INVITED PAPER

Query-Adaptive Fusion for Multimodal Search

Search systems need to have the flexibility to adapt to each query so a search strategy that is likely to provide the most useful retrieval results can be employed.

By Lyndon Kennedy, *Student Member IEEE*, Shih-Fu Chang, *Fellow IEEE*, and Apostol Natsev

ABSTRACT | We conduct a broad survey of query-adaptive search strategies in a variety of application domains, where the internal retrieval mechanisms used for search are adapted in response to the anticipated needs for each individual query experienced by the system. While these query-adaptive approaches can range from meta-search over text collections to multimodal search over video databases, we propose that all such systems can be framed and discussed in the context of a single, unified framework. In our paper, we keep an eye towards the domain of video search, where search cues are available from a rich set of modalities, including textual speech transcripts, low-level visual features, and high-level semantic concept detectors. The relative efficacy of each of the modalities is highly variant between many types of queries. We observe that the state of the art in query-adaptive retrieval frameworks for video collections is highly dependent upon the definition of classes of queries, which are groups of queries that share similar optimal search strategies, while many applications in text and web retrieval have included many advanced strategies, such as direct prediction of search method performance and inclusion of contextual cues from the searcher. We conclude that such advanced strategies previously developed for text retrieval have a broad range of possible applications in future research in multimodal video search.

KEYWORDS | Multimedia indexing and retrieval; multimodal search; query-adaptive fusion; query-class-dependent models

I. INTRODUCTION

As the reach and quality of search technologies continue to grow and mature, the developers of search systems have come to an important understanding that all queries are not created equal, meaning that applying a single standard search method for all possible queries is inadequate. Users can enter an infinite number of possible queries to represent an equally expansive set of information needs. A successful retrieval system needs to be able to interpret the users' queries, extrapolate their intentions, and then employ a search strategy that will be likely to return relevant results. In short, a successful search mechanism needs to adapt to each individual incoming query.

A key aspect of this emerging need for adaptive query strategies is that modern retrieval systems nearly universally incorporate cues from multiple diverse information sources. For example, early text retrieval systems relied on simple matching between counts of query keywords and their frequencies in the documents in the search set and this approach was incorporated in early incarnations of web search engines; however, the PageRank algorithm [44] showed that the structure of links between pages on the web is an equally important source of information about documents and their relative relevance to a query. Now, all serious web search engines incorporate a similar link structure analysis. Likewise, early image and video retrieval systems had separate tools for issuing queries via image examples, text keywords, and sketches, where each tool could be deployed independently [18], [19], [35], [36]. However, more recent systems allow for the use of many diverse query inputs and search methods simultaneously with the facility to combine the results from multiple modalities into a single, fused multimodal ranking of the relevance of images or videos in the search set [29]–[33].

A complication that arises when so many information sources and search methods are available to process an incoming query is that the relative utility of the available ranking approaches can be highly variable from query to query. For example, a search for a named person (like "Condoleezza Rice" or "Saddam Hussein") in a news video database, would be best served by a text search over the speech recognition transcripts, rather than an example

Manuscript received July 1, 2007.
L. Kennedy and **S.-F. Chang** are with the Department of Electrical Engineering, Columbia University, New York, NY 10027 USA (e-mail: lyndon@ee.columbia.edu; sfchang@ee.columbia.edu).
A. Natsev is with the IBM Watson Research Center, Hawthorne, NY 10532 USA (e-mail: natsev@us.ibm.com).
Digital Object Identifier: 10.1109/JPROC.2008.916345

image search. On the other hand, a search for a sports scene (like "tennis court" or "basketball players") would be greatly helped by the inclusion of example image search. Similar difficulties can arise in searching the web. Users of modern search engines have a variety of expectations for the results of different types of queries. Research-oriented queries (like "support vector machines") would be expected to return pages giving details and background related to the research query, whereas bookmark-replacement queries (like "IBM") would be expected to return a link to the homepage of the searched topic at the top of the results list.

Clearly, a *query-independent* strategy, where the same search mechanism is applied for every incoming query, regardless of the intentions of the user, is inadequate and can poorly serve many types of searches. The solution is to design a *query-adaptive* strategy, where the exact formulation of the search algorithm changes based on the intentions of the user. The user intentions, however, are difficult to measure and predict, since users in many search systems enter queries of only a few words. Understanding the user's intentions from such sparse input can be akin to mind reading. Many research efforts have proposed various methods for adapting retrieval strategies according to the query provided by the user in a number of search applications. One such strategy is to predefine a set of query "classes," which are sets of queries in which the optimal search strategies are similar for all queries contained within the class [3]–[5], [7], [11], [13], [14], [17]. Incoming queries can be classified into one of the classes based on some light natural language analysis of the query. This approach is particularly prevalent in video search systems. Another strategy is to statistically measure the results returned by each of the available search tools and predict the relative quality of each of the available tools [27], [34]. In the final fused combination of all the available tools, each tool can be weighted according to its estimated quality of performance. Yet more new strategies are proposing to take measurements of the user's context (such as location, search history, or task status) and make further adaptations based on these information sources [8], [9], [47].

