

Query Expansion using Probabilistic Local Feedback with Application to Multimedia Retrieval

Rong Yan
Intelligent Information Mgmt. Dept.
IBM TJ Watson Research Center
19 Skyline Dr., Hawthorne, NY
yanrong@us.ibm.com

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

ABSTRACT

As one of the most effective query expansion approaches, local feedback is able to automatically discover new query terms and improve retrieval accuracy for different retrieval models. However, the performance of local feedback is heavily dependent on the assumption that most top-ranked documents are relevant to the query topic. Although this assumption might be sensible for ad-hoc text retrieval, it is usually violated in many other retrieval tasks such as multimedia retrieval. In this paper, we develop a robust local analysis approach called probabilistic local feedback (PLF) based on a discriminative probabilistic retrieval framework. The proposed model is effective for improving retrieval accuracy without assuming the most top-ranked documents are relevant. It also provides a sound probabilistic interpretation and a convergence guarantee on the iterative result updating process. Although derived from variational techniques, this approach only involves an iterative process of simple operations on ranking features and thus can be computed efficiently in practice. Our multimedia retrieval experiments on TRECVID'03-'05 collections have demonstrated the advantage of the proposed PLF approaches which can achieve noticeable gains in terms of mean average precision over various baseline methods and PRF-augmented results.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Retrieval models

General Terms

Algorithms, Performance, Theory

Keywords

Multimedia Retrieval, Query Expansion, Probabilistic Local Feedback

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.

CIKM'07, November 6–8, 2007, Lisboa, Portugal.

Copyright 2007 ACM 978-1-59593-803-9/07/0011 ...\$5.00.

1. INTRODUCTION

Most users begin their retrieval process without knowing the detailed collection information and the retrieval environment. Therefore their query terms can be significantly different from the terms used to describe the same concepts in the documents [20]. As one of the most effective ways to mitigate this “word mismatch” issue and improve the retrieval accuracy, automatic query expansion helps users to formulate a better query by adding new terms/phrases based on initial retrieval results or entire collection information. Numerous query expansion techniques have been developed before and most of them fall into two categories, i.e., global analysis and local analysis. The basic idea of global analysis is to expand the query description using global collection statistics based on the concurrence analysis of the entire collection. For example, a representative global technique is to automatically create a domain-specific thesaurus based on global term-to-term similarities [13] and use it to expand additional query terms based on their similarities to the query keywords. Other well-known global techniques include Latent Semantic Indexing [3], PhraseFinder [7] and so forth. However, global analysis has to obtain statistics for each pair of terms in the entire collection, which is a computationally demanding task especially for large text collections.

As an alternative, we can resort to a local analysis strategy that explores the information from initially retrieved documents in real time in order to determine which terms to expand. The essence of local analysis, also known as local feedback or pseudo-relevance feedback (PRF), is to utilize the top retrieved documents as positive examples to select discriminative query terms to improve the retrieval performance. In these approaches, a small number of top-ranked documents are assumed to be relevant, from which the expansion terms are discovered and used to formulate a second-cycle query. In practice, the idea of local analysis has been implemented in various forms for different retrieval models. In the vector space model, new query vectors are constructed by moving the initial query vectors towards the center of the feedback documents [15]. The classical probabilistic model takes feedback documents as positive examples and estimates the model parameters using Bayesian rules [14]. Language modeling approaches explore the local analysis idea by estimating additional query language models [9, 17] or relevance models [10] from a set of feedback documents. By combining global analysis and local analysis, Xu et al. [20] proposed an effective local analysis algorithm called local context analysis (LCA). In this work, several noun groups are selected from the top ranked documents based on the passage co-

occurrence of query terms and introduced into the original query. Recent TREC evaluation results have demonstrated that local analysis approaches are usually effective in improving retrieval accuracy and they outperform the global analysis approaches on average. However, these approaches suffer from a drawback that their performance is heavily dependent on the quality of the initial retrieval outputs. If the retrieval quality of the first set of results is poor, local analysis tends to fail with too many noisy terms introduced in the second round. Another concern is that although it is intuitive to execute multiple local analysis runs to obtain better results, these methods are not always guaranteed to converge to a fixed set of retrieval results.

From another perspective, while query expansion enjoyed great success over the last decade, it is usually discussed within the domain of ad-hoc text retrieval. But the idea of automatically enriching query semantics is largely applicable to many other retrieval scenarios. In this work, we investigate the query expansion methods in a broader context, which aims to augment retrieval results by expanding various types of ranking features beyond word features. For instance, queries in multimedia retrieval can be reformulated to incorporate additional “visual term” features from high-level semantic concepts, which are learned from the visual modality using manually annotated development data. However, broader applications can bring additional challenges to the local analysis approaches, such as more heterogenous document features and poorer initial performance. For example, the state-of-the-art automatic video retrieval systems can only achieve a mean average precision of 12% in the TRECVID’05 evaluation [16], which is obviously insufficient to support the PRF approaches.

In this paper, we develop a robust local analysis approach called probabilistic local feedback (PLF) based on a discriminative probabilistic retrieval framework. Because PLF only requires the top-ranked documents contain *more* relevant documents than the bottom-ranked documents, it can automatically expand ranking features to improve the initial retrieval accuracy even if most top-ranked documents are irrelevant. Formally, this model can be described as an undirected graphical model that treats document relevance and weights of ranking features as a set of latent variables. Therefore, it provides a sound probabilistic interpretation and a convergence guarantee on the iterative result updating process. Although derived from variational techniques, the final form of this approach only involves an iterative process of simple operations on ranking features, and thus it can be computed *efficiently* in practice¹. To investigate whether the proposed approach can perform robustly on poor initial retrieval results, we evaluate it on several multimedia retrieval collections, i.e., the TREC video track (TRECVID) ’03-’05 collections. Our retrieval experiments have demonstrated the effectiveness of the proposed approaches, which can achieve noticeable performance gains over various baseline methods and their PRF-augmented results.

2. DISCRIMINATIVE MODELS FOR INFORMATION RETRIEVAL

In the following discussions, we present a discriminative retrieval framework as a basic platform for the proposed

¹Please find more discussions on the issue of efficiency in Section 3.2

approach. Let us begin by introducing the basic notations and terminologies used in this work. The term *document* is used to refer to the basic unit of retrieval throughout this paper. A search collection \mathcal{D} contains a set of documents $\{d_1, \dots, d_j, \dots, d_{M_D}\}$. Given a query Q provided by users, let $y \in \{1, -1\}$ indicate if the document D is relevant or irrelevant to Q . For the query Q and each document D , we can generate a bag of ranking features from N text keywords or information sources, denoted as $f_i(D)$. For instance, in ad-hoc text retrieval, $f_i(D)$ can be defined as the tf.idf weight of the i^{th} term. In multimedia retrieval, $f_i(D)$ can be the outputs of uni-modal retrieval experts on text/visual modalities or the detection results of predefined semantic concepts. Generally speaking, the goal of information retrieval is to combine these ranking features $f_i(D)$ and generate a ranked list of documents that satisfies users’ information need.

Conventional relevance-based probabilistic models [5] rank documents by sorting the conditional probability that each document would be judged relevant to the given query, i.e., $P(y = 1|D, Q)$. Most well-known text retrieval models, such as the BIR model [14], proceed by inverting the position of y and D based on the Bayes rule and estimating the generative probabilities of document D in the relevant and irrelevant documents. However, the underlying model assumptions of these approaches such as term independency, could be invalid in practice. In contrast, discriminative models can directly model the classification boundary and typically make fewer model assumptions. They have been applied in many domains of text processing such as text classification and information extraction. Moreover, it is possible for retrieval systems to provide different types of ranking features from completely irrelevant sources including both query-dependent features and query-independent features. Generative models might have difficulties manually postulating different model distributions for these outputs. Nalapati [11] has shown that with presence of heterogenous features, discriminative models are superior to generative models in a home-page finding task.

Taking these factors into account, we decided to adopt discriminative models as the basic retrieval framework. Formally, we model the posterior probability of the relevance as a logistic function on a linear combination of ranking features, i.e.,

$$P(y_j | \lambda, D_j) = \frac{1}{Z_j} \exp \left(y_j \sum_{i=0}^N \lambda_i f_i(D_j) \right), \quad (1)$$

where Z_j is the normalization factor to construct a correct probability distribution and λ_i is the combination parameter for the output of i^{th} ranking features $f_i(D_j)$. This logistic regression model presented in Eqn(1), a.k.a. the maximum entropy model, summarizes our basic retrieval framework. It naturally provides a probabilistic interpretation for the retrieval outputs. The weights λ_i can either be learned from existing training data [14, 11] or assigned by predefined rules (e.g., if they are set to the tf.idf weights of query terms, this probabilistic model will reduce to a vector space model [2]). Once the parameters are estimated, documents can be presented to users in a descending order of $P(y_j = 1 | \lambda, D_j)$, or equivalently the weighted sum of retrieval outputs $\sum_{i=0}^N \lambda_i f_i(D_j)$. Note that, this linear combination model can be recovered to a large number of standard retrieval models with appropriate choices

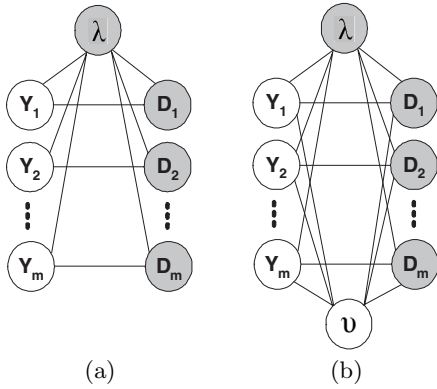


Figure 1: (a) The graphical model representation for the discriminative probabilistic retrieval model, where the document relevance is only determined by the initial weights λ and the ranking features, (b) The graphical model representation for PLF, which assumes the weights ν of unweighted ranking features to be latent variables. The nodes with known values are shaded, while other nodes are unshaded.

of query/document features, such as the tf.idf-based vector space model and the Okapi-family retrieval models.

To concisely represent the retrieval model, we can summarize document relevance variables into a vector and remove the normalization factor by using a proportional sign,

$$p(\mathbf{y}|\lambda, \mathbf{D}) \propto \prod_{j=1}^{M_D} \exp\left(y_j \sum_{i=0}^N \lambda_i f_i(D_j)\right). \quad (2)$$

where all the bold letters represent a vector of the corresponding scalar variables, e.g., \mathbf{y} means $\{y_1, \dots, y_{M_D}\}$. In the following discussions, we assume λ and $f_i(D_j)$ are already known, and only the distribution of variable y needs to be estimated. The graphical model representation of the discriminative retrieval model is shown in Figure 1(a), where the document relevance is determined by the initial combination weights λ together with the corresponding ranking features $f(D_j)$, and thus the document relevance variables Y_j are conditionally independent to each other. Therefore, Eqn(1) is essentially equivalent to Eqn(2) with a normalization factor that is only relevant to $\{\lambda, \mathbf{D}\}$.

3. PROBABILISTIC LOCAL FEEDBACK

In this section, we propose a new retrieval model called probabilistic local feedback(PLF), followed by describing its inference approach as well as its connections to other methods. Finally we conclude with an illustrative example on a two-dimensional synthetic dataset.

3.1 Model Description

Although the basic retrieval model defined in Eqn(2) provides a principled probabilistic framework for retrieval, it may suffer from the fact that only a very small proportion of the ranking features are utilized (i.e., associated with a non-zero weight) and the other features remain unweighted in the retrieval process. Simply stated, the reason is twofold: 1) general users tend to provide short queries [20], and 2) external training data are usually sparse or they do not exhibit common patterns on most of the features. However, it

does not imply these *unweighted ranking features* are useless for every query. Instead, since general users only have limited knowledge on underlying data collections, they might not be able to discover the most relevant features to express their own information need. Such a “mismatch” issue can greatly impede the accuracy of retrieval results.

