Probabilistic Models for Combining Diverse Knowledge Sources in Multimedia Retrieval

Rong Yan

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
yanrong@cs.cmu.edu

October 25, 2006

Abstract

In recent years, the multimedia retrieval community is gradually shifting its emphasis from analyzing one media source at a time to exploring the opportunities of combining diverse knowledge sources from correlated media types and context. This thesis presents a conditional probabilistic retrieval model as a principled framework to combine diverse knowledge sources. An efficient rank-based learning approach has been developed to explicitly model the ranking relations in the learning process. Under this retrieval framework, we overview and develop a number of state-of-the-art approaches for extracting ranking features from multimedia knowledge sources. To incorporate query information in the combination model, this thesis develops a number of *query analysis* models that can automatically discover mixing structure of the query space based on previous retrieval results. To adapt the combination function on a per query basis, this thesis also presents a *probabilistic local context analysis*(pLCA) model to automatically leverage additional retrieval sources to improve initial retrieval outputs. All the proposed approaches are evaluated on multimedia retrieval tasks with large-scale video collections as well as meta-search tasks with large-scale text collections.

1 Introduction

Recent improvement in processor speed, network systems, and the availability of massive digital storage has led to an explosive amount of multimedia data online [SO03, HBC$^+$03]. Already, according to network infrastructure company CacheLogic, more than 40 percent of Internet traffic is being taken up by peer-to-peer swaps that involve video content. Add to that the growing amount of legitimate content from companies such as Apple Computer, YouTube, and Google Video, and the scale of consumers' demand for video begins to emerge. However, if all these multimedia data are not manageable and accessible by general users,

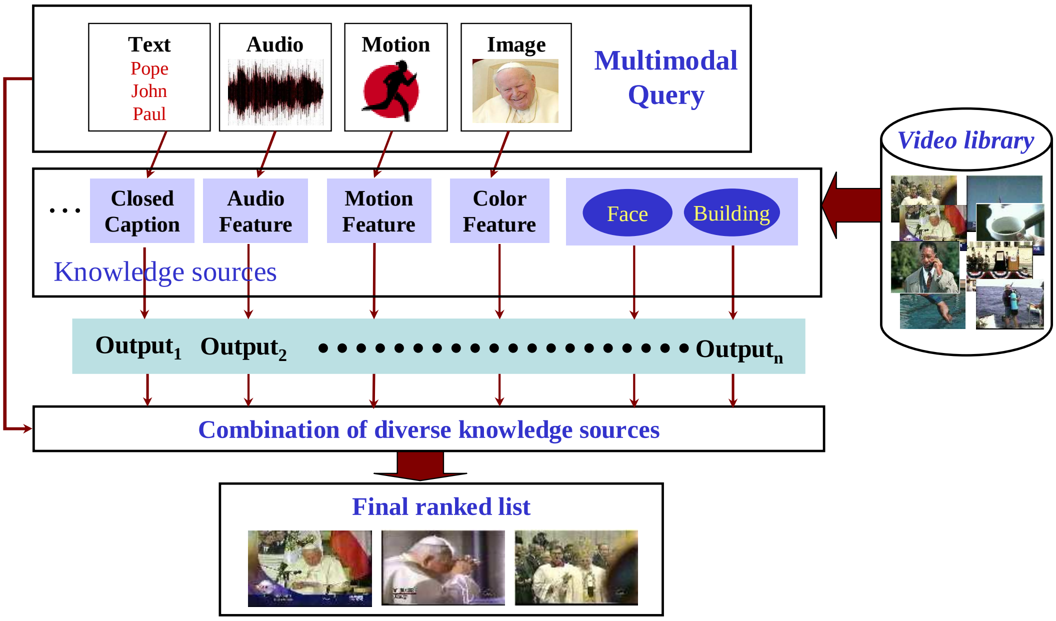


Figure 1: Design of multimedia retrieval systems.

they will become much less useful in practice. To achieve this, *multimedia information retrieval* systems, which aim to search a large number of multimedia data for documents relevant to an information need, offer an important platform to access and manage the vast amount of multimedia content online. Increasingly, this technique has drawn more and more attention from the extant web search engines (e.g. *Google*, *Yahoo!*, *Blinx.tv* and so on).

Given the luxury of accessing data from multiple streams, the multimedia retrieval community has been gradually shifting its emphasis from analyzing one media source at a time to exploring opportunities to select and combine diverse knowledge sources from correlated media types and context, especially as recent years have seen the rapid development of large-scale semantic concept detection techniques and retrieval approaches on various modalities. For example, searching multimedia collections for "the election of a US president" might need to leverage information from these contexts: restricting the search collections to news video, identifying the segments with persons on the screen and looking for persons speaking the word "election" in the audio stream. In order to develop a knowledge source combination strategy, two basic issues must be addressed, i.e., *what* to combine (i.e., identify available knowledge sources from multimedia data) and *how* to combine (i.e., develop effective combination strategies to merge multiple knowledge sources). Hence in the remainder of this chapter, we will elaborate both issues followed by summarizing the contributions of this thesis.

1.1 What to Combine

The early multimedia retrieval systems [Lew02, SLN$^+$02, WCGH99] usually model documents as a set of (low-level) detectable features generated from different modalities, such as tf.idf weights for text transcripts and color/texture histograms for images. The top ranked documents are retrieved by finding the most similar documents to the query examples on low-level feature space. However, the effectiveness for these low-level representations is usually limited owing to two reasons. First, text information alone in multimedia documents, e.g., speech transcripts and closed captions, is not predictive enough to support multimedia retrieval. More importantly, text information is not always be available in the multimedia data such as surveillance or soccer video. Second, although retrieval outputs can be augmented with other modalities such as visual features, these low-level representations typically result in the notorious *semantic gap* between users and retrieval systems, due to their inability to capture the semantic meaning of video content.

In order to better support multimedia retrieval, an intermediate layer of hundreds of *semantic concepts* has been introduced [NKH00] in an effort to capture the semantic content of multimedia documents. The typical concepts include a wide range of topics such as those related to people(face, anchor, etc), acoustic(speech, music, significant pause), objects(image blobs, buildings, graphics), location(outdoors, city, studio setting), genre(weather, financial, sports) and production(camera motion, blank frames) [CMC05]. The task of automatic semantic concept detection has been investigated by many researchers in recent years [BDF$^+$02, NKFH98, LTS03, YN05, YCH04, JLM03, WCCS04, SP02, SSL02, VJZ98, VFJZ99]. Their successes have demonstrated that a large number of high-level semantic concepts are able to be inferred from the low-level multi-modal features with a reasonable detection accuracy.

This thesis overviews and compares a number of state-of-the-art approaches for extracting ranking features from various multimedia knowledge sources, such as text retrieval, image retrieval and semantic concept detection in large-scale multimedia collections. These studies offer a useful guideline for researchers to select suitable algorithms to deal with different knowledge sources in multimedia systems. Meanwhile several novel approaches have been proposed to extract ranking features in a more effective way, e.g., SVM ensembles to handle rare classes, semi-supervised cross feature learning to leverage multimodal information, undirected graphical models to model concept relations and dual-wing harmoniums to discover hidden concepts.

1.2 How to Combine

Although a large body of work has been devoted to extract various ranking features from multimedia sources, relatively less attention has been given to this problem: *find a suitable strategy to combine diverse knowledge sources based on user information needs.* This task is not only a significant challenge but also offers great promise to provide considerable performance improvement. Until recently it remained unclear how these heterogenous knowledge sources could be systematically combined in the context of multimedia retrieval. The mainstream approaches rely on query independent combination strategies that can be either predefined as some combination functions or learned from some development sets. However in these approaches, the combination parameters would not be able to change across various queries and hence result in a considerable loss of flexibilities. For example, the query “finding George Washington” and the query “finding White House” should not share equivalent weights, because the former one prefers the outputs from face recognition and text retrieval whereby the latter one prefers the outputs from image retrieval.

In this thesis, we propose using a conditional probabilistic retrieval model as the basic framework to combine diverse knowledge sources in multimedia retrieval, which translated the retrieval task into a (supervised) learning problem with the parameters learned discriminatively. In contrast to the typical choices of generative models, a discriminative learning model is suggested for estimating the combination parameters in order to deal with heterogenous ranking features. We also propose an efficient rank learning approach called "ranking logistic regression" that can explicitly model the ranking relations in the learning process with much less computational efforts. Based on this probabilistic retrieval model we develop a two-stage learning approach so that the query information can be modeled in the knowledge source combination with a solid probabilistic foundation. It contains 1) a query analysis stage which can discover the mixing structure of the query space based on the past relevance judgments and 2) a context analysis stage which can automatically leverage additional ranking features to refine the initial retrieval results.

The query analysis approaches aim to adapt the combination functions for each unseen query by learning from past retrieval results. However, given the virtually infinite number of unseen queries, it is impractical to learn the combination function simply on a per query basis. A feasible alternative is to learn on some predefined query classes, i.e., associating combination weights with a few pre-defined classes which consist of queries with similar characteristics. In this case, it is legitimate to collect truth data for each query type because the number of types is very limited, while the learned weights can be reused for other unseen queries as long as they belong to some of the predefined classes. The effectiveness of this method has been demonstrated by previous work [YYH04, CNL$^+$04]. To extend the idea of query-class based retrieval, we also propose an approach called *probabilistic latent query analysis*(pLQA) inspired by the algorithm *probabilistic latent semantic analysis*(pLSI) [Hof99], with the goal of automatically discovering the mixing structure of the query space without explicitly defining query classes. Three pLQA models have been discussed which evolve from a basic version(BpLQA) to an adaptive version (ApLQA) that operates on the query feature space and a kernel version (KpLQA) that builds on a Mercer kernel representation. This formulation offers a probabilistic interpretation for latent query types, provides guideline to estimate number of query types and allows the mixing of multiple query types in a single query. A further extension of pLQA is called *hierarchical pLQA* model (HpLQA) that can model the distributions of query-specific combination components in a single query class via a hierarchical Bayesian model.

However, the effectiveness of query analysis is limited by the amount of training data. As a complementary method, context analysis aims to adapt the combination function specifically to the current query by means of treating the combination weights of ranking features as missing variables. In particular, a *probabilistic local context analysis*(pLCA) model is proposed to automatically leverage useful ranking features to improve the initial retrieval outputs. Formally, it can be described as an undirected graphical model that treats the document relevances and the weights of "un-weighted" features as a set of latent variables. In this model, the marginal dependence between initial retrieval results and latent concept weights allow the usefulness of each semantic concept to be determined in the retrieval process. In the case where human feedback is available, we also propose a pLCA variant to adjust the combination parameters based on human relevance feedback. An approximate inference algorithm is developed to accelerate the parameter updating process.

To evaluate the effectiveness of multimedia retrieval on large scale data collections, the proposed approaches are applied to the TRECVID collections, which contain queries from broad domains and multiple video collections of heterogenous topics. Well-established evaluation methodologies and metrics are adopted to demonstrate the effectiveness of our methods in the evaluation. Note that the proposed approaches are also applicable in other domains such as combining audio-visual outputs based on the environments, combining multiple search engines based on queries and combining multiple answers from various sources based on questions. To demonstrate the effectiveness of the proposed approaches outside the multimedia data, we also extend the experiments to a meta-search task that combines the outputs from multiple search engines to form a better ranked list.

In the following sections, we first overview the approaches of extracting ranking features from multimedia streams and then summarize our explorations on combining knowledge sources.

2 Overview on Ranking Feature Generation

The first step of multimedia retrieval is to extract a set of ranking features from multiple modalities, which will be combined into a ranking list in the subsequent process. In this section, we describe and compare a wide range of ranking feature generation approaches for multimedia data. Specifically, ranking features can be categorized into two types: query-dependent and query-independent features. The query dependent features are usually generated amidst the retrieval process by uni-modal retrieval components on low-level video features. They are used to indicate the similarity between query keywords/examples and multimedia documents in terms of a specific modality. Two examples of query dependent features are text retrieval outputs over the closed captions and image retrieval outputs over the color histogram. Therefore, we first investigate the performance of the state-of-the-art text retrieval and image retrieval approaches in multimedia retrieval systems. Numerous factors of text/image retrieval have been discussed in detail, including retrieval models, text sources, expansion window size, query expansion, visual features, similarity measures and their combination strategies. The retrieval performance with various settings are evaluated in TRECVID video collections and these experiments have confirmed the following conclusions.

In text retrieval, Okapi models usually provide better performance than vector space models, as is similar to previous experiments in pure-text collections. However unlike text collections, text retrieval in video corpus is relatively insensitive to the choice of document length normalization schemes. Among five predefined query types, text retrieval are most effective in the queries of finding persons and specific objects. Among all available text sources, closed caption provides the best performance, but speech transcript also achieves comparable results in term of average precision even with a word error rate around 20%. VOCR is shown to be useful in person-finding queries but not in others. Putting all text sources together is often superior to using any single source except for some rare cases. Expanding the text retrieval results to neighbor shots is an important strategy to mitigate the issue of timing misalignment between video clips and relevant text keywords. But for news video retrieval, it is beneficial to limit the expansion size inside the story boundary. Manual query expansion with a careful keyword selection can considerably improve text retrieval performance, especially for the queries related to sports events. But on the other hand, automatic query expansion based on WordNet or local feedback might degrade the retrieval performance if it is not handled properly.

In image retrieval, color-based features (especially color moment features) are one of the best choices with high effectiveness and low computational cost. The edge histogram can occasionally provide a comparable performance with color-based features but its performance is not as consistent. With respect to each query type, image retrieval is particularly useful for specific object type and sport type queries, of which the information need can be captured by the visual appearance of a limited number of image examples. But it produces relatively poor performance on the other query types. Being robust to outliers and efficient to compute, the L_1 distance is shown to be one of the most effective distance metrics in practice. To combine multiple query images, using the harmonic mean and maximum function outperforms the other fusion functions in terms of mean average precision. Combining the outputs of text retrieval and image retrieval can consistently improve the retrieval performance based on any single modality. Meanwhile, it is more robust and effective if the combination is done in a query-dependent way.

In contrast to query dependent features, the query independent features can be extracted and indexed in the databases before the retrieval process. The most widely used query independent features for multimedia retrieval are the indices of high level semantic concepts. The pool of semantic concepts usually covers a wide range of topics including objects, sites, events, specific personalities, named entities and so on. Therefore, we study the general approaches and discussed several open research directions for automatic semantic concept detection in multimedia collections. Typical semantic concept detection approaches are made up of four major steps: manual annotation, low-level feature extraction, supervised learning and multi-modal fusion. The successes of previous approaches have demonstrated that a large number of high-level semantic concepts can be directly inferred from low-level multimodal features without demanding a prohibitive amount of human annotation efforts. But several open research directions still need to be explored, such as balancing training distribution, leveraging unlabeled data, modeling relationship between concepts and extracting hidden concepts from video collections. Moreover, we contribute several novel approaches for semantic concept detection, e.g., SVM ensembles to handle rare class, semi-supervised cross feature learning to leverage multimodal information, undirected graphical models to model concept relations and dual-wing harmoniums to discover hidden concepts. Finally, our case study results confirmed that a few thousand semantic concepts could be sufficient to support high accuracy video retrieval systems. When sufficiently many concepts are used, even low detection accuracy can potentially provide good retrieval results as long as we find a reasonable way to combine them.

3 Basic Probabilistic Model for Multimedia Retrieval

3.1 Basic Formulation

In this section, we proposed using a relevance-based probabilistic retrieval model as a principled framework to combine diverse knowledge sources in multimedia retrieval. Formally, let us denote Q as the current query, $\mathcal{D} = \{D_1, ..., D_j, ..., D_{M_D}\}$ as a set of multimedia documents. Each document D_j is represented as a bag of ranking features $\{f_1, ..., f_N\}$. Let $y \in \{-1, 1\}$ indicates the relevance/irrlevance between a pair of document D and query Q. Consequently we denote the conditional probability of relevance as $P(y = 1|D, Q)$, or equally $P(y_+|D, Q)$. Under this representation, the multimedia retrieval problem can be formalized as follows: *Given a query Q, a document D_j, and their ranking features $f_i(D_j, Q)$, estimate a combination function of $F(Q, D_j, f_1, ..., f_N)$ to predict the conditional probability of relevance $P(y_+|D_j, Q)$.*

For the text-based retrieval, conventional relevance-based probabilistic models [Fuh92] rank documents by sorting the conditional probability that each document would be judged relevant to the given query, i.e., $P(y_+|D, Q)$. The underlying principle using probabilistic models for information retrieval is called *Principal Ranking Principle* [Rob77] that suggests sorting the documents D by the log-odds of relevance, where the odds ratio $O(y|D, Q)$ is defined as the ratio between $P(y_+|D, Q)$ and $P(y_-|D, Q)$. Following the idea of the binary independent model [RJ77], we directly model log-odd ratio to be a weighted linear combination function of ranking features, i.e.,

$$\log O(y|D, Q) = \sum_{i=0}^{N} \lambda_i(Q) f_i(D, Q). \tag{1}$$

In this section, we assume the combination weight $\lambda_i(Q)$ is independent of the query Q and rewrite the retrieval model as follows,

$$\log O(y|D, Q) = \sum_{i=0}^{N} \lambda_i f_i(D, Q).$$

or equivalently,

$$P(y_+|D, Q) = \sigma\left(\sum_{i=0}^{N} \lambda_i f_i(D, Q)\right) = \left[1 + \exp\left(-\sum_{i=0}^{N} \lambda_i f_i(D, Q)\right)\right]^{-1}, \tag{2}$$

where $\sigma(x) = 1/(1+e^{-x})$ is the standard logistic function and λ_i is the combination parameter for the i^{th} ranking feature $f_i(D, Q)$. Eqn(2) summarizes the basic multimedia retrieval framework and it naturally provides a probabilistic interpretation on the retrieval outputs. Interestingly, this retrieval model shares its connections with the traditional vector space models and the statistical language modeling approaches. The detailed discussions can be found in the full version of the thesis.

With all the ranking features available, the following step is to estimate the corresponding combination parameters λ_i for each ranking feature. We decide to use discriminative learning methods to estimate the combination parameters, because it can directly model the classification boundary with less model assumptions as compared to the generative learning approaches. Specially, we consider using logistic regression to estimate weights from training relevance judgment [Gey94]. Formally, given the relevance judgment y_{tj} for the training queries Q_t and documents D_j, the maximum-likelihood estimation for the combination parameters λ is as follows,

$$\begin{aligned} \lambda^* &= \arg\max_{\lambda} \prod_{t=1}^{M_Q} \prod_{j=1}^{M_D} P(y_+|D_j, Q_t)^{y'_{tj}} (1 - P(y_+|D_j, Q_t))^{1-y'_{tj}} \\ &= \arg\max_{\lambda} \sum_{t=1}^{M_Q} \sum_{j=1}^{M_D} \log \left[\sigma \left(y_{tj} \sum_{i=1}^{N} \lambda_i f_i(D_j, Q_t) \right) \right]. \end{aligned} \tag{3}$$

The maximum likelihood estimator λ^* has to be found numerically. Various numerical optimization algorithms, e.g., newton's method, conjugate gradient and iterative scaling, can be applied to optimize the log-likelihood of logistic regression. In the implementation, we adopt a well-known optimization algorithm called iterative reweighted least squares (IRLS).

Before the learning process, we need to use some feature selection method to choose a subset of predictive ranking features for the learning algorithm. In our experiments, we apply a χ^2 test [YP97] to make the selection. The χ^2 statistics is generally computed to measure the dependence of two random variables. In our case we use it to measure the dependence between each ranking feature and the relevance judgment. If a ranking feature is suggested to be independent to the document relevance (over all queries), this feature will be treated as query-specific and be eliminated in the learning process.

3.2 Ranking-based Learning Models

The aforementioned retrieval models cast information retrieval into a binary classification problem. Despite its great successfulness, such a classification framework might have difficulties in dealing with the retrieval task. For example, because there are only a small fraction of relevant examples in the collection, a classification algorithm that always provides negative prediction will unfortunately achieve a high predictive accuracy. Moreover, the classification accuracy has no relationship with the retrieval measure such as average precision. Therefore, recent years have seen a few attempts to develop learning algorithms that can explicitly account for ranking relations in information retrieval [Joa02, FISS98, Bea05, GQXN05]. Most of these rank learning approaches attempt to model the pairwise ranking preferences between every pair of relevant and irrelevant training examples. They are built on a solid foundation because it has been shown that minimizing the discordant pairs of examples are closely related to the commonly used ranking criteria. However, the effort of modeling every pair of examples often leads to a prohibitive learning process and thus limits their applications in practice.

In this section, we extend our basic model into its ranking-based counterpart based on a general rank learning framework. The basic idea is to switch the learning criterion from optimizing the classification mistakes to optimizing the number of discordant pairs Q between the predicted ranking and the target ranking. By introducing an convex loss function $L(\cdot)$, we can obtain the following unified margin-based rank learning framework,

$$\begin{aligned}\min_f RR_{rank}(f) &= \sum_{q_t \in Q} \sum_{d_j \in D_{qt}^+} \sum_{d_k \in D_{qt}^-} L(f(d_j, q_t) - f(d_k, q_t)) + \nu\Omega(\|f\|) \\ &= \sum_{q_t \in Q} \sum_{d_j \in D_{qt}^+} \sum_{d_k \in D_{qt}^-} L\left(\sum_{i=1}^{n} \lambda_i [f_i(d_j, q_t) - f_i(d_k, q_t)]\right) + \nu\Omega(\|f\|),\end{aligned} \quad (4)$$

where $f(d_j, q) = \sum_{i=1}^n \lambda_i f_i(d_j, q)$, $L(\cdot)$ is the empirical loss function, $\Omega(\cdot)$ is some monotonically increasing regularization function, $\|f\|$ is the norm of function f and ν is the regularization constant. This framework can be generalized to a large family of rank learning approaches such as Ranking SVMs [Joa02] and RankBoost [FISS98]. In our case, we are actually taking the logit loss $L_R(x) = -\log \sigma(x) = \log(1 + \exp(-x))$ as the empirical loss function.

To make the optimization less computational intensive, we further propose an approximate but efficient rank learning framework by approximating the pairwise risk function. The new loss function is designed based on the the following inequality,

$$RR_{prox}(f) \geq RR'_{rank}(f) \geq \frac{1}{2}[RR_{prox}(f) - RR_{prox}(-f)], \quad (5)$$

where $RR'_{rank}(f)$ is the pairwise ranking risk defined in Eqn(4) without the regularization factor and $RR_{prox}(f)$ is the approximate ranking risk function based on a shifted retrieval function $f^\alpha(d_j, q) = \sum_{i=1}^n \lambda_i [f_i(d_j, q) - \alpha_i]$,

$$RR_{prox}(f) = \sum_{q_t} \left\{ \sum_{d_j \in D_{qt}^+} M_D^- L\left(f^\alpha(d_j, q_t)\right) + \sum_{d_k \in D_{qt}^-} M_D^+ L\left(-f^\alpha(d_k, q_t)\right) \right\}.$$

In particular, we designed a new learning algorithm called ranking logistic regression(RLR) by plugging in the logit loss function, which has a similar form as classical logistic regression except the positive data are weighted stronger to balance the positive/negative data distribution and meanwhile the median value of each feature is shifted to zero. Formally, the new optimization function can be written as,

$$\max_\lambda \sum_{q_t} \left\{ \sum_{d_j \in D_{qt}^+} M_D^- \log \sigma\left(\sum_i \lambda_i f_{ijt}^*\right) + \sum_{d_k \in D_{qt}^-} M_D^+ \log \sigma\left(-\sum_i \lambda_i f_{ikt}^*\right) \right\},$$

where $f_{ijt}^* = f_i(d_j, q_t) - \alpha_i^*$ and α_i^* is the optimal feature shift that can be computed in a closed form. This learning algorithm serves as the basic learning framework for the model extension in the following sections.

4 Query Analysis

The aforementioned discriminative retrieval framework and its ranking counterpart provides a principled platform to support the task of retrieval source combination. However, in its current form, the combination parameters λ_i are still completely independent of the queries. In other words, the model will associate every possible query with the same set of combination parameters no matter what types of information needs users are expressing. In the previous chapter, we already showed that adopting such a query-independent knowledge combination strategy is not flexible enough to handle the variations of heterogeneous information needs.

Therefore, it is desirable to leverage the information from query description by developing more advanced retrieval methods. However, given the virtually infinite number of queries, it is impractical to learn combination functions on a per query basis. A trade-off needs to be found between the difficulty of providing training data and the ability of capturing the idiosyncrasy of each query. Following this argument, we propose a series of query analysis approaches which attempt to discover the mixing structure of past retrieval results and use the current query description as evidence to infer a better combination function.

4.1 Query-class Based Retrieval Model

In this section, we propose a retrieval approach called query-class based retrieval model, which aims to associate combination weights with a few pre-defined query classes and uses these weights to combine multi-modal retrieval outputs. In more detail, a user query is first classified into one of the four predefined query classes, i.e. finding named persons, named objects, general objects and scenes. Once the query is categorized, the ranking features from multiple modalities can be combined with query-class associated weights. In this case, it is legitimate to collect truth data for each query class because the number of classes is very limited, while the learned weights can be reused for other unseen queries as long as they belong to some of the predefined classes. The effectiveness of this model has been confirmed by our experiments on multimedia retrieval and many subsequent studies [CNL+04, CNG+05, Huu05, YXW+05, KNC05].

The design of query classes should follow two guidelines. First, the queries in the same query class should share similar combination functions. Second, queries should be automatically classified into one of the query classes with reasonable accuracy. After investigating the general queries for multimedia retrieval, we define the following four query classes according to the expressed intent:

Named person (P-query) queries for finding a named person, possibly with certain actions, e.g., “Find shots of Yasser Arafat” and “Find shots of Ronald Reagan speaking”.

Named object (E-query) queries for a specific object with a unique name, which distinguishes this object from other objects of the same type. For example, “Find shots of the Statue of Liberty” and “Find shots of Mercedes logo” are such queries.

General object (O-query) queries for a certain type of objects, such as “Find shots of snow-covered mountain" and "Find shots of one or more cats". They refer to a general category of objects instead of a specific one among them, though they may be qualified by adjectives or other words.

Scene (S-query) queries depicting a scene with multiple types of objects in certain spatial relationships, e.g., "Find shots of roads with lots of vehicles" and "Find shots of people spending leisure time on the beach". They differ from O-queries mainly by the number of the object types involved.

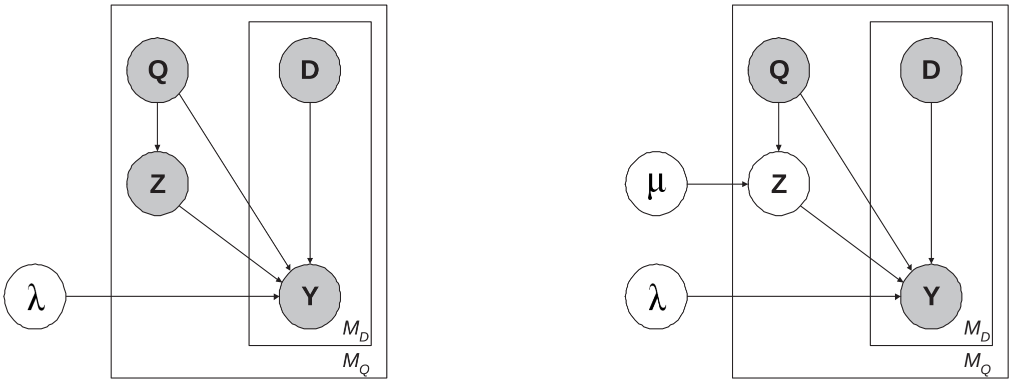
According to our definition, each query class should favor a specific set of ranking features. For example, face presence, size, position information and face recognition are critical to a P-query but of little value to other query classes. For both a P-query and a E-query, the text transcript is particularly important since such queries are more likely to have a perfect match in textual features, and so is video OCR because proper names may appear on the screen as overlaid text. On the other hand, visual features like color, texture, and shape can be helpful to the O-query and S-query. Overall, such a query classification approach captures the characteristics of queries regarding the feature effectiveness and therefore is promising for leading to a better performance.

After each query is associated with a single query class, the parameters can be estimated similarly as Eqn(3) except the training data are restricted in the given query class. The optimization of the log-likelihood function can be achieved similarly as optimizing a logistic regression problem.

4.2 Probabilistic Latent Query Analysis

Despite recent successes, query-class combination methods still have plenty of room for improvement. One major issue is that query classes usually need to be defined using expert domain knowledge. This manual design can work well when only a few query classes are needed, but it will become difficult for tens or hundreds of query classes, where each query in a class has to share similar characteristics and thus a similar combination strategy. If the query classes are not carefully defined, a large learning error from both the query-class assignment and the combination parameter estimation might result. Furthermore, current query-class methods do not allow mixtures of query classes, but at times such a mixture treatment could be helpful. For instance, the query "finding Bill Clinton in front of US flags" should be associated with both a "Person" class and a "Named Object" class rather than only one of these. Finally, determining the number of query classes remains to be an unanswered problem in these methods due to the nature of manual design. Some previous approaches [VGJL95, KNC05] can discover query classes by using clustering techniques. However, these approaches typically separate the processes of query class categorization and combination optimization into two sequential steps without being jointly optimized. Moreover, it is not straightforward for general clustering methods to handle mixtures of query classes in a principled manner.

Based on these considerations, it is desirable to develop a data-driven probabilistic combination approach that allows query classes and their corresponding combination parameters



(a) (b)

Figure 2: Graphical model representation for (a) query-class combination models where the query classes are manually defined, (b) probabilistic latent query analysis(pLQA) where the query classes are defined as latent variables. The nodes with known values are shaded, while other nodes are unshaded. The plates stand for replicas of the subgraphs where the number of replicas is on the corner.

to be automatically discovered from the training data itself, rather than handcrafted using human knowledge. Therefore, we propose a new combination approach called probabilistic latent query analysis (pLQA) to merge multiple retrieval sources based on statistical latent-class models. The proposed approaches have advantages over query-independent and query-class combination methods in several ways: (1) they unify combination weight optimization and query-class categorization into a single discriminative learning framework; (2) they are able to automatically discover the latent query classes directly from training data; (3) they can handle mixture of query classes in one query and (4) they can determine the number of query classes with an statistical model selection principle. Experiments are conducted on two retrieval applications, i.e., multimedia retrieval on the TRECVID'02-'05 collections [SO03] and meta-search on the TREC-8 collection [VH99]. The results show that the proposed approaches can uncover sensible latent classes from training data, and also demonstrate higher effectiveness in combining multiple retrieval sources. In the rest of this section, we first discuss the basic form of the pLQA method and then extend pLQA to its adaptive version and kernel version.

4.2.1 Basic pLQA

It would be ideal if we could learn specific combination parameters for every possible query. However, given the virtually infinite number of query topics, it is impractical to learn the combination weights on a per query basis because we cannot collect enough training data individually. We need to come up with a trade-off to balance the difficulties of collecting training data and the ability to capture the idiosyncracy of the query space. To achieve this, we make the following assumptions in our models: (1) the entire query space can be described by a finite number of query classes, where queries from each class share the same combination function; (2) the query description can be used to indicate which class a query belongs to. Under the first assumption, the basic probabilistic retrieval model presented in the last chapter can be naturally extended to a finite mixture of conditional probabilistic models. Formally, we can introduce a multinomial latent query class variable z to indicate which mixture the combination function is drawn from. Based on the second assumption, the choice of z is solely depending on the query Q. Putting all these together, we have the joint probability of relevance y and latent variable z as,

$$P(y_+, z|Q, D; \mu, \lambda) \;=\; P(z|Q; \mu)P(y_+|Q, D, z; \lambda), \tag{6}$$

where μ is the parameter for multinomial distributions, λ is the combination parameter for query classes. The mixture component $P(y_+|Q, D, z; \lambda)$ corresponds to a single logistic regression model and the mixing proportion $P(z|Q; \mu)$ controls the switches among different classes based on the query-dependent parameters μ_Q. By marginalizing out the hidden variables z, the corresponding mixture model can be written as, (M_z is the number of query classes)

$$P(y_+|Q, D; \mu, \lambda) = \sum_{z=1}^{M_z} P(z|Q; \mu) \cdot \sigma \left(\sum_{i=1}^{N} \lambda_{zi} f_i(D, Q) \right). \tag{7}$$

In the following discussions, we refer the model presented in Eqn(7) to as the *basic pLQA* (**BpLQA**) model where each latent query class represents a group of similar queries sharing the same combination weights. Its parameters can be estimated by the EM algorithm.

Figure 3(ab) compare the probabilistic graphical model representations of BpLQA and the query-class combination methods. One of the their major differences is the semantic of the mixing proportions $P(z|Q, \mu)$. In the query-class method, query classes have to be derived from manually defined generation rules before the learning process. However, query classes in the BpLQA model are expressed as latent variables and thus can be estimated directly from training data. Note that when the number of latent variables is reduced to 1, BpLQA degrades to the case where retrieval source combination is not relevant to queries, i.e., a query-independent combination approach.

4.2.2 Adaptive pLQA

Discovering the underlying structure of query space by itself is not sufficient to handle the retrieval source combination, because a practical combination model should be able to predict combination parameters for unseen queries outside the training collection. Unfortunately, BpLQA cannot easily generalize the multinomial parameters μ to any of these unseen queries, because each parameter $\mu_{\cdot t}$ in BpLQA specifically corresponds to the t^{th} training query. Since it is impossible to enumerate all possible queries in the training set, we need to come up with a solution to predict the mixing proportions $P(z|Q_t; \mu)$ for any unseen queries that do not belong to the training collection. To generalize the parameters to new documents, Hofmann [Hof99] suggested a "fold-in" process for the latent class model by re-learning all training documents with the new document to generate an updated parameter estimation. However, this "fold-in" process by plugging in new queries and re-estimating the entire model is not reasonable in our task, because it requires a long time to process and more importantly, we do not have any relevance judgment for new queries to learn from.

To address this problem, we propose an adaptive approach aiming at parameterizing the mixing proportion $P(z|Q_t;\mu)$ using a specific set of features directly extracted from query topics, or called query features. They are able to capture important characteristics of users' information need. Formally, we can represent each query as a bag of query features $\{q_1, ...q_L\}$. The mixing proportions $P(z_k|Q;\mu)$ can then be modeled using a soft-max function $\frac{1}{Z}\exp(\sum_l \mu_{zl} q_l)$, where $Z = \sum_z \exp(\sum_l \mu_{zl} q_l)$ is the normalization factor that scales the exponential function to be a probability distribution. By substituting the mixing proportion back into Eqn(7), the BpLQA model can be rewritten as,

$$P(y_+|Q,D) = \frac{1}{Z}\sum_z \exp(\sum_l \mu_{zl} q_l)\sigma\left(\sum_{i=1}^{N} \lambda_{zi} f_i(Q,D)\right). \tag{8}$$

Note that, since μ_{zl} is associated with each query feature instead of each training query, this modification allows the estimated μ_{zl} to be applied in any unseen queries as long as they can be formulated as vectors of query features. In the following discussions, we refer the model expressed in Eqn(8) to as the *adaptive pLQA*(**ApLQA**) model.

The only remaining issue for defining the ApLQA model is to design a set of predictive query features. There are two useful principles to guide the design of suitable query features: 1) they should be able to be automatically generated from query descriptions or the statistics of ranking features, and 2) they should be predictive to estimate which latent classes the query belong to. For example, we can consider the presence/absence of specific person names in query topics, and the mean retrieval scores of each retrieval source as query features. There exists a trade-off for choosing the number of query features. Introducing more query features can represent the information need more precisely, but meanwhile the learning process will also demand more training data to achieve a stable estimation. It is still an open research direction to choose the optimal number of query features. Note that, in contrast to creating query classes that must be mutually exclusive, defining query features is much more flexible, eliminating the need to partition the query space into non-overlapping regions. Moreover, the number of query features can be much larger than the number of query classes with the same amount of training data.

4.2.3 Kernel pLQA

By introducing explicit query features into the combination function, ApLQA can handle unseen queries that do not appear in the training data. However, the assumption of linear query feature combination in ApLQA might not be the best choice in general because the number of query features is often limited. Also, there exists some useful query information that cannot be described by explicit query feature representation. For example, the edit distance between two queries is a helpful hint for combination but it cannot be easily represented as a explicit query feature. Therefore, we develop an extension of the ApLQA model called the *kernel pLQA*(**KpLQA**) model that lends itself to the use of implicit feature space via Mercer kernels based on the representer theorem [KW71]. This extension is also motivated by a significant body of recent work that has demonstrated kernel methods are effective in a variety of applications.

In more detail, the kernel representation allows simple learning algorithms to construct a complex decision boundary by projecting the original input space to a high dimensional feature space, even infinitely dimensional in some cases. This seemingly computationally intensive task can be easily achieved through a positive definite reproducing kernel K and the well-known “kernel trick”. With this kernel representation, we can derive the corresponding log-likelihood function by substituting original mixing proportion term $P(z|Q_t)$ to be,

$$P(z|Q_t) = \frac{1}{Z}\exp(\sum_{k=1}^{M_D} \alpha_{zk} K(Q_t, Q_k))$$

ApLQA is a special case of KpLQA if each element of K is chosen to be the inner product between the query features of two queries. However, the flexibility of kernel selection has offered more powers to the KpLQA model. For example, the kernel function can have different forms such as the polynomial kernel $K(u,v) = (u \cdot v + 1)^p$ and the Radial Basis Function (RBF) kernel $K(u,v) = \exp(-\gamma\|u-v\|^2)$. The latter one has the ability to project the query features into an infinite dimensional feature spaces. Moreover, we can transform the distance metric between queries (e.g., edit distance between queries) into the implicit feature space in form of a Mercer kernel, instead of designing explicit features for each query. Also, it is possible to linearly combine multiple Mercer kernels obtained from various places to form another positive definite Mercer kernel.

4.2.4 Hierarchical pLQA

The pLQA model assumes that the combination parameters of each training query can be completely represented by the parameters of a latent query class λ_z. However, this model assumption is not always sensible because specific queries usually contain outlier noises beyond a limited number of latent query classes. Therefore, it might be worthwhile to explicitly model the query-specific variation in the phase of parameter estimation. To be more precise, we can express the combination parameters of the current query ω as the sum of two terms,

$$\omega = \lambda_z + \epsilon,$$

where λ_z is the deterministic combination parameters corresponding to the z^{th} query class and ϵ is a random variable drawn from Gaussian distribution $\mathcal{N}(0, \Sigma)$ that captures the query-specific variation in a single query class. Therefore, ω is a random variable drawn from a Gaussian distribution with mean λ_z and variance σ^2. Accordingly, the joint probability of relevance y, latent variable z and combination parameters ω can be written as,

$$P(y_+, \omega, z|Q, D; \mu, \lambda) \;=\; P(z|Q,\mu)P(\omega|z,\lambda)P(y|Q,D,\omega) \tag{9}$$

By marginalizing out the ω in the joint probability, the corresponding conditional probability of relevance can be written as,

$$\begin{aligned} P(y_+|Q,D;\mu,\lambda) &= \sum_z \int_\omega P(z|Q,\mu)P(\omega|z,\lambda)P(y|Q,D,\omega) \\ &= \sum_z \int_\omega d\omega P(z|Q,\mu)P(\omega|z,\lambda)\cdot\sigma\left(\sum_{i=1}^N \omega_i f_i(D,Q)\right). \end{aligned}$$

In this thesis, we call above retrieval model the *hierarchical pLQA model.* Note that when the variance of distribution $P(\omega|z,\lambda)$ goes to 0, the hierarchical pLQA model will naturally degrade to the pLQA model described in Eqn(7). Unfortunately, the exact inference in the hierarchical pLQA model is intractable. Therefore we usually resort to some approximate inference approaches such as *variational methods* to optimize the objective function.

5 Context Analysis

Query analysis offers a useful way to incorporate query information into the knowledge source combination, e.g., learning query independent combination models and query-class based combination models. However, since these learning approaches can only capture general patterns that distinguish relevant and irrelevant training documents, their power is usually limited by the number of available manual relevance judgments. If a high-level semantic concept is either very rare or has an insignificant discriminative pattern in the training data, it will simply be ignored by the learning algorithm without showing any effects in the combination function. In the rest of the paper, we call these *unweighted semantic concepts* in short. In fact, these unweighted semantic concepts constitute a major proportion of the available high-level semantic concepts. For example, even after learning with a large development collection including 500 hours of video from TRECVID'03-'05, a five query-class combination model still ignores more than 80% of the semantic concepts due to their inability to show strong patterns in the training documents. However, many unweighted semantic concepts are not completely worthless and occasionally they are helpful for the queries in related domains. For instance, the infrequent appearance of the concepts "ocean" and "sand" usually result in their absence in the learned retrieval function. But they can become highly predictive if the current query is "finding people on the beach".

Following the above discussion, we find that it is more desirable to develop retrieval approaches that can adaptively leverage unweighted semantic concepts on a per query basis without the support of training data. In this paper, we propose a new retrieval approach called probabilistic local context analysis(pLCA), which can automatically leverage useful high-level semantic concepts to improve the initial retrieval output. Formally, it can be described as an undirected graphical model that treats the document relevances and the combination weights of concepts as a set of latent variables. In this model, the marginal dependence between initial retrieval results and latent concept weights allow the usefulness of each semantic concept to be determined in the retrieval process. Moreover, we also propose a pLCA variant that can incorporate human relevance feedback into the learning process.

Our video retrieval experiments on TREC'03-'05 collections have demonstrated the effectiveness of the proposed pLCA approaches, which can achieve noticeable performance gains over various baseline methods.

5.1 Model Description

We begin with a review of the basic multimedia retrieval models, of which the underlying idea is to utilize discriminative models to combine multiple retrieval sources. Formally, we model the posterior probability of the relevance as a logistic function on a linear combination of ranking features, i.e.,

$$P(y|D,\lambda) = \sigma\left(y\sum_{i=0}^{N}\lambda_i f_i(D)\right), \tag{10}$$

where $\sigma(x) = 1/(1+e^{-x})$ is the standard logistic function and λ is the estimated combination parameter for the output of i^{th} ranking features $f_i(D)$. This logistic regression model, a.k.a. the maximum entropy model, summarizes our basic retrieval source combination framework. Once the parameters are estimated, documents can be presented to users in descending order of $P(y_+|D)$, or equivalently by the weighted sum of retrieval outputs $\sum_{i=0}^{N}\lambda_i f_i(D)$. By summarizing all the document prediction into a vector representation and eliminating the normalization factor, we can have

$$P(\mathbf{y}|D,\lambda) \propto \prod_{j=1}^{M_D}\exp\left(y_j\sum_{i=0}^{N}\lambda_i f_i(D_j)\right). \tag{11}$$

This formulation is equivalent to the previous representation due to the independence assumption between document relevances.

In order to automatically leverage additional semantic concepts, we propose a novel retrieval model called probabilistic local context analysis(pLCA) by considering the combination weights of unweighted semantic concepts as latent variables in an undirected graphical model. In more detail, we assume that we have generated all the ranking features for each document D_j and a set of initial combination parameters λ which can be estimated from a variety of methods such as manual definition and automatic learning. Then the combination weights ν corresponding to the unweighted semantic concepts, i.e., the semantic concepts that are not used to generate the initial outputs (equivalently, their associated λ are set to 0), are left as latent variables. Finally, both λ and ν will impose a joint effect on the document relevance variables Y_j. We expect that the distribution of unknown ν_i can be influenced by the initial retrieval results through Y_j and thus the useful ranking features can be selected without manual intervention.

Based on the model description, we can derive its corresponding graphical model representation as shown in Figure 3. In analogy to a conditional random field [LMP01], we can derive the conditional probability of relevance $\mathbf{y}$, latent weights ν given initial weights λ and

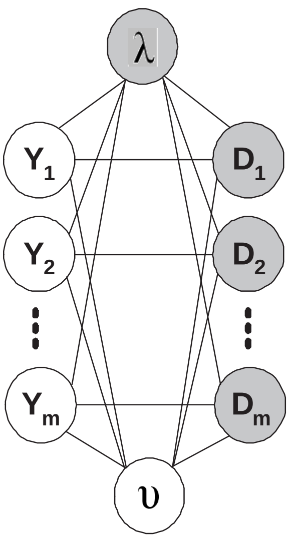


Figure 3: The graphical model representation for pLCA, which assumes the weights ν of unweighted semantic concepts to be latent variables. The nodes with known values are shaded, while other nodes are unshaded.

documents $\mathbf{D}$ as follows,

$$p(\mathbf{y}, \nu | \lambda, \mathbf{D}) \propto \prod_{l} p_0(\nu_l) \prod_{j=1}^{M_D} \exp\left(y_j \sum_{i \in W} \lambda_i f_i(D_j) + y_j \sum_{l \in U} \nu_l f_l(D_j) \right). \tag{12}$$

where $W = \{i : \lambda_i \neq 0\}$ contains the indices of initially weighted semantic concepts, $U = \{i : \lambda_i = 0\}$ contains the indices of unweighted semantic concepts, ν_l is a latent combination weight for the l^{th} unweighted concepts[1]. The prior distribution $p_0(\nu_l)$ represents how likely it is that an unweighted semantic concept is relevant to the user's information need only based on the query description. For example, the query of "finding the map of Iraq" can induce a high weight on the semantic concept of "map". However, this prior term is not able to capture the semantic concepts that are not explicitly mentioned in the query. Therefore, to further refine the document relevances, we use the potential $\exp(y_j \lambda_i f_i(D_j))$ to capture the effects from initial combination parameters and the potential $\exp(y_j \nu_l f_l(D_j))$ to model the connections from additional unweighted concepts. Since both λ_i and $f_i(D_j)$ are already given, we can pre-compute the initial retrieval results $f^\lambda(D_j) = \sum_i \lambda_i f_i(D_j)$ and simplify the Eqn(12) to be,

$$p(\mathbf{y}, \nu | \lambda, \mathbf{D}) \propto \prod_{l} p_0(\nu_l) \prod_{j=1}^{M_D} \exp\left(y_j f^\lambda(D_j) + y_j \sum_{l \in U} \nu_l f_l(D_j) \right). \tag{13}$$

[1] We use the proportional sign to indicate the intractability to compute the normalization constant on the right hand side. The same as follows.

5.2 Parameter Estimation And Inference

By marginalizing out the latent variables ν, we can present the documents to users in a descending order of the following conditional probability of y,

$$p(\mathbf{y}|\lambda, \mathbf{D}) = \int_{\nu} \prod_{l} p_0(\nu_l) \prod_{j=1}^{M_D} \exp\left(y_j f^{\lambda}(D_j) + y_j \sum_{l \in U} \nu_l f_l(D_j) \right) d\nu. \tag{14}$$

But because of the presence of the log-partition function in the undirected graphical model, it is usually intractable to compute the posterior probability in Eqn(14) with an exact inference approach. Therefore, we resort to variational methods to provide an approximate inference for the intractable posterior distributions. Specifically we can derive the following fixed point equations,

$$\begin{aligned} \gamma_j &= \left[1 + \exp\left(2f_j^{\lambda} + 2\sum_l \beta_l f_{jl} \right) \right]^{-1}, \\ \beta_l &= \nu_l^0 + \sum_j (2\gamma_j - 1)\sigma_l^2 f_{jl}. \end{aligned} \tag{15}$$

These fixed point equations can be interpreted as follows: the first equation attempts to assign a relevance judgment for each document, and the second equation aims to estimate the usefulness of each semantic concepts given the previous judgment. These equations are invoked iteratively until the change of KL-divergence is small enough. Upon convergence, we use the final $q(y_j|\gamma_j)$ as a surrogate to approximate the posterior probability $p(y_j|\mathbf{a}, \mathbf{D})$ without explicitly computing the integral. Since $q(y_j|\gamma_j)$ is a Bernoulli distribution, we can simply rank the documents in an descending order of the parameter γ_j as the retrieval outputs. Note that this iterative update process typically converges in a small number of iterations and thus the proposed pLCA approach can be implemented efficiently in the real retrieval system.

The update process of pLCA shares some similar characteristics as 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, pLCA does not require the assumption that most of top-ranked documents have to be relevant. Instead, it can work reasonably well as long as top-ranked documents contain more relevant documents than the bottom-ranked documents, which is sensible for video retrieval. pLCA also provides a sound probabilistic interpretation and a convergence guarantee on the iterative parameter updating process, which is usually missing in the PRF approaches. Moreover, it is not necessary for pLCA to specify a certain number of positive documents for the refinement process, since the initial prediction confidence has been naturally integrated in the probabilistic model.

5.3 Incorporate Human Feedback

The aforementioned pLCA approach can automatically leverage useful semantic concepts in an unsupervised manner. But in order to further improve the retrieval performance, pLCA can be augmented by incorporating additional human relevance feedback. Typically, relevance feedback algorithms proceed by first requesting users to annotate a small number of video documents from the initial retrieval results and then feeding them back to update the retrieval models. It can be viewed as a learning component in a retrieval system, which learns from a small amount of relevant examples to adjust the ranking function adaptively for additional annotations.

In this section, we mainly discuss how to use the additional annotation to modify the combination parameters ν in the pLCA model. Formally, we can denote the manual relevance judgment as $\{y_1, ..., y_K\}(y_k \in \{-1, 1\})$ associated with a set of documents $\{D_1, ..., D_K\}$. Given that a small number of documents have been annotated by the users, we can obtain a similar set of fixed point equations as before except that the variational parameters γ_j do not need to be updated any more on those annotated documents. Therefore, we can have

$$\begin{aligned} \gamma_j &= \left[1 + \exp\left(2f_j^\lambda + 2\sum_l \beta_l f_{jl}\right)\right]^{-1}, \\ \beta_l &= \nu_l^0 + \sum_j (2\gamma_j - 1)\sigma_l^2 f_{jl} + \sum_k y_j \sigma_l^2 f_{lk}. \end{aligned} \tag{16}$$

These update rules naturally incorporate the additional feedback information into the learning process. If we ignore all the variational decision γ_j and only consider the relevance judgment on the feedback documents, the update rules will degrade to a Rocchio-like updating process.

6 Conclusions and Future Directions

6.1 Conclusions

Multimedia information retrieval systems, which aim to search a large number of multimedia data for documents relevant to an information need, offers an important platform to access and manage the vast amount of multimedia contents online. In recent years, the research community of multimedia retrieval has been gradually shifting its emphasis from analyzing one media source at a time to exploring the opportunities to select and combine diverse knowledge sources from correlated media types and context. But it has always been a significant challenge to develop principled combination approaches and capture useful factors such as query information and context information in the retrieval process.

This thesis presents a conditional probabilistic retrieval model as a principled framework to combine diverse knowledge sources. As to our best knowledge, this is the first complete probabilistic model for multimedia retrieval that can handle multiple forms of ranking features, including query dependent features (uni-modal retrieval outputs) and query independent features (semantic concept indexing). It can also integrate multiple ranking features as well as query information and context information in a unified framework with a solid probabilistic foundation. In order to deal with heterogenous ranking features, a discriminative learning approach is suggested for estimating the combination parameters. Moreover, in order to incorporate the ranking information into the learning process, we also develop a general margin-based rank learning framework for the information retrieval task. An efficient approximation is proposed for the margin-based rank learning framework which can significantly reduce the computational complexity with a negligible loss in the performance.

Under this retrieval framework, we overview and compare a number of state-of-the-art approaches for extracting ranking features from various multimedia knowledge sources. These studies offer a useful guideline for researchers to select the suitable algorithms to deal with difference knowledge source in multimedia systems. Meanwhile, we develop several novel machine learning approaches for extracting ranking features, e.g., *SVM ensembles* to handle rare class, *semi-supervised cross feature learning* to leverage multimodal information, *undirected graphical models* to model concept relations and *dual-wing harmoniums* to discover hidden concepts.

To incorporate the query information into the combination process, we present two type of query analysis approaches. The first type is called query-class dependent retrieval model, of which the basic idea is to first classify each query into one of the predefined classes and then apply the associated combination weights to fuse the outputs from multiple retrieval sources. In order to automatically detect query classes from development data, we further propose a series of retrieval models called probabilistic latent query analysis (pLQA) to merge multiple retrieval sources, which unifies the combination weight optimization and query class categorization into a discriminative learning framework. Four pLQA models have been discussed which evolve from a basic version(BpLQA) to an adaptive version (ApLQA) that operates on the query feature space, a kernel version (KpLQA) that builds on a Mercer kernel representation, a hierarchical version that bases itself on a hierarchical Bayesian model. In contrast to the typical query-independent and query-class combination methods, pLQA can automatically discover latent query classes from the training data rather than relying on manually defined query classes.

Although query analysis offers a useful way to incorporate the factor of query information into combination models, their power is often limited by the number of available manual relevance judgment. In order to automatically leverage useful high-level semantic concepts without training data, we propose an automatic multimedia retrieval approach called probabilistic local context analysis(pLCA). This approach can be represented as an undirected graphical model by treating document relevances and combination weights of semantic concepts as latent variables. Thus, it allows the information from initial retrieval results to influence the selection of semantic concepts for the given query. Built on a sound probabilistic foundation, pLCA is effective for improving video retrieval without either learning a number of training data or assuming most top-ranked documents to be relevant. As an extension, we also propose a variant of the pLCA approach that can take human relevance feedback into account.

To evaluate the performance of the proposed approaches, we provide a thorough study on the standard multimedia collections and offer baseline performances for other researchers to compare with. We also want to emphasize that although the combination approaches developed in this thesis is motivated by the multimedia retrieval problem, their contributions and potential applications are not only limited to this domain. For example, most of the proposed approaches are also examined on the task of meta-search over large-scale text collections. The applicability of the proposed methods can be extended to many other areas involving knowledge source combination, such as question answering, web IR, cross-lingual IR, multi-sensor fusion, human tracking and so forth.

6.2 Future Directions

The proposed conditional probabilistic retrieval model offers many new opportunities to develop principled retrieval approaches for combining multimedia knowledge sources especially for multimedia data. The approaches we describe in this thesis only reveal a small tip of the full potentials of the proposed model. Many interesting future research directions can be explored as follows,

Model knowledge relations Typically, we assume each ranking feature is independently generated from a single knowledge source. However, knowledge sources are not isolated with each other. For example, the concept “outdoors” should have connections with the concept “sky”. Therefore, it is desirable to go beyond the independent assumption to model the relationship between ranking features. The tree-augmented term weighting scheme proposed by Van Rijsbergen [Rij79] can be viewed as a starting point to follow.

Handle missing features In practice, some ranking features might not always be available for entire collections, especially when the number of features is large. In this case, how to compensate the missing modalities and adjust the retrieval models accordingly will become an key issue. From another perspective, currently the semantics of ranking features have to be manually defined before the learning process, which usually requires a time-consuming human annotation process to support. It will become very interesting if we can discover latent knowledge sources within the learning process of combination models.

Refine query classes Introducing query classes into the combination model provides a practical way to handle the query information. However, the organization of query classes and the definition of query features still have room to improve. For instance, we can design a hierarchy layout for the query classes so as to capture the top-down organizations between general classes and specific classes. We can also investigate on designing better query features for ApLQA and introducing new kernels to KpLQA using pairwise distance metrics such as WordNet distance or edit distance.

Introduce user context Standard information retrieval systems treat each query in the same way without considering the user context. However, in a real retrieval scenario, a user’s information need is often iteratively refined based on a series of short-term retrieval sessions and long-term preference biases. Such an iterative refinement process suggests that the retrieval outputs should depend on the previous actions that the user has taken. For example, if the user is looking for “George Bush” in the previous query and “Rice” in the current query, the latter one will have higher possibilities to be referred to “Condoleezza Rice” rather than ”Rice University”. Indeed, the user context can be constructed from all the information about previous actions and user environments, such as previous query keywords, explicit/implicit feedback, personal profile, query time, query location and so on. These factors can be formally incorporated into the conditional probabilistic retrieval model via a prior function. It would be very interesting to extend the proposed framework so as to optimize long-term retrieval utility over a dynamic user context.

Exploit social intelligence Recent years have seen the emergence of social intelligence in a wide range of web-based applications. The most popular form of social intelligence is Google, which ranks pages on the basis of links they get. In other words, it simply uses the brains of others. Also that is exactly what Wikipedia is doing. In doing so, it is not only competing successfully with The New York Times, but also old standard, Encyclopedia Britannica. Following the similar ideas, we could consider how to leverage the power of social intelligence to improve the information retrieval quality, such as crowd-sourcing the video concept annotation to the huge amount of web users, utilizing the external knowledge shared on the web and exploiting the social network to uncover users' inherent connections.

Apply to other domains The applicability of the proposed approaches is not limited to the multimedia retrieval problems. They can be used in many other areas involving knowledge source combination, such as question answering, web retrieval, cross-lingual information retrieval, multi-sensor fusion, human tracking and so forth. We would like to explore these opportunities in our future work.

References

[BDF$^+$02] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3, 2002.

[Bea05] C. Burges and et al. Learning to rank using gradient descent. In *Proceedings of the 22nd intl. conf. on machine learning*, pages 89–96, 2005.

[CMC05] S. F. Chang, R. Manmatha, and T. S. Chua. Combining text and audio-visual features in video indexing. In *IEEE ICASSP 2005*, 2005.

[CNG$^+$05] T.-S. Chua, S.-Y. Neo, H.-K. Goh, M. Zhao, Y. Xiao, and G. Wang. Trecvid 2005 by nus pris. In *NIST TRECVID-2005*, Nov 2005.

[CNL$^+$04] T. S. Chua, S. Y. Neo, K. Li, G. H. Wang, R. Shi, M. Zhao, H. Xu abd S. Gao, and T. L. Nwe. Trecvid 2004 search and feature extraction task by nus pris. In *NIST TRECVID*, 2004.

[FISS98] 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.

[Fuh92] N. Fuhr. Probabilistic models in information retrieval. *The Computer Journal*, 35(3):243–255, 1992.

[Gey94] Fredric C. Gey. Inferring probability of relevance using the method of logistic regression. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 222–231, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[GQXN05] J. Gao, H. Qi, X. Xia, and J.-Y. Nie. Linear discriminant model for information retrieval. In *Proceedings of the 28th international ACM SIGIR conference*, pages 290–297, New York, NY, USA, 2005. ACM Press.

[HBC$^+$03] A. G. Hauptmann, R.V. Baron, M.-Y. Chen, M. Christel, P. Duygulu, C. Huang, R. Jin, W.-H. Lin, T. Ng, N. Moraveji, N. Papernick, C.G.M. Snoek, G. Tzanetakis, J. Yang, R. Yan, , and H.D. Wactlar. Informedia at TRECVID 2003: Analyzing and searching broadcast news video. In *Proc. of TRECVID*, 2003.

[Hof99] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proc. of the 22nd Intl. ACM SIGIR conference*, pages 50–57, 1999.

[Huu05] B. Huurnink. AutoSeek: Towards a fully automated video search system. Master's thesis, University of Amsterdam, October 2005.

[JLM03] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 119–126, 2003.

[Joa02] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the 8th ACM SIGKDD intl. conf. on knowledge discovery and data mining*, pages 133–142, New York, NY, USA, 2002. ACM Press.

[KNC05] L. Kennedy, P. Natsev, and S.-F. Chang. Automatic discovery of query class dependent models for multimodal search. In *ACM Multimedia*, Singapore, November 2005.

[KW71] G. Kimeldorf and G. Wahba. Some results on tchebycheffian spline functions. *J. Math. Anal. Applic.*, 33:82–95, 1971.

[Lew02] Michael Lew, editor. *Intl. Conf. on Image and Video Retrieval.* The Brunei Gallery, SOAS, Russell Square, London, UK, 2002.

[LMP01] 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.

[LTS03] C. Lin, B. Tseng, and J. Smith. VideoAnnEx: IBM MPEG-7 annotation tool for multimedia indexing and concept learning. In *IEEE International Conference on Multimedia and Expo*, 2003.

[NKFH98] M. R. Naphade, T. Kristjansson, B. Frey, and T.S. Huang. Probabilistic multimedia objects (multijects): A novel approach to video indexing and retrieval in multimedia systems. In *Proc. of ICIP*, 1998.

[NKH00] M. R. Naphade, I. Kozintsev, and T. S. Huang. On probabilistic semantic video indexing. In *Proceedings of Neural Information Processing Systems*, volume 13, pages 967–973, Denver, CO, Nov. 2000.

[Rij79] C. J. Van Rijsbergen. *Information Retrieval.* Butterworth-Heinemann, Newton, MA, USA, 1979.

[RJ77] S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Informaiton Science*, 27, 1977.

[Rob77] S. E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, 33(4):294–304, 1977.

[SLN$^+$02] J. R. Smith, C. Y. Lin, M. R. Naphade, P. Natsev, and B. Tseng. Advanced methods for multimedia signal processing. In *Intl. Workshop for Digital Communications IWDC*, Capri, Italy, 2002.

[SO03] A.F. 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.

[SP02] M. Szummer and R. Picard. Indoor-outdoor image classification. In *IEEE International Workshop in Content-Based Access to Image and Video Databases, Bombay, India*, Jan 2002.

[SSL02] N. Serrano, A. Savakis, and J. Luo. A computationally efficient approach to indoor/outdoor scene classification. In *International Conference on Pattern Recognition, Qubec City, Canada*, Aug. 2002.

[VFJZ99] A. Vailaya, M. Figueiredo, A. Jain, and H. Zhang. A bayesian framework for semantic classification of outdoor vacation images. In *SPIE Conference on Electronic Imaging, San Jose, California*, 1999.

[VGJL95] E. M. Voorhees, N. K. Gupta, and B. Johnson-Laird. Learning collection fusion strategies. In *Proc. of the 18th ACM SIGIR conference on Research and development in information retrieval*, pages 172–179, 1995.

[VH99] Ellen M. Voorhees and Donna Harman. Overview of the eighth text retrieval conference (trec-8). In *TREC*, 1999.

[VJZ98] A. Vailaya, A. Jain, and H.J. Zhang. On image classification: City vs. landscape. In *IEEE Workshop on Content-Based Access of Image and Video Libraries, Santa Barbara, CA*, Jun 1998.

[WCCS04] Yi Wu, Edward Y. Chang, Kevin Chen-Chuan Chang, and John R. Smith. Optimal multimodal fusion for multimedia data analysis. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 572–579, 2004.

[WCGH99] H. Wactlar, M. Christel, Y. Gong, and A. G. Hauptmann. Lessons learned from the creation and deployment of a terabyte digital video library. *IEEE Computer*, 32(2):66–73, 1999.

[YCH04] J. Yang, M. Y. Chen, and A. G. Hauptmann. Finding person x: Correlating names with visual appearances. In *Intl. Conf. on Image and Video Retrieval (CIVR'04)*, Ireland, 2004.

[YN05] R. Yan and M. R. Naphade. Semi-supervised cross feature learning for semantic concept detection in video. In *IEEE Computer Vision and Pattern Recognition(CVPR)*, San Diego, US, 2005.

[YP97] Yiming Yang and Jan O. Pedersen. A comparative study on feature selection in text categorization. In *Proc. of the 14th ICML*, pages 412–420, 1997.

[YXW$^+$05] J. Yuan, L. Xiao, D. Wang, D. Ding, Y. Zuo, Z. Tong, X. Liu, S. Xu, W. Zheng, X. Li, Z. Si, J. Li, F. Lin, and B. Zhang. Tsinghua university at TRECVID 2005. In *NIST TRECVID 2005*, Nov 2005.

[YYH04] 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.