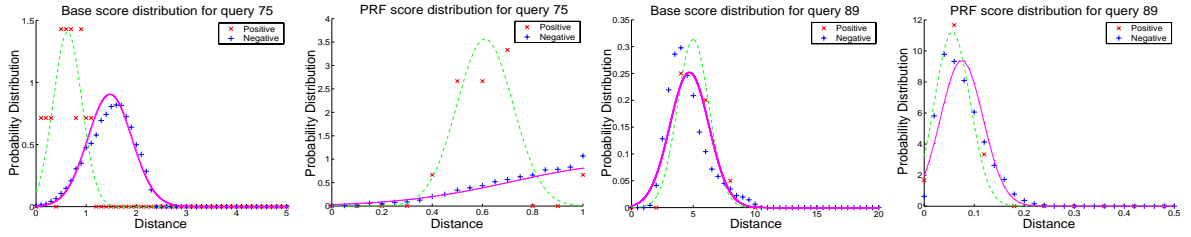


Figure 2: Probabilistic distribution of positive distance and negative distance for Query 75 and Query 89 in TREC02 data. Each red plus sign are the normalized number of positive examples in each bin in a histogram, while blue cross sign are that of negative distance. The green dash line and magenta solid line are the estimated Gaussian distribution for positive distance and negative distance respectively.

tasks. As expected, the negative distance can be perfectly fit to the Gaussian model, however, the positive distance are only approximately Gaussian distributed due to the small sampling size. To take a closer look, we found that the major performance improvement of the NPRF approach is due to the increase of the normalized mean distance $\frac{\mu^- - \mu^+}{\sigma^-}$. In query 89, base similarity score produces a poor performance since its positive mean μ^+ is even larger than negative mean μ^- . In this case, NPRF can greatly reduce the normalized positive mean distance μ^+/σ^- , since it has the ability to adapt similarity scores across queries. However, as a trade-off, it will over score the "false positives" which are far away from the negative training examples. In query 75, base similarity score does already achieve a high average precision and leaves no room for NPRF to improve further. Therefore, the normalized positive mean distance μ^+/σ^- does not change too much itself. Unfortunately, figure 2 shows a considerable reduction of μ^+/σ^- , which indicates that more false positives has been assigned a lower retrieval score than the positive examples. This "false positive" problem greatly degrade the performance of NPRF approach.

3.2 Probabilistic Output and Combination

Fusion of different retrieval algorithms is an effective way to address the "false positive" problem in NPRF. As mentioned before, combining base score and NPRF score might offer a reasonable trade-off. More interestingly, it has been found that combination of NPRF score and retrieval scores from different modalities can recover most of the performance hurt since most false positives can be filtered out by additional information. Also, these combinations can reduce the prediction variance and offer more stable results. In this section, we study how to combine different retrieval algorithms into one and present our combination schemes via the estimation of posterior probabilities.

Platt et al [2] suggest using a parametric sigmoid model to fit the posterior directly,

$$p(+|t) = \frac{1}{1 + \exp(At + B)} \quad (3)$$

However, we prefer an approximation which leads to reasonable prediction effectiveness with less computational effort. One solution is to set the parameters manually based on empirical testing. Especially when the output of the retrieval algorithm is bounded by some interval $[\min, \max]$, one can always set the parameters to make $p(+|\min)$ close to 0 and $p(+|\max)$ close to 1. Experimental results show that this ad-hoc parameter setting can lead to reasonable performance.

Another form of approximation can be derived from the rank distribution,

$$p(+|t = t_0) = 1 - \frac{\text{Rank}(e)}{N} \quad (4)$$

where N is the number of all the examples in the collection. This approximation allows a simpler form of probability estimation which is also called "weighted borda voting" in the literature.

4. EXPERIMENTAL RESULTS

The video data came from the video collection provided by the TREC Video Retrieval Track. The definitive information about this collection can be found at the NIST TREC Video Track web site [10]. On the image processing side, two types of low-level image features including color features and texture features were used in our system. The color feature is the cumulative color histogram for the HSV (Hue-Saturation-Value) color space [6] using 16 bins each channel. The texture features are obtained from the convolution of the image pixels with various Gabor wavelet filters [5], which is quantized into 16 bins. Their central and second-order moments are generated as the texture feature.

Apart from the image retrieval, we also extract information from other modalities, especially the text information like speech and video OCR transcripts, movie titles and external video summaries. The first type of textual information is from speech and Video OCR transcripts. The retrieval of these transcripts is done using the OKAPI BM-25 formula[3]. Externally provided video summaries are another source of textual information. For each query, the posterior probability of a video shot is set to 1 if any keyword of the query can be found in the video summaries for the corresponding movie, otherwise the posterior probability is set to 0.

We used the *SVM^{Light}* as the implementation of the SVM algorithm. The setting for the feedback learning algorithm is RBF kernel SVMs with parameter 0.05. For the probabilistic output, the parameters (A, B) is manually set to be $(-10, -2)$. For the combination of retrieval scores, the weight for speech transcript λ_{sp} are set to 1 and the weight for video summary information λ_v was set to 0.2. Specifically, the combination rule becomes $s_i^f = (1 - \lambda_b)s_i^{NPRF} + \lambda_b s_i^{Base} + \lambda_{sp} s_i^{sp} + \lambda_v s_i^v$

Next we presented the performance of the NPRF approach. Table 1 lists the comparison between NPRF and the base similarity score in terms of precision, recall and mean average precision. As can be seen from the figure, NPRF

	Precision	Recall	MAP
Base	0.1108	0.3000	0.1415
NPRF	0.1320	0.3318	0.1522

Table 1: Comparison between base similarity score and NPRF score in terms of precision, recall and mean average precision(MAP)

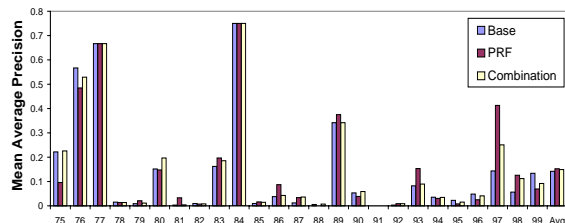


Figure 3: Comparison between base metric, NPRF score and their combination for individual queries. The mean average precision is shown versus the query IDs

achieves a performance improvement over the base retrieval algorithm in all three performance measures. The NPRF approach can achieve reasonable precision improvement, recall improvement and MAP improvement beyond the base similarity score.

Our next experiment was designed to examine the effect of the NPRF approach for individual queries. Figure 3 compares the mean average precision per query of the base similarity score, the NPRF approach and their combination when $\lambda_b = 0.5$. Compared to the base similarity score, the NPRF score results in a large increase for query 23, and query 19 but mostly loses in queries 1, and query 25. The combination of both achieves a fairly good trade-off between them. As expected, only 11 of the 25 queries can achieve a higher MAP with the NPRF approach over the base retrieval algorithm and 10 has lower MAP, but their combination seems to benefit most of the queries, which has 14 queries higher and only 3 lower than the base retrieval algorithm. This again indicates the importance of the combination strategies.

5. CONCLUSION

This paper presented a novel technique, negative pseudo-relevance feedback(NPRF), to improve the performance of content-based video retrieval. Different from the canonical relevance feedback technique, our approach does not require users to provide judgment within a retrieval process. In this work, the task of video retrieval is framed as a concept classification problem. Specifically, the positive examples are provided by users' query and negative examples can be obtained from the worst matching examples identified based on a generic similarity metric. A margin-based learning algorithm, support vector machine (SVM), is used to learn the updated similarity scores. Theoretical analysis shows that the benefit of the NPRF approach derives from the ability to adapt similarity score across queries and thus separate the means of the negative/positive score distributions. Since some extreme outliers might be misclassified as false

positives, we suggest that smoothing with either the initial similarity score or the score from different modalities can safeguard against these egregious errors. Experiments on the data from the 2002 TREC Video track evaluations confirmed the effectiveness of the NPRF approach on a collection of over 14000 shots in 40 hours of video.

Acknowledgments

This research is partially supported by the advanced Research and Development Activity (ARDA) under contract number MDA908-00-C-0037 and MDA904-02-C-0451. We also thank Jian Zhang and Weihao Lin for stimulating discussion.

6. REFERENCES

- [1] C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):955–974, 1998.
- [2] J. Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In B. S. A. Smola, P. Bartlett and D. Schuurmans, editors, *Advances in Large Margin Classifiers*. MIT Press, 1999.
- [3] S. E. Robertson, S. Walker, M. Hancock-Beaulieu, A. Gull, and M. Lau. Okapi at TREC4. In *Text REtrieval Conference*, pages 21–30, 1992.
- [4] Y. Rui, T. S. Huang, and S. Mehrotra. Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Trans. on Circuits and Systems for Video Technology*, 8:644–655, September 1998.
- [5] B. Schiele and J. L. Crowley. Object recognition using multidimensional receptive field histograms. In *ECCV*, 1996.
- [6] M. A. Stricker and M. Orengo. Similarity of color images. In *Storage and Retrieval for Image and Video Databases (SPIE)*, pages 381–392, 1995.
- [7] J. P. Tarel and S. Boughorbel. On the choice of similarity measures for image retrieval by example. In *ACM Intl. Conf. on Multimedia*, pages 107–118, 2002.
- [8] K. Tieu and P. Viola. Boosting image retrieval. In *Intl. Conf. on Computer Vision*, pages 228–235, 2001.
- [9] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *ACM Intl. Conf. on Multimedia*, pages 107–118, 2001.
- [10] TREC2002. TREC2002 video track, <http://www-nlpir.nist.gov/projects/t2002v/t2002v.html>.
- [11] L. Wu, C. Faloutsos, K. P. Sycara, and T. R. Payne. Multimedia queries by example and relevance feedback. *IEEE Data Engineering Bulletin*, 24(3), 2001.
- [12] R. Yan, A. Hauptmann, and R. Jin. Multimedia search with pseudo-relevance feedback. In *Intl Conf on Image and Video Retrieval*, pages 238–247, 2003.
- [13] R. Yan, A. Hauptmann, and R. Jin. Pseudo-relevance feedback for multimedia retrieval. In A. Rosenfeld, D. Doermann, and D. DeMenthon, editors, *Video mining*. Kluwer Academic Publishers, 2003.
- [14] H. Yu, J. Han, and K. C. Chang. PEBL: Positive example based learning for web page classification using SVM. In *Proceedings of the 2002 ACM SIGKDD Conference (KDD 2002)*, pages 239–248, 2002.