In order to address this problem, we aim to automatically expand additional ranking features that are closely related to the original query description and can better capture the content of underlying data collections. This expansion is achieved by treating the weights of unweighted ranking features as latent variables rather than simply setting them to be 0. In more detail, let us assume that the initial retrieval results are generated from some ranking features $f_i(D_j)$ and a set of initial weights λ_i . Based on this setting, we further introduce a new latent weight ν for each unweighted ranking feature, i.e., the feature of which the associated λ is 0, where ν is a random variable ranged from $[-\infty, \infty]$. For example, if we give a query “finding a building” to a multimedia retrieval system with 10 different semantic ranking features available (such as “outdoors” and “building” which are learned from pre-collected visual examples), and initially the retrieval system decides only the weight λ for “building” is 1 based on word spotting, then we can introduce the latent weights ν for all the other 9 semantic features instead of simply setting them to be zero.

To enable both λ and ν to exert a joint effect on the document relevance variables Y_j , we follow the definition of conditional random fields [8] to derive the conditional probability of relevance \mathbf{y} , latent weights ν given initial weights λ and documents \mathbf{D} as,

$$p(\mathbf{y}, \nu|\lambda, \mathbf{D}) \propto \prod_l p_0(\nu_l) \prod_{j=1}^{M_T} \exp\left(y_j \sum_{i \in W} \lambda_i f_i(D_j) + y_j \sum_{i \in U} \nu_i f_i(D_j)\right). \quad (3)$$

where $W = \{i : \lambda_i \neq 0\}$ contains the indices of initially weighted ranking features, $U = \{i : \lambda_i = 0\}$ contains the indices of unweighted ranking features, ν_l is a latent combination weight for the l^{th} unweighted concepts.² For the sake of efficiency, we select only the top M_T documents in the initial ranking to update in Eqn(3), where M_T is a smaller number than the number of documents in the entire collection M_D . The settings and experiments w.r.t. M_T are described in the experiment section.

A comparison between the models described in Eqn(2) and Eqn(3) shows that both retrieval models share the same components $\exp(y_j \lambda_i f_i(D_j))$ to capture the effect of the initial retrieval weights. But the proposed model further introduces an additional set of components $\exp(y_j \nu_i f_i(D_j))$ so as to model the connections between unweighted concepts and document relevance. The prior distribution $p_0(\nu_l)$ offers another level of modeling flexibility, which represents how likely an unweighted ranking feature is relevant to information needs based on human prior knowledge or external knowledge sources. For example, we can model the prior as a normal distribution $\mathcal{N}(\nu_l^0, \sigma^2)$ where ν_l^0 reflects our prior belief on the weight for l^{th} ranking feature and σ^2 is a pre-defined constant variance. In case when no prior knowledge is available, we can set all the ν_l^0 to be 0. But if we have

²We use the proportional sign to indicate the intractability to compute the normalization factor on the right hand side.

a semantic lexicon at our disposal (e.g., WordNet), we can define $\nu_l^0 > 0$ if the l^{th} term is closely connected to a query term in the semantic lexicon, otherwise $\nu_l^0 = 0$.

In the following discussions, we refer the model presented in Eqn(4) to as the *probabilistic local feedback* model. Its graphical model representation is shown in Figure 1(b). In fact, the construction under undirected graphical model semantics is of crucial importance for the correct functionality of PLF. The conditional dependence between λ and ν allows the posterior probability of $p(\nu|\lambda, \mathbf{D})$ to be updated according to the initial weights and the ranking features in the search collection. Therefore, the proposed PLF model is able to estimate the effectiveness for each unweighted feature and discover useful features from the query context. In contrast, if we switch the model representation to be a directed graph, the latent combination weight ν will be then independent to the initial ranking results λ given y_j is unknown. Since no additional information from λ can flow to the nodes of ν , a directed graphical model will trivially produce the same retrieval outputs as the original ones.

Note that because both λ_i and $f_i(D_j)$ are already known, we can pre-compute the initial retrieval results $f^\lambda = \sum_i \lambda_i f_i$ and simplify the Eqn(3) to be,

$$p(\mathbf{y}, \nu | \lambda, \mathbf{D}) \propto \prod_l p_0(\nu_l) \prod_{j=1}^{M_T} \exp \left(y_j f^\lambda(D_j) + y_j \sum_{l \in U} \nu_l f_l(D_j) \right). \quad (4)$$

This simplification gives us an option to handle initial outputs provided by arbitrary retrieval systems, even if their outputs are not in form of $\sum_i \lambda_i f_i$. However, in order to mitigate the effect of the heterogeneity of retrieval output distributions and deal with the case when the retrieval scores are not available in practice, it could be beneficial to re-normalize f^λ based on document ranks before running the following inference step. Given that only M_T documents are selected, we can linearly normalize the conditional probability $P(y_j | \lambda, D_j)$ of the j^{th} ranked document to be $\frac{M_T+1-j}{M_T+1}$, or equivalently, $f^\lambda(D_j) = \frac{1}{2} \log \left(\frac{M_T+1-j}{j} \right)$. Similarly, we also shift the mean of each ranking feature f_l to be 0. These normalization schemes are applied in the rest of this paper. We will develop and evaluate the other possible normalization schemes in the future.

3.2 Probabilistic Inference

According to the probabilistic ranking principle, documents need to be ranked in a descending order of the conditional probability of relevance, i.e., $p(\mathbf{y} | \lambda, \mathbf{D})$. By marginalizing out the latent variables ν , we can compute the conditional probability of document relevance y as follows,

$$p(\mathbf{y} | \lambda, \mathbf{D}) \propto \int_\nu \prod_l p_0(\nu_l) \prod_{j=1}^{M_T} \exp \left(y_j f^\lambda(D_j) + y_j \sum_{l \in U} \nu_l f_l(D_j) \right) d\nu. \quad (5)$$

However, because of the presence of the normalization constant on the right hand side, it is usually intractable to compute the posterior probability in Eqn(5) with an exact inference approach. Therefore, we resort to variational methods to provide an approximate inference for the intractable posterior distributions.

Specifically, we adopt the mean field approximation [12] in our derivation, which takes a factorized form of all singleton

marginals over the variables. The first step of the mean field approximation is to construct the following family of variational distributions,

$$q(\mathbf{y}, \nu) = \prod_j q(\nu_l | \beta_l) \prod_j q(y_j | \gamma_j),$$

as a surrogate to approximate the posterior distribution $p(\mathbf{y}, \nu | \lambda, \mathbf{D})$, where $q(\nu_l | \beta_l)$ is a Gaussian distribution with mean β_l and the same variance σ as the prior $p_0(\nu)$, $q(y_j | \gamma_j)$ is a Bernoulli distribution where $y_j = 1$ with a sample probability of γ_j and otherwise $y_j = -1$. After some mathematical manipulations (see details in Appendix), we can find that the variational distribution closest to $p(\mathbf{y}, \nu | \lambda, \mathbf{D})$ must satisfy the following fix point equations,

$$\begin{aligned} \gamma_j &= \left[1 + \exp \left(2f_j^\lambda + 2 \sum_l \beta_l f_{jl} \right) \right]^{-1} \quad j = 1 \dots M_T, \\ \beta_l &= \nu_l^0 + \sum_j (2\gamma_j - 1) \sigma_l^2 f_{jl} \quad l \in U. \end{aligned} \quad (6)$$

These equations are invoked iteratively until the change of KL-divergence is small enough. To understand how this approach is expected to work, it helps to run the updates for only one step. In our implementation, we start with the first update rule and initialize β_l to 0. The first equation attempts to adjust the relevance score for each of the M_T document being selected, where top-ranked documents have positive scores and bottom-ranked documents have negative scores. Given the previous judgments, the second equation aims to estimate the weights of each ranking feature. The weight is larger if the feature can better distinguish the positive and negative documents. To explain why PLF can identify the correct features to expand, let us suppose the l^{th} feature is a discriminative feature that should be expanded, such as $f_{jl} > f_{kl}$ when d_j is a relevant document and d_k is an irrelevant document. Since most of the documents are irrelevant and the mean of f_l is set to 0, the expected sum $\sum_j (2\gamma_j - 1) f_{jl}$ for irrelevant documents is close to 0. Therefore, as long as top-ranked documents contain more relevant documents than bottom-ranked documents, PLF can automatically expand f_l with a positive weight.

A nice property of these two update rules is that their convergence is almost always guaranteed. Upon convergence, we use the final $q(y_j | \gamma_j)$ as a surrogate to approximate the posterior probability $p(y_j | \lambda, \mathbf{D})$ without explicitly computing the integral. Since $q(y_j | \gamma_j)$ is a Bernoulli distribution, we can simply rank the documents in a descending order of the parameter γ_j as the retrieval outputs.

Note that, this iterative update process only involves simple mathematical operations on the ranking features, such as sum, product and exponential function. Moreover, it typically converges in a small number of iterations. Thus the proposed PLF approach can be implemented efficiently in a real retrieval system. For example, with a single 2.66GHz Intel CPU, it only takes less than 1 second to generate the final retrieval outputs for the default settings described in our experimental section.

Remark: The update process of PLF shares some characteristics similar to the traditional pseudo-relevance feedback(PRF) techniques in the sense that both of them aim to refine the retrieval outputs based on initial rankings. However unlike PRF, PLF does not require the assumption that most of top-ranked documents have to be relevant. Instead,

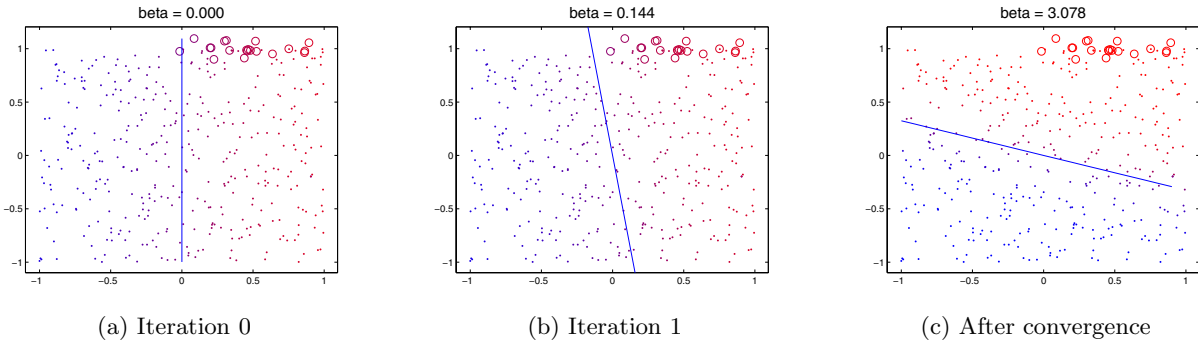


Figure 2: An illustrative example of PLF on a 2-D synthetic data set. X-axis stands for the initial retrieval outputs $f^\lambda(D_j)$ and Y-axis stands for the value of an un-weighted ranking feature $f(D_j)$. “o” and “.” denote the relevant and irrelevant documents. Red/blue colors denote the positive/negative predictions from PLF. (a) shows the synthetic dataset with the initial decision boundary as a solid line. Most of the top-ranked documents are irrelevant to the query. Above the figure, we show the value of the variational weights β . (b) shows the decision boundary after one iteration of PLF. (c) shows the decision boundary after its convergence. The relevant documents are ranked at the top after expansion.

it can work reasonably well as long as the top-ranked documents contain more relevant documents than the bottom-ranked documents. PLF also provides a sound probabilistic interpretation and a convergence guarantee on the iterative parameter updating process, which is usually missing in the PRF approaches.

PLF is also related to previous work that applied statistical learning algorithms to automatically improve existing ranking functions or retrieval models. For example, Collins et al. [1] considered a discriminative reranking approach using additional features of the trees to improve upon the initial ranking for natural language parsing. Tieu et al. [19] used boosting algorithms to choose a small number of features from millions of highly selective features for image retrieval. In the task of collaborative filtering, Freund et al. [4] proposed the RankBoost algorithm which learns to rank a set of objects by combining multiple “weak” classifiers to build up a more accurate composite classifiers. Taskar et al. [18] developed a learning method based on undirected probabilistic models to induce “unseen” features in the testing set. Their approach introduced a continuous hidden variable for each unseen feature to describe its influence on the class. The probabilistic inference over test data can estimate the distribution of hidden variables and thus outperform the learning method using training data alone.

3.3 Illustrative Examples

To show the ability of PLF to automatically expand useful ranking features and improve initial search results, we prepared a synthetic dataset shown in Figure 2(a) where X-axis shows the initial retrieval outputs $f^\lambda(D_j)$ and Y-axis shows the value of an unweighted ranking feature $f(D_j)$. In this figure, “o” and “.” represent the relevant and irrelevant documents. Blue and red colors represent the positive and negative predictions from PLF. There are a total of 20 relevant documents and 380 irrelevant documents. The documents are ranked in a descending order of their corresponding variational parameters γ_j , or equivalently, $f_j^\lambda + \sum_l \beta_l f_{jl}$. Figure 2(a) also uses a solid line to indicate the initial decision boundary, which shows the documents ranked at the medium positions. It is purely determined by $f^\lambda(D_j)$ at

Data Set	t03	t04	t05
Query Num	25	24	24
Doc Num	75850	48818	77979

Table 1: Labels of video collections and their statistics. t^{} indicate the search collection labels.**

the starting point. From this graph, we can observe that neither the initial prediction $f^\lambda(D_j)$ or the ranking feature $f(D_j)$ is perfect in predicting relevant documents, but both of them provide informative evidence to do so. More importantly, since the initial retrieval results provide a better-than-random performance, it can serve as a good initial setting to decide whether the unweighted feature should be expanded.

Figure 2(b) plots the decision boundary after running one step of the fix point equations in Eqn(8). It can be found that the variational parameter β becomes a positive number and the decision boundary is shifted to a more accurate position. Figure 2(c) shows the final decision boundary after the fixed point equations converge. It produces a much better retrieval results than the initial setting. This demonstrate the effectiveness of PLF even when the initial retrieval results are not so accurate.

4. EXPERIMENTS

We evaluate the proposed approach following the guidelines of the manual retrieval task in the TREC video retrieval evaluation (TRECVID) [16], which requires an automatic video retrieval system to search relevant documents without any human feedback. The main motivation for us to choose the multimedia collections as the testbed is to investigate whether PLF can perform robustly for the noisy initial retrieval outputs. Although the proposed approach can also be applied to the traditional text collections without any technical difficulties, we will leave it to the future work due to the time limits on the data preparation process.

Data	Initial	Expansion	MAP	P30	P100	Person	SObj	GObj	Sport	Other
t03	Text	None	0.146(+0%)	0.171	0.118	0.371	0.230	0.068	0.031	0.007
		PRF	0.149(+2%)	0.181	0.120	0.327	0.286	0.067	0.036	0.009
		PLF*	0.170(+16%)	0.197	0.125	0.399	0.321	0.067	0.035	0.008
	QClass	None	0.200(+0%)	0.236	0.137	0.466	0.336	0.088	0.106	0.015
		PRF	0.195(-2%)	0.232	0.137	0.441	0.333	0.088	0.104	0.017
		PLF	0.204(+2%)	0.231	0.136	0.472	0.354	0.087	0.103	0.015
	ApLQA	None	0.210(+0%)	0.249	0.144	0.463	0.358	0.106	0.103	0.017
		PRF	0.202(-3%)	0.252	0.144	0.434	0.354	0.102	0.106	0.017
		PLF*	0.230(+9%)	0.260	0.144	0.480	0.443	0.106	0.108	0.016
t04	Text	None	0.078(+0%)	0.178	0.107	0.188	0.012	0.033	0.046	0.044
		PRF	0.073(-5%)	0.175	0.105	0.160	0.007	0.037	0.042	0.053
		PLF	0.083(+6%)	0.184	0.108	0.189	0.006	0.035	0.072	0.047
	QClass	None	0.094(+0%)	0.199	0.125	0.194	0.080	0.046	0.108	0.045
		PRF	0.094(+0%)	0.180	0.125	0.198	0.080	0.050	0.098	0.044
		PLF	0.102(+9%)	0.207	0.120	0.211	0.058	0.047	0.120	0.055
	ApLQA	None	0.110(+0%)	0.220	0.119	0.255	0.053	0.044	0.110	0.051
		PRF	0.097(-12%)	0.187	0.119	0.219	0.030	0.045	0.099	0.044
		PLF	0.114(+4%)	0.238	0.124	0.258	0.056	0.041	0.130	0.056
t05	Text	None	0.073(+0%)	0.207	0.175	0.141	0.015	0.097	0.075	0.016
		PRF	0.079(+7%)	0.232	0.184	0.153	0.019	0.102	0.082	0.016
		PLF*	0.080(+10%)	0.231	0.185	0.154	0.026	0.104	0.081	0.016
	QClass	None	0.116(+0%)	0.292	0.211	0.173	0.031	0.100	0.322	0.017
		PRF	0.111(-4%)	0.281	0.211	0.169	0.031	0.094	0.294	0.018
		PLF	0.124(+7%)	0.304	0.223	0.193	0.035	0.100	0.337	0.017
	ApLQA	None	0.129(+0%)	0.294	0.212	0.209	0.042	0.104	0.328	0.017
		PRF	0.125(-3%)	0.294	0.212	0.210	0.048	0.103	0.283	0.017
		PLF	0.137(+6%)	0.315	0.220	0.225	0.044	0.116	0.335	0.017

Table 3: Comparison of three baseline approaches and their PRF/PLF-augmented retrieval outputs. The first column indicates the search collections and the second column indicates the baseline approaches. * means statistical significance over the baseline with p-value < 0.05 (sign tests).

Query	Useful features and their weights
Hu Jintao	Leader:1.0, Crowd:0.39, Airplane:0.07
Tony Blair	Leader:1.0, Commercial:-0.37, Crowd:0.93
Helicopter	Airplane:0.23, Sky:1.0
Fire/flame	Car:0.51, Building:0.64, Urban:0.93
Basketball	Crowd:0.53, Commercial:-0.35
Map of Iraq	Maps:1.00, Computer Screen:0.13

Table 2: Examples of useful ranking features (out of 75 visual concepts) and associated combination weights β found by PLF for six TRECVID’05 query topic. These features are provided by their corresponding semantic concept detectors built on development data.

4.1 Experimental Setting

In the following experiments, the retrieval units were video shots defined by a common shot boundary reference. The query topics contain multimodal information including text descriptions, image examples and video examples. We used the query topics and video collections from TREC’03-’05 to evaluate the proposed learning algorithms. Each of these video collections is split into a development set and a search set chronologically by source. The development sets are used as the training pool to develop automatic multimedia retrieval algorithms and the search sets mainly serve as the testbeds for evaluating the performances of retrieval systems. All of our experiments are evaluated on the search sets where for each query topic, the relevance judgment on search sets was provided officially by NIST. The development sets are only used to build the models for semantic concepts and learn the combination function in baseline methods. The computation of PLF has no relations to the development

sets. Table 1 lists the labels of each search collection and their statistics of query/document numbers.

As the building blocks of multimedia retrieval, we generated a number of ranking features on each video document including 75 high-level semantic concepts learned from development data (including face, anchor, commercial, studio, graphics, weather, sports, outdoor, person, crowd, road, car, building, motion and so forth), and 5 uni-modal retrieval experts (text retrieval, face recognition, image-based retrieval based on color, texture and edge histograms). For each retrieval expert, we transformed their raw scores into ranks and normalized the mean value to 0 in order to avoid the problems brought by inconsistent scales of various retrieval outputs. Each semantic concept is associated with a short text description, such as “*Car: segment contains video of an automobile*” and ground truth annotation on the development data. The detailed descriptions on the feature generation can be found in [6].

In order to improve the robustness of the PLF model, we apply a χ^2 test [23] to filter out some irrelevant ranking features before the inference process. The χ^2 statistics are generally computed to measure the dependence between two random variables. In our work, we use it to measure the independence between each feature and document relevance. If a feature tends to be independent of the relevance labels, this feature will be eliminated from the inference process. Only those features with a strong indication of their dependence are kept in the model. Under the assumption that irrelevant features are less likely to be strongly correlated with the relevance labels, the χ^2 test is able to eliminate most of the irrelevant features and improve the learning robustness, although a small proportion of relevant features might also be mistakenly discarded. In our experiments,

Data	Initial	Var.	MAP	P30	P100	Person	SObj	GObj	Sport	Other
t03	Text	1	0.170	0.197	0.125	0.399	0.321	0.067	0.035	0.008
		0.1	0.170	0.197	0.125	0.399	0.322	0.067	0.035	0.008
		10	0.168	0.191	0.121	0.401	0.316	0.064	0.034	0.009
	QClass	1	0.204	0.231	0.136	0.472	0.354	0.087	0.103	0.015
		0.1	0.205	0.235	0.137	0.472	0.354	0.087	0.105	0.015
		10	0.186	0.212	0.125	0.424	0.314	0.086	0.103	0.014
	ApLQA	1	0.230	0.260	0.142	0.480	0.443	0.106	0.108	0.016
		0.1	0.230	0.259	0.142	0.479	0.443	0.106	0.108	0.016
		10	0.211	0.237	0.133	0.424	0.409	0.103	0.106	0.015
t04	Text	1	0.083	0.184	0.104	0.189	0.006	0.035	0.072	0.047
		0.1	0.083	0.183	0.104	0.188	0.006	0.036	0.072	0.048
		10	0.085	0.203	0.105	0.196	0.006	0.031	0.075	0.051
	QClass	1	0.102	0.207	0.120	0.211	0.058	0.047	0.120	0.055
		0.1	0.103	0.207	0.120	0.212	0.058	0.047	0.120	0.055
		10	0.098	0.209	0.126	0.204	0.062	0.043	0.113	0.052
	ApLQA	1	0.114	0.238	0.124	0.258	0.056	0.041	0.130	0.056
		0.1	0.114	0.238	0.124	0.259	0.056	0.041	0.131	0.055
		10	0.113	0.238	0.125	0.250	0.053	0.041	0.128	0.058
t05	Text	1	0.080	0.231	0.185	0.154	0.026	0.104	0.081	0.016
		0.1	0.081	0.228	0.184	0.155	0.026	0.105	0.081	0.016
		10	0.080	0.225	0.187	0.152	0.026	0.107	0.081	0.016
	QClass	1	0.124	0.304	0.223	0.193	0.035	0.100	0.337	0.017
		0.1	0.125	0.315	0.223	0.192	0.035	0.103	0.340	0.017
		10	0.109	0.271	0.198	0.162	0.029	0.094	0.299	0.017
	ApLQA	1	0.137	0.315	0.220	0.225	0.044	0.116	0.335	0.017
		0.1	0.138	0.312	0.220	0.225	0.047	0.117	0.337	0.017
		10	0.128	0.306	0.216	0.222	0.036	0.112	0.282	0.017

Table 4: Comparison of variances σ^2 in the prior distribution.

we set the cutoff threshold to be 5.02, which corresponds to a confidence interval of 2.5% in the χ^2 distribution. Because the choice of threshold is relatively insensitive to the retrieval results, we do not show any experiments in this paper varying the χ^2 threshold.

4.2 Retrieval Results

To illustrate the ability of PLF to automatically expand useful ranking features, Table 2 lists six TRECVID’05 query topics (in the first column) together with their expanded ranking features and combination weights β found by PLF (in the second column). The query-class based combination method [22] is used to provide the initial search results for PLF. For each query, we normalized the highest combination weight to be 1 and discarded the ranking features when the absolute values of their combination weights are less than 0.05. It can be observed that most of the ranking features suggested by PLF are reasonable and closely related to the query topics. For example, for the query of “Hu Jintao”, it is easy to understand that the results can be augmented by using the visual concepts of “Government Leader” and “Crowd”. The appearance of “Airplane” can be explained by the fact that the truth video clips often contain arrival/departure scenes of Hu Jintao in the airport. These kinds of concepts are very difficult to find by merely analyzing the query description. Note that, the learned combination weights can be either positive or negative. For instance, the concept of “Commercial” is assigned a negative weight for the query of “Basketball”, which means that the basketball scenes usually do not contain any commercials.

Next, we present the retrieval performance of PLF as well as three baseline approaches, i.e., text retrieval (where only one ranking feature from text retrieval expert is used), query-class based combination [22] and adaptive probabilistic latent query analysis (ApLQA) [21]. The initial retrieval

outputs of PLF is provided by the baseline method that we are comparing with. We also compare PLF with a pseudo-relevance feedback algorithm, which assumes a subset of the top-rank examples to be positive and updates the combination parameters via the second equation in Eqn(3). For PRF, the number of feedback documents is chosen as the best configuration ranged from 50 to 500 at a step of 50 in the search collection (so PRF has an unfair advantage over PLF). To determine the parameter ν_l^0 in Eqn(3), we directly match query terms with the text description of semantic concepts. If there is a match between them, we set the corresponding ν_l^0 to be 1 in our experiments. This information is incorporated in both the baseline methods and the PRF method in order to provide a fair comparison. Among all 75 ranking features, we only consider expanding the unweighted ranking features that do not appear in the baseline retrieval functions. Unless stated otherwise, the prior variance σ^2 is set to 1 and M_T is set to 300, which means only the top 300 documents from the baseline outputs are updated.

Table 3 provides a detailed comparison between three baseline approaches and their PLF-augmented outputs on TRECVID’03-’05. All the retrieval results are reported in terms of the mean average precision(MAP) up to 1000 documents and precision at top 30, 100 documents. From this table, we can observe that the performance of PRF is inconsistent across multiple collections. This is partially because the initial results of multimedia retrieval are not sufficiently accurate to meet the requirement of PRF in general. In contrast, PLF can produce robust improvement for all the collections, because its assumption that requires *more* relevant documents on top is much easier to satisfy in practice. It is almost always superior to the baseline methods no matter which initial retrieval output and data collection are used. On average, it provides a roughly 1-2% absolute

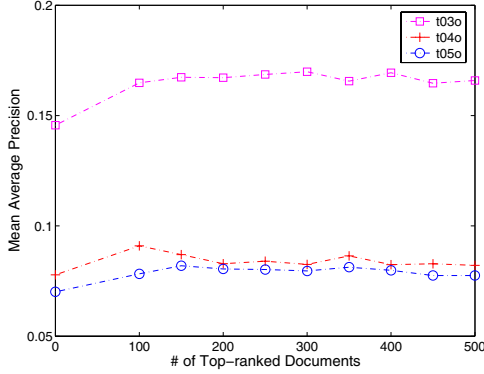


Figure 3: Learning curves vs. the number of updated documents M_T .

improvement in terms of mean average precision (or equivalently 10% relative improvement) over the baseline performance. The major performance growth factor for PLF can be traced to a higher precision/recall on the top-ranked documents. To analyze the results in more detail, we also grouped the queries in each collection and reported their MAP in five different categories, i.e., named person, special object, general object, sports and general queries. By comparing the retrieval performance with respect to each query type, we find that PLF benefits most from the person-type and special-object-type queries, as well as the sport-type queries in *t05*. This is because these query types can provide initial retrieval outputs of better quality, and thus they are able to be improved by making better use of additional ranking features.

To evaluate the sensitivity of PLF with respect to its parameters, we designed a series of experiments with text retrieval as initial outputs. Table 4 compares PLF with the prior variance σ^2 varied from 0.1 to 10. As we can observe, the setting of $\sigma^2 = 0.1$ is on par with the setting of $\sigma^2 = 1$. However, if we modify σ^2 to be a large value, e.g., 10 in our case, it might occasionally result in a large loss in terms of mean average precision. This suggests that PLF becomes more stable with a smaller variance in its prior potential, because large variances might dilute the useful information encoded in the prior distribution. Figure 3 depicts the learning curve of PLF with the number of updated documents M_T grown from 0 to 500 at a step of 50. In these three collections, although the highest mean average precision are achieved at different parameter settings, the fluctuation of mean average precision is typically less than 1% especially when M_T is larger than 200. It shows that PLF is not too sensitive to the variation on the number of updated documents.

5. CONCLUSION

In this paper, we propose a robust local analysis approach called probabilistic local feedback (PLF) based on a discriminative probabilistic retrieval framework. The proposed approach is effective to improve retrieval accuracy without assuming most top-ranked documents are relevant. This approach can be represented as an undirected graphical model by treating document relevance and weights of unweighted features as latent variables. Thus, it allows the information

from initial weights and collection statistics to jointly influence the expansion of ranking features. It also provides a sound probabilistic interpretation and a convergence guarantee on the iterative result updating process. Although derived from variational techniques, this approach can be computed efficiently in practice. Our multimedia retrieval experiments on three TRECVID collections have demonstrated the advantage of the proposed PLF approach, which achieves noticeable gains in terms of mean average precision over various baseline methods and PRF-augmented results. We expect that introducing external semantic knowledge sources in PLF could result in a further improvement on the retrieval performance. We also plan to investigate the effectiveness of PLF on the ad-hoc text retrieval task.

Appendix: Derivation of Eqn(8)

To approximate the distribution $p(\mathbf{y}, \nu | \lambda, \mathbf{D})$ using mean field methods, the first step is to construct the following family of variational distributions,

$$q(\mathbf{y}, \nu) = \prod_j q(\nu_l | \beta_l) \prod_j q(y_j | \gamma_j),$$

as a surrogate to approximate the posterior distribution $p(\mathbf{y}, \nu | \mathbf{a}, \mathbf{D})$, where $q(\nu_l | \beta_l)$ is a Gaussian distribution with mean β_l and the same variance σ as the prior potential $p_0(\nu)$, $q(y_j | \gamma_j)$ is a Bernoulli distribution where $y_i = 1$ with a sample probability of γ_j and otherwise $y_i = -1$. The independence between variables in the variational distributions results in the following efficient inference algorithm, which aims to optimize the KL divergence between $q(\mathbf{y}, \nu)$ and $p(\mathbf{y}, \nu | \lambda, \mathbf{D})$. This optimization can alternatively be cast into the maximization of the following lower bound,

$$\begin{aligned} 0 &\geq KL(q(\mathbf{y}, \nu) || p(\mathbf{y}, \nu | \lambda, \mathbf{D})) \\ &= E_q [\log p(\mathbf{y}, \nu | \lambda, \mathbf{D})] - E_q [\log q(\mathbf{y}, \nu)] \\ &= E_q \left[\sum_l \log p_0(\nu_l) + \sum_j y_j (f_j^\lambda + \sum_l \nu_l f_{jl}) \right] + H(q) \\ &= -\frac{(\beta_l - \nu_l^0)^2}{2\sigma_l^2} + \sum_j (2\gamma_j - 1)(f_j^\lambda + \sum_l \beta_l f_{jl}) \\ &\quad + \sum_j \gamma_j \log \gamma_j + \sum_j (1 - \gamma_j) \log(1 - \gamma_j), \end{aligned} \quad (7)$$

where f_{jl} denotes $f_l(D_j)$, f_j^λ denotes $f^\lambda(D_j)$, $E_q[f]$ refers to the expectation of $f(x)$ with respect to the distribution of $q(x)$, $H(q)$ refers to the entropy of the distribution q . It could be found that the gap between the inequality is exactly the K-L divergence between the variational posterior distribution $q(\mathbf{y}, \nu)$ and true posterior distribution $p(\mathbf{y}, \nu | \lambda, \mathbf{D})$. Therefore, we can alternatively solve a simpler optimization problem, i.e., to maximize the variational lower bound in order to find the best variational distributions to approximate the true posterior distribution.

By taking the derivative of the variational low bound with respect to the variational parameters γ, β to be zero, we can derive the following fixed point equations,

$$\begin{aligned} \gamma_j &= \left[1 + \exp \left(2f_j^\lambda + 2 \sum_l \beta_l f_{jl} \right) \right]^{-1} \quad j = 1 \dots M_T, \\ \beta_l &= \nu_l^0 + \sum_j (2\gamma_j - 1) \sigma_l^2 f_{jl} \quad l \in U. \end{aligned} \quad (8)$$

This completes the derivation of the fix point equations in Eqn(8). After running these update rules to convergence, we can use the resulted variational distributions to approximate the true distribution $p(\mathbf{y}, \nu | \lambda, \mathbf{D})$. Therefore, the marginal probability $p(y_j | \lambda, \mathbf{D})$ can be approximated by,

$$p(y_j | \lambda, \mathbf{D}) \approx \int_{\lambda} \int_{\mathbf{y} \setminus y_j} \prod_j q(\nu_l | \beta_l) \prod_j q(y_j | \gamma_j) = q(y_j | \gamma_j)$$

6. REFERENCES

- [1] M. Collins. Discriminative reranking for natural language parsing. In *Proc. of the 17th Intl. Conf. on Machine Learning*, pages 175–182, 2000.
- [2] W. B. Croft and D. J. Harper. Using probabilistic models of document retrieval without relevance information. *Document retrieval systems*, pages 161–171, 1988.
- [3] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 1990.
- [4] Y. Freund, R. D. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. In *Proc. of the 15th Intl. Conf. on Machine Learning*, pages 170–178, San Francisco, CA, USA, 1998.
- [5] N. Fuhr. Probabilistic models in information retrieval. *The Computer Journal*, 35(3):243–255, 1992.
- [6] A. Hauptmann, M.-Y. Chen, M. Christel, C. Huang, W.-H. Lin, T. Ng, N. Papernick, A. Velivelli, J. Yang, R. Yan, H. Yang, and H. D. Wactlar. Confounded expectations: Informedia at trecvid 2004. In *Proc. of TRECVID*, 2004.
- [7] Y. Jing and W. B. Croft. An association thesaurus for information retrieval. In *Proceedings of RIAO-94, 4th International Conference “Recherche d’Information Assistee par Ordinateur”*, pages 146–160, New York, US, 1994.
- [8] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th Intl. Conf. on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA, 2001.
- [9] J. Lafferty and C. Zhai. Probabilistic relevance models based on document and query generation. In *Language Modeling for Information Retrieval, Kluwer International Series on Information Retrieval*, volume 13. Springer, 2003.
- [10] V. Lavrenko and W. B. Croft. Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference*, pages 120–127, New York, NY, USA, 2001. ACM Press.
- [11] R. Nallapati. Discriminative models for information retrieval. In *Proc. of the 27th SIGIR conference on Research and development in information retrieval*, pages 64–71, 2004.
- [12] C. Peterson and J. Anderson. A mean field theory learning algorithm for neural networks. *Complex Systems*, 1:995–1019, 1987.
- [13] Y. Qiu and H.-P. Frei. Concept based query expansion. In *Proceedings of the 16th annual international ACM SIGIR conference*, pages 160–169, New York, NY, USA, 1993. ACM Press.
- [14] S. E. Robertson and K. S. Jones. Relevance weighting of search terms. *Journal of the American Society for Informaiton Science*, 27, 1977.
- [15] J. J. Rocchio. Relevance feedback in information retrieval. In *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice Hall, Englewood Cliffs, NJ, 1971.
- [16] A. Smeaton and P. Over. TRECVID: Benchmarking the effectiveness of information retrieval tasks on digital video. In *Proc. of the Intl. Conf. on Image and Video Retrieval*, 2003.
- [17] T. Tao and C. Zhai. Regularized estimation of mixture models for robust pseudo-relevance feedback. In *Proceedings of the 29th annual international ACM SIGIR conference*, pages 162–169, New York, NY, USA, 2006. ACM Press.
- [18] B. Taskar, M. F. Wong, and D. Koller. Learning on the test data: Leveraging unseen features. In *Proc. of the 20th International Conference on Machine Learning*, 2003.
- [19] K. Tieu and P. Viola. Boosting image retrieval. In *Intl. Conf. on Computer Vision*, pages 228–235, 2001.
- [20] J. Xu and W. B. Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Transaction Information System*, 18(1):79–112, 2000.
- [21] R. Yan and A. G. Hauptmann. Probabilistic latent query analysis for combining multiple retrieval sources. In *Proceedings of the 29th international ACM SIGIR conference*, Seattle, WA, 2006.
- [22] R. Yan, J. Yang, and A. G. Hauptmann. Learning query-class dependent weights in automatic video retrieval. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 548–555, 2004.
- [23] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *Proc. of the 14th ICML*, pages 412–420, 1997.