In this paper, we provide a review of work on adaptive search strategies in a broad range of applications. We conduct this review with a special emphasis on multimodal video search applications. We keep an eye towards recent developments outside of the domain of video retrieval in order to gain insights into important lessons that can be learned and applied in deciding new directions forward in multimodal video retrieval research. We will show that many current techniques in play in state-of-the-art video retrieval systems, which mostly rely on query-class-dependent mechanisms, are providing workable solutions to query-adaptation problems. However, we will also see that there are practical problems arising due to the need for large amounts of training data to develop such systems. We will propose to follow suit with other query-adaptive applications and to leverage the benefits from contextual awareness and difficulty prediction techniques by integrating these methods into future video retrieval systems.

The remaining sections of this paper are organized as follows. In Section II, we will introduce a general framework for query-adaptive search processing, which we will use to frame our discussion of a broad range of search techniques in various applications. We will provide detailed review of the literature on query-adaptive search systems in Section III. Section IV will analyze the successes and shortcomings of various common aspects of these search applications and we will offer conclusions in Section V.

II. FRAMEWORK FOR QUERY-ADAPTIVE SEARCH PROCESSING

A. Terminology

In this paper, we will examine query-adaptive search systems from a variety of domains, media types, and applications. In order to facilitate comparisons between the diverse types of research, we will adopt a standard vocabulary to describe many of the common aspects and components across each of the systems.

Modality: A modality is traditionally defined as a sense through which a human can receive some piece of information. In the multimedia retrieval literature, however, the use of the term "modality" has been expanded to mean any source of information about the contents of a searchable database that can be leveraged algorithmically for retrieval. So in image and video search applications, the traditional "vision" modality can be decomposed into various modalities which measure various low-level aspects of the visual data (such as the color distributions, the textures, and the directions of the edges) as well as some mid-level visual concept categorization (such as people, location, and objects). Likewise, the traditional "text" modality can be obtained by the textual information associated with the content or textual speech recognition transcripts. The "audition" modality may be processed by measurements of speaker pitch and volume or audio features extracted from the music signals. Even inter-document relationships, like the links between pages on the web, can be interpreted as modalities to be incorporated in processing and searching a database.

Document: A unit within a database to be searched and retrieved. Examples include a page on the web, a shot within a video corpus, or an image in a personal photo collection. In many cases, documents can encompass information from many different modalities such as speech recognition transcripts and visual concept detection results in a video shot or the body text of a web page and the anchor text of referring pages on the web.

Query: An input to the search system provided by the user, which states the user's information need. In web

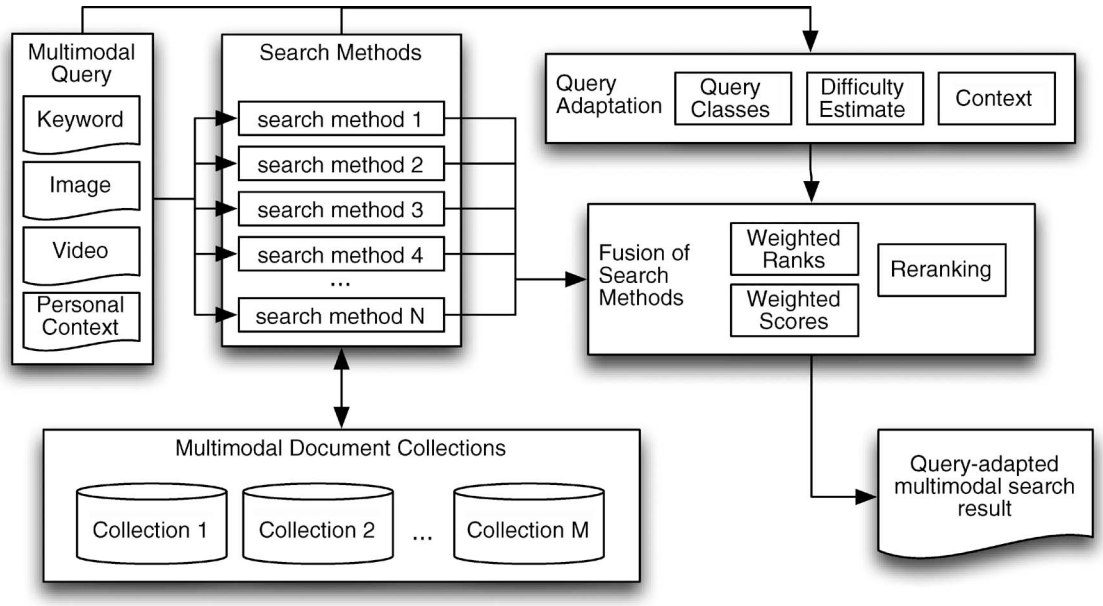

Fig. 1. ***General framework for query-adaptive multimodal fusion.***

search systems, a query is typically a few text keywords, while in video retrieval systems, queries can include example images and video clips in addition to text keywords. Queries are typically very sparse, including just a few keywords or example images.

Context: Information about the user's current state or personal history and preferences that can be relevant to clarifying his or her query and improving retrieval results. In web or video browsing applications, the current document that the user is viewing can be an important piece of context. In mobile applications, the physical location of the user can be relevant.

Search method: An algorithm used to match users' queries and contexts against the database to rank documents according to their relevance. Many search methods only focus on pieces of queries or specific modalities of documents. For example, in many video search systems, the example images and videos given by the searcher are matched against the visual contents of the database by a single search method, while a separate search method handles the query text keywords and matches them against the speech recognition transcript. A fusion method is then needed to combine the results from various applicable search methods to give a multimodal result.

Multimodal fusion: A combination of the results from various search methods and modalities to give a final result. In video search systems, the scores resulting from unimodal search methods are typically weighted and summed.

Meta-search and federation: Meta-search uses several different search engines to index and search over collections of documents, where the collections may be slightly or entirely overlapping. The results from the multiple search engines are then combined and returned to the user. Federation uses the same search engine to search over separate, nonoverlapping collections of documents and then combines the results and returns them to the user. In the literature, the term, meta-search, has often encompassed both meta-search (as described here) and federation; however, in this paper, we will use this separate terminology to distinguish between the two.

Query class: A set of queries where the optimal multimodal fusion or meta-search strategy is similar. In broadcast news video, searches for shots of named persons are best handled by giving more weight to the text search method, while a search for a sports scene can give more weight to an example-based image search method, so these types of queries would require their own classes.

B. General Framework

In Fig. 1, we show a general framework for query-adaptive retrieval, where the various components are general across many applications, but their inclusion or exclusion and exact design can vary.

At the core of these retrieval systems is a suite of individual search methods which can be built to search over various aspects of the data index in many various configurations. Such search methods may be separate full search engines in meta-search or single-modality search methods for video retrieval.

Where there are search methods, there is obviously a need to allow for the input of queries. Queries are different across many applications and are typically limited to a few textual keywords; however, they can also encompass full

natural language sentences, example images, sounds, and/or video clips, and even contextual cues about the user's state while issuing the query.

The principal motivation for introducing a query-adaptation mechanism into the search system is the fact that each of these component search methods has varying search performances in response to different query inputs, which necessitates the need for two components. First, there is a component needed to fuse the information provided by the various search methods and cues. This is widely done using a simple weighted linear fusion of the results. And second, there is a component needed to analyze the incoming query and predict the appropriate search method to associate with it, which is typically done by predefining a set of query classes and classifying the query with some light language processing. Newer methods might look at the statistics of the results returned from various search methods to infer directly which search methods to prefer and which to ignore.

Finally, before any of these components can be activated and the system can be applied, there is a need for a collection of previously seen queries, along with the relevance labels that are associated to documents in a training collection. This is used to anticipate the types of queries that will be encountered by the system and to train the methods for classifying or adapting to incoming queries and optimizing the combinations of search methods. Note the training collection ideally is different from the training data sets that are used for developing individual search methods.

We stress that this proposed unified framework for query-adaptive retrieval need not be rigidly applied across all applications and domains. It should perhaps go without saying that the various applications that we fold into this framework are quite different in many aspects. For example meta-search can have different and overlapping indexes and different indexing and ranking schemes, while multimodal search uses different search techniques to address various informational aspects of a single source. As a result, the specific considerations for any such system would vary according to the task at hand. Our intention here, though, is to use this framework as a construct for guiding our discussion of diverse sets of applications by emphasizing the similarities between these problems, rather than the differences. We use the lessons learned through this exercise to expose opportunities for techniques and algorithms developed in one application (such as meta-search) to be applied to problems in another application (like multimodal search).

III. APPLICATIONS OF QUERY-ADAPTIVE SEARCH

In this section, we will summarize and compare a number of query-adaptive search systems in a broad range of domains and applications. We will first describe applications in the text, web, and video search domains and discover a broad trend of incorporating predefined query classes as a strategy for adapting search mechanisms for varying queries. We will then explore some more alternative approaches to query adaptation, such as query difficulty prediction and contextual awareness, which extend beyond query classes and can infer adaptive strategies more directly.

A. Search Over Text Collections and the Web

In retrieval from text collections or the web, we can observe the existence of diverse information sources, such as disjointed subcollections of text documents or linking and structural cues in web documents, which will necessitate adaptive strategies for utilizing the available information sources for varying queries. We will examine two case studies: one in which the retrieval strategy from text documents is dependent upon the inferred "topic" of the query and another in which the search method for the web is dependent upon the inferred intention or "task" of the user.

Adapting by Query Topic: Since the beginning of research in information retrieval over text documents, systems have been dominated primarily by the notion that the approximate semantic content of a document can be encapsulated by the counts of the words that appear in the document and that relevant documents can be discovered by a comparison of the query terms provided by the user and the relative frequency of those terms in the document [20], [23], [24], [26]. Despite the convergence across many text-based information retrieval systems towards the same fundamental principles for retrieval, it was observed that the results from various systems were still unique and somewhat complementary. Indeed, if the results were different, but still comparable in performance, then it would be possible to perform a meta-search and combine the results from different engines and come up with an even finer-grained result to return to the user [45], [46].

In [11], Voorhees *et al.* look at adapting retrieval strategies in response to varying queries in a federated retrieval scenario, where the set of documents to be searched is stored in separate indexes, which are searched separately and need to have their results fused before returning to the user. To simulate such a situation, the authors take a natural subdivision of the documents used in the early TREC evaluations and split them into five different collections according to the five different sources that contributed the documents. Each collection is indexed independently with the same search system and, given a natural language text query, returns a ranked set of documents. A naïve approach to such an application might be to assume that relevant documents are uniformly distributed across all of the collections; however, this is typically the exception, rather than the rule. Intuitively, it is quite likely that the individual collection will have their

own topical idiosyncrasies; one collection might be mostly focused on medical journal articles, while another might contain news stories, or technical help articles, so the reality is that relevant documents are highly likely to be skewed towards a few of the collections, depending on the content of the query. The authors propose to improve upon the naïve assumption of the collections being equally relevant, by learning the relevance of the collections to various types of queries, using a collection of training queries, which have relevance labels assigned.

There are two approaches to predicting the fusion strategy from the varying collections that are tested. In the first approach, for each incoming query, the nearest training queries are found according to the cosine similarity between the vector space representations of each of the queries (or the counts of various words in each query). The average number of relevant documents from each of the search collections is then calculated over the set of found related training queries. The query can then be run over each of the collections to get ranked lists for each collection. The ranked lists are fused using a round-robin selection process, where the merged list is calculated by iteratively selecting the top-ranked documents from each collection's ranked list, where the next collection to choose a document from is determined probabilistically with the probability of selecting a collection being proportional to the average number of relevant documents coming from the collection in the training queries. In the second approach, the training queries are first clustered into a set of classes. The clustering is conducted by constructing a similarity score between training queries by calculating the size of the overlap of the sets of top-ranked documents returned by each query. The intuition is that two queries that return many of the same documents in highly ranked positions will be topically similar. Then, each cluster of queries is represented by the centroid of the queries in term vector space. An incoming query is then mapped to a single cluster of training queries via its cosine similarity with the centroids of the clusters. From there, the average number of relevant documents coming from each of the training queries in the cluster can again be used to determine fusion weights for resulting ranked lists from each of the subcollections.

Adapting by Query Task: In the early stages of the world wide web, many web search engines came about through the direct application of the simple keywords-based retrieval models from the information retrieval research community over the newly created web document set. The search results of such engines were often of mixed quality, since the importance and relevance of many documents was not conveyed by the content, alone. Furthermore, the retrieval mechanisms were easy to understand, and therefore they were easily subverted by advertisers and other participants who wished to drive traffic to irrelevant sites by simply adding popular keywords to the pages. The web search landscape changed significantly, however, with the introduction of the PageRank algorithm [44], which took into consideration the fact that the web has a unique structure of linkages between the various documents that it contains. Indeed, it is rather intuitive that sites on the web that have a great deal of incoming links are in some sense "authoritative." The existence of a link to a particular site indicates that the site was determined to be of some importance by some human author on the web. This act of annotating the web is of considerable utility in web search, when compared to simply mining the terms contained in the pages, and, therefore, sites with many inbound links can be thought of as being of considerable importance and are, therefore, *a priori* more likely to be relevant to queries than other sites, which have few or no incoming links. Indeed, the power of the link-structure on the web has come to such widespread understanding that subversive entities on the web have devised methods to build out so-called "link farms," which are machine-generated sites aimed at synthetically building up the number of incoming links to particular sites, in order to game the importance assessments gathered by search engines. This has come to the point where search engines must actively detect and remove such systems from their indexing schemes.

In the years that have followed since the introduction of PageRank, many other important signals or cues from the structure of the web have been incorporated into nearly all major search engines, including the use of the text inside anchor tags from linking sites to propagate some annotation to the linked-to site, the consideration of content held in title or heading tags, and the structure, content, and depth of the URL for a given page.

Of course, the relative importance of each of these cues can vary depending upon the objectives of a given query. In [5], Kang and Kim propose to decompose the various types of web queries into three specific types: topic relevance, homepage finding, and service finding. The topic relevance task is a search conducted in the spirit of the typical information retrieval evaluation framework, where the intent of the query is to uncover some piece of information that is previously unknown to the searcher. Documents in this case are to be ranked in decreasing likelihood of fulfilling the needs of the searcher. The homepage finding task, on the other hand, is a known-item finding task, where the intention of the query is to find a specific page that is already known to the searcher, but the exact URL is unknown. This may be a query such as "find the IBM homepage." The goal here is to find the exact homepage for IBM. This mode of interaction is in tune with the often-observed behavior of users using web search engines instead of maintaining bookmarks. The service finding task is more directed at finding web documents which would enable a transaction desired by the user, such as "buy plane tickets."

The authors conduct experiments over a standard 10-GB collection of web documents. The collection is

indexed and searched using three primary information sources: 1) the content information, which is the traditional term-frequency indexing of documents, but here can also include text from the title of the document or from the anchor text of linking documents; 2) the link information, which includes the PageRank scoring of each document's authoritativeness; and 3) the URL information, which accounts for the depth of the URL, where shorter URLs, with fewer slashes and less directory structure appended to the end of the URL, indicate a greater likelihood of a homepage and the terms used in the URL structure can be indexed and matched against query keywords. The authors test the system for search over the collection using 50 topic relevance queries and 145 homepage-finding queries and find that link information is significantly more powerful for the homepage-finding application than it is for the topic relevance task (indeed, it can be harmful for the topic relevance task). The authors also decompose the experiments further to test the relative efficacy of the use of the full textual content of the page versus the use of just the title and the anchor text snippets. They find that the anchor text and title are succinct enough that they are very powerful for the homepage finding task, while they are too sparse to adequately encapsulate the depth of a document to be useful in the topic relevance task. These results are quite sensible, since homepages often tend to be the highly linked entry points for many major sites. Also, the titles and anchor texts provide useful and succinct annotations of exactly which homepage is to be found at the URL. On the other hand, topic relevance tasks require much greater depth of information and the relevant documents are more likely to be buried deeper within a site.

Having observed these divergences in the applicability of various document modalities and search methods in two distinct types of queries, the authors also set out to devise a scheme for automatically determining the intention of the user (i.e., whether the query is for topic relevance or homepage finding) to then be able to apply the best strategy for the query. The authors note that typical queries into web search engines are very short and are not grammatically constructed sentences, but rather strings of keywords, so the interpretation of the searcher's intentions can be quite difficult. To address the problem, the authors actually propose to begin with the document collection. They heuristically divide the collection of documents in the index into two types: homepages and topic pages. The homepages are chosen to be simply the pages that are at the root level of a site, where the URL has no directory structure beyond the domain name. They further add in virtual pages for all texts contained in hyperlinks within the collection, since linked-to pages are also possibly homepages and homepages are typically more textually sparse than topic relevance pages, so the added hyperlink text will act to significantly augment the content information for homepages. The remaining documents are contained in the topic page set. They then construct language models for each of the collections, by counting the likelihood of randomly sampling each of the terms from a given collection. The result is that many named entities and businesses that are associated with homepages are more frequent in the homepage collection than in the topic collection. The queries are classified as either homepage or topic-relevance based on which language model they fit best. This is interesting, since the characteristics of the queries are not learned from logs of queries, but rather inferred from the types of documents that are expected to be relevant to the query. The authors also incorporate statistics on interrelationships between multiple terms in each collection, so when a multiterm query is issued, if the terms are more likely to co-occur in one collection than the other, then it can be inferred that the preferred collection is matched to the query type. Other features used for predicting the query type are the anchor text usage rate and part-of-speech information. For anchor text usage rate, the frequency of the occurrence of query terms as parts of anchor text segments is measured, and the heuristic assumption is that terms occurring in anchor text segments are highly likely to be related to the homepage-finding task. For the part-of-speech information, the occurrence of verbs (other than forms of "be"), is heuristically taken to be an indicator of a topic relevance task. Each of the query-type predictors are then combined through a weighted summation to arrive at a final classification decision, and the appropriate search mechanism is applied. The results indicate that this approach is highly successful and classifying queries and that proper distinction between query intentions leads to significant gains in search performance.

Discussion: A promising strategy for adapting search methods can rely upon predefined hard classes or types of queries, which can be defined using any given criteria, such as the inferred topic of the query or the intended task of the user. We have also observed that classification of incoming queries can be done using classification schemes learned from either the queries seen in training data, or the actual content of documents expected to be associated with the given class. In the following section, we will see that the query classification is also useful in retrieval over video and image collections.

B. Multimodal Search Over Image and Video Collections

Videos and images can encompass a rich set of information sources and modalities that can all be exploited to enhance indexing and retrieval. In this section, we will review several applications of query-adaptive retrieval systems for image and video collections and see that virtually all of these systems rely on predefined classes of queries, such as those introduced in the previous section. The primary distinguishing factor

between these various systems is the way in which each system defines these query classes, either through human-provided intuition or through the mining of query logs.

Search over image and video collections offers many challenges, but also some opportunities, when compared to text and web search applications. The most glaring challenge in search over visual collections is the often-mentioned "semantic gap," or the disconnect between the information that is actually contained in the document (i.e., just the pixel intensities) and the higher level semantic meaning that a human observer, or search system user, would like to attach to the document. Of course, this semantic gap is somewhat present in text search as well. A bag-of-words representation of a text document does not fully capture all the subtleties of the semantics of the document; however, the size of this words-to-semantics gap is considerably smaller than the pixels-to-semantics gap in image search. On the other hand, visual collections also present many opportunities for applications and methods that go beyond the feasible approaches of text-based collections. Many natural collections of visual data contain a richness of information from a variety of different sources that open up many possibilities for advanced retrieval techniques. For example, searching over images gathered from the web opens up the opportunity to associate text with the images via the filenames of the images or the text around the image on the webpage. Broadcast news video collections can be reasonably searched using text-based methods over the closed captioning or speech recognition transcripts, but the video stream is also a rich source of information which can be leveraged for a variety of enhancements, such as extracting the story structure of the news broadcast or linking related stories via detecting duplicated visual scenes across diverse news sources. So, while extracting high-level semantics directly from visual data remains a challenging area of research, the use of visual information to augment rough semantics gathered from related text and other sources is a robust and promising application area.

Multimedia Search Components and Meta-Search: Many of the earliest image and video retrieval systems recognized the diversity of possible information sources available in visual databases and incorporated a variety of tools for enabling search over visual content [12], [18], [19], [35], [36]. One such tool is the query-by-example search method, whereby users could provide external examples, or use images from within the database to find other images that were similar in various low-level features, such as color distribution or texture. Other systems included query-by-sketch search methods, where users could draw an approximation of the images that they are seeking and find images with similar spatially distributed color schemes. Other systems, still, made use of any textual information that was available to be associated with images (such as filenames and webpage body text) and enabled search via text keywords. A commonality across many of these systems, though, is the seeming recognition on the part of the designers that no single combination of all the available search methods would be appropriate for every query seen by the system. So, despite the ability of many of the search engines to fuse across several different search methods, the selection of tools to use and the degree of fusion across those tools was often left to the users, who would be expected to gain some level of expertise in using the search system.

Benitez *et al.* [13] proposed a meta-search application for image and video retrieval, which disseminated incoming queries to many of the openly available visual search engines of the time and combined the results to be returned to the user. Fig. 2 shows the system architecture of the meta-search system. A baseline approach to such a task might trust the results from each engine equally and afford them equal weight when combining the results to be returned to the user. However, the authors proposed to track the quality of the results from previous queries to the system to trace the relative performance of each engine for each type of query. This tracking of quality is done via a relevance feedback mechanism, wherein users of the system are able to mark returned images as either "relevant" or "irrelevant" to the query that they have issued. So, when a query image is submitted to various search engines and aggregated and displayed back to the searcher, the feedback from the user can be used to evaluate the relative strength of the results from each of the underlying search methods. After a large collection of queries and relevance feedback inputs have been solicited from users, the query images are then clustered based on their color and texture features. When a new query image is issued, it is compared against the various clusters of previously issued query images and the nearest cluster is selected as the class of the query. From there, the system can evaluate which of the underlying visual search systems offered the most relevant images in response to queries belonging to that cluster in the past. The system can then choose to only issue queries to the engines most likely to give relevant responses or to afford higher weightings to the results from those engines when fusing the returned images from the search engines. The clusters of query images, then, are effectively *classes* of queries, where the search strategy for a new incoming query mapped into the class is determined by the performance of the component search engines in the past. Interestingly, the classes of queries are dynamic and the associated search strategies are also variable, as the distribution of queries into classes is updated with new clustering results as more and more queries are processed by the system. The evaluation shows that selecting search engines based on clustering of previous queries and the relevance labels given by searchers significantly improves the precision of results returned to the users during later queries, when compared to a baseline flat fusion of all the search engines or random selection of engines.

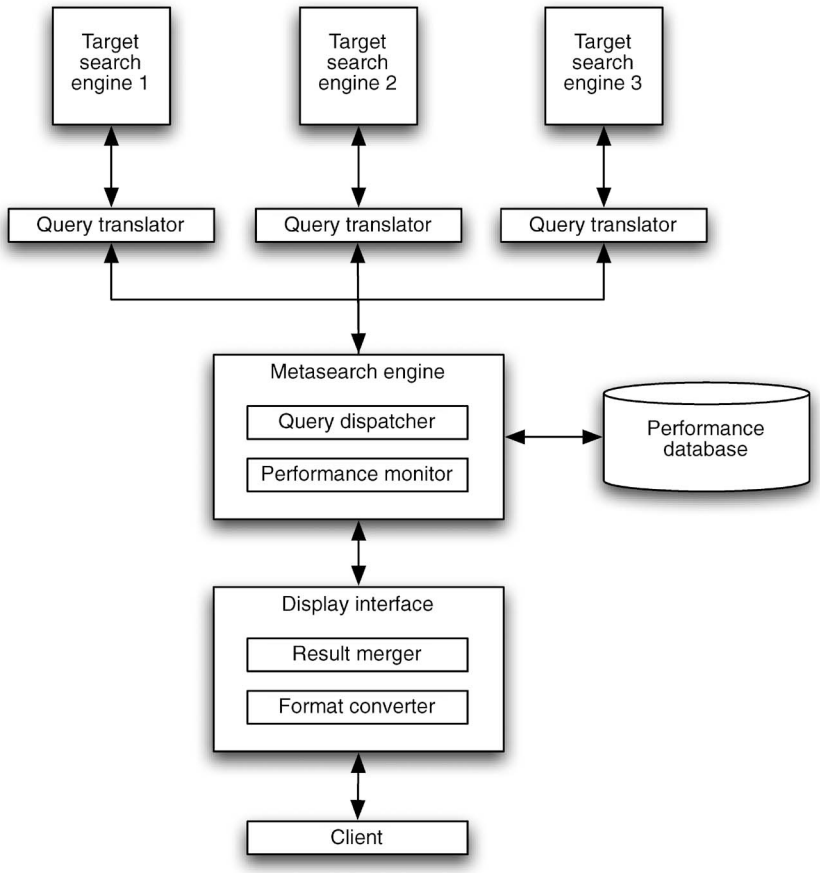

Fig. 2. ***System architecture of image meta-search engine [13].***

TRECVID Search: Components and Fusion: In recent years, much of the research in image and video retrieval has been driven in large part by the NIST TRECVID video retrieval evaluations [28], which are open benchmarks, wherein research teams can establish comparisons between a variety of image and video search techniques via evaluations on commonly shared data on identical tasks. In the years of 2003 to 2006, the TRECVID benchmark data was largely focused on broadcast news videos. In each year, NIST would provide a common set of broadcast news videos, typically on the order of 100 hours in size. The videos were accompanied with a common segmentation of the videos into shot units along with speech recognition transcripts [25] and machine translation of the speech recognition into English in the case of foreign news sources. One of the key evaluation activities has been for the "search" task, wherein NIST provides a set of 24 standard query topics that each team participating in the evaluation must submit to their system for processing. The results from all participating teams are then evaluated and a pooled set of ground truth labels is determined by NIST. The query topics provided typically consist of a natural language textual description of the required results (like "Find shots of Condoleezza Rice," or "Find shots of helicopters in flight") along with five to ten example query images or video clips. The task of the search system, then, is to find shots from within the search data that match the needs expressed by the query.

Search across TRECVID collections is typically done using a few core tools: *text search*, allowing keyword-based queries against the speech recognition transcripts of the videos; *image matching*, allowing the user to provide example images and find other images in the search set with similar low-level features; and, in more recent systems, *concept-based search*, using a set of pretrained visual concept detectors which can be selected and fused based on text keyword queries. These methods are representative of the core tools used in state of the art systems, but, in principle, any set of tools can be used. In most successful systems, each tool is applied independently and the results are combined through a weighted summation of either scores or ranks, resulting in a fused multimodal ranking of documents in the search set. The problem here is how to choose the weight for each of the individual search components. In the earlier development of TRECVID search systems, the weighting assigned to

each of the available search methods was selected over some validation set to maximize the average performance of the system over a broad range of queries. This is a *query-independent* strategy, which we have established can be suboptimal for many queries.

Query-Class-Dependent Models: To improve upon these query-independent systems, Chua *et al.* [7] and Yan *et al.* [17] independently proposed *query-class-dependent* approaches, which relied upon human-defined classes of queries. Chua *et al.* defined six classes of queries: "Person," "Sports," "Finance," "Weather," "Disaster," and "General," while Yan *et al.* chose five classes: "Named Persons," "Named Object," "General Object," "Scene," and "Sports." The general frameworks of both proposed systems follow similar strategies, beyond the selection of the query classes. Optimal fusion strategies are determined for each class (either by learning weights over a training set or by hand-tuning weights). Also, importantly, both systems employ a strategy for mapping unseen test queries into query classes via some light natural language processing (since the queries in each class tend to be linguistically similar). The findings in both systems show significant improvements in using these query-class-dependent strategies instead of query-independent ones.

In examining these two query-class-dependent applications, we can see that it is unclear how, exactly, system developers should identify and define classes of queries. In these earlier works, the classes were defined by humans sifting through a set of typical user queries and intuitively identifying patterns. Query classes discovered in this manner, however, are prone to problems induced by human factors and it is difficult to know whether or not a proposed query-class scheme is optimal or not. Indeed, we can see that the two independently proposed classes of queries discussed above have very little in common: there are only two classes, "Named Person" and "Sports," shared between the two sets of classes. Further analysis shows that memberships in the various classes are unbalanced or not representative of the types of queries typically used. For example, the "Finance," "Weather," and "Disaster" classes contain very few queries, while the "General" class encompasses virtually all of the queries. It is therefore necessary to provide a principled framework for automatically discovering classes of queries.

In our prior paper [14], we undertook the task of designing a method for determining useful groupings of queries by mining previous queries and their associated relevance judgments. We took the stance that, intuitively, the best way to fuse various unimodal search methods is to give more weight to methods which perform well for a query and give less weight to the methods that perform poorly; therefore, it is reasonable to suggest that well-formed query classes should be constrained to contain queries with similar performances across each of the available search methods. To utilize this constraint, we proposed a data-driven method to automatically discover query classes, using a set of example queries with labeled ground truth relevance. Fig. 3 shows the overall system architecture. We can automatically discover query classes by clustering in a "performance space," which is defined by the performance of the query in various unimodal search methods. The query classes discovered by this approach are virtually guaranteed to have more consistent optimal fusion weights within each class; however, attempting to use these classes in a real search scenario will be problematic since queries coming into a search system will typically have unknown performance across the search tools, so it will be impossible to reliably select the best class for an unseen incoming query. To address this problem, we propose to also represent queries in a "semantic space," which captures the intent of the user by measuring the semantic content of the query through methods such as counting named entities, parts of speech, and measuring lexical relationships between keywords across various queries. The most important aspect of this semantic space is the fact that it is measurable at query time, so through clustering training queries in a space composed of both performance and semantic measurements, we can arrive at query classes which have consistent performance across various search methods *and* can be used reliably in real query classification applications. After the classes of queries are determined by the clustering process, the optimal weighting of search methods for each class is determined and a mapping from the semantic space of an incoming query to the clusters is also learned, much in the same way that these weights and mappings are learned in the previously-proposed hand-defined query-class-dependent models. Through evaluation, we see that this automated approach to discovering and defining query classes provides search results that consistently and significantly outperform the search results of hand-defined classes.

In a similar work, Yan and Hauptmann [3] also propose a framework for automatically determining class-dependent search models. The framework proposes a "probabilistic latent query analysis" (pLQA) model, wherein the query classes are found as a latent variable under the assumptions that queries within the same class share the same optimal weighting of individual search methods and that the plain-text description of the query has enough information to classify the query into a class. The proposed approach offers many benefits beyond the framework proposed in our prior work. Whereas our prior work uses the "performance space" as a proxy for estimating the optimal combination weights for a given query during class discovery, the pLQA model uses the combination weights directly during class discovery. And while our prior model discovers query classes first and then fusion models second in a two-stage process, the pLQA model optimizes class membership and fusion models in a single joint process. Further advantages come

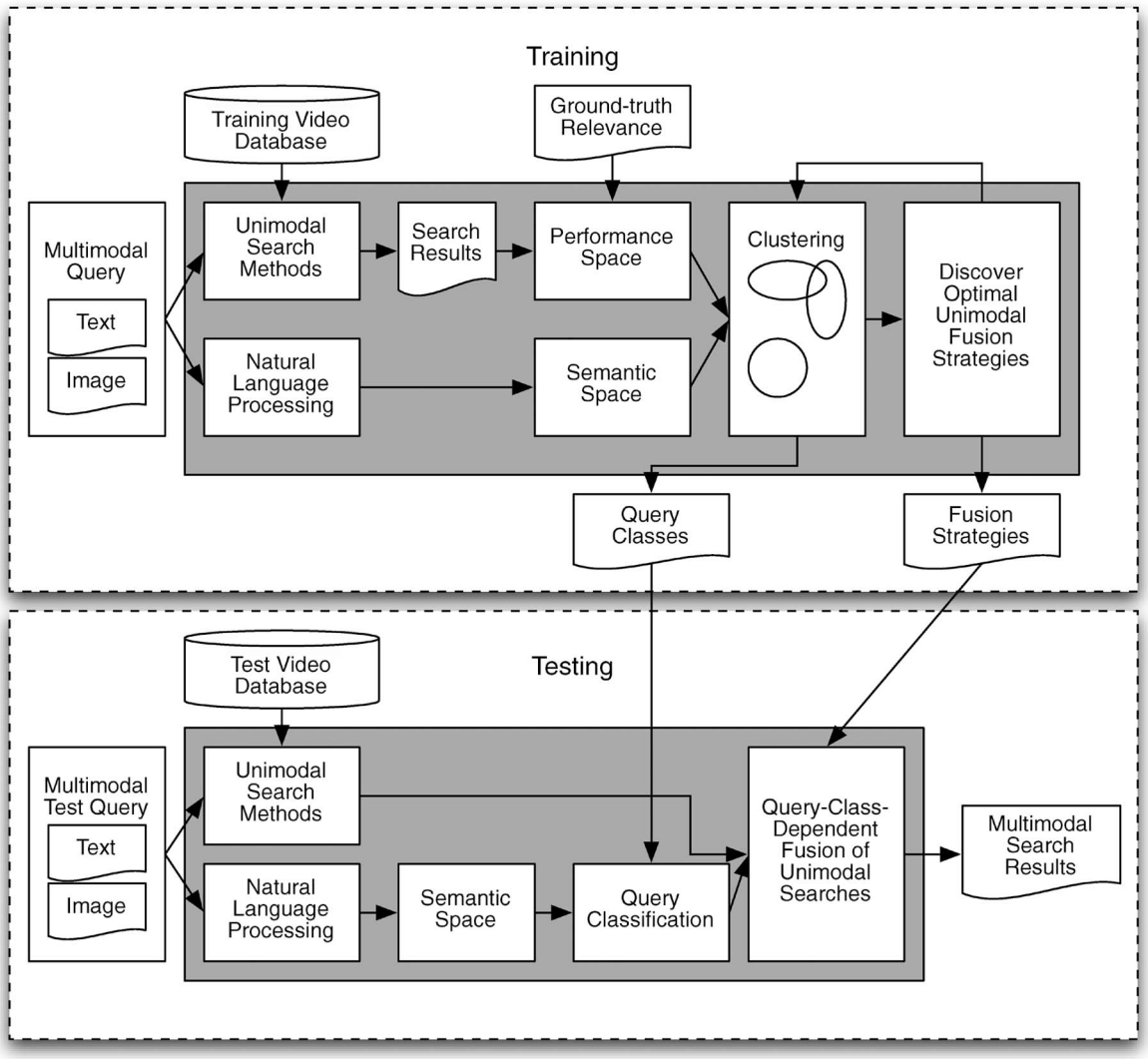

Fig. 3. ***Framework for automatic discovery of query classes by joint performance/semantic space clustering.***

from the inclusion of a principled approach for determining the number of query classes and the ability to map a single query into a mixture of several different query classes. Evaluation of the pLQA model on a broad range of TRECVID queries shows significant performance gains over both hand-defined query classes, which are greater in magnitude than the improvements shown in our prior work.

In [59], Xie *et al.* propose a system that dynamically creates query classes as incoming queries are received (as opposed to the cases above, where query classes are learned once from the training data and kept static across all queries). To achieve this, the system measures the semantic distance between the incoming query and the pool of training queries. This semantic distance is measured with various textual query features, such as named entities and concepts, as provided by a question-answering system. The top k closest queries from the training pool are then pulled together as a dynamic query class. An adaptive fusion model is then learned on the fly, such that it optimizes average retrieval performance over this subset of training queries, and the model is applied to the query. In the end, the system only shows slight improvements over static query class methods similar to the above-described approaches; however, this dynamic query class approach is unique and shows promise. One possible shortcoming is its reliance entirely on semantic similarity between queries (and not the performance-based similarity discussed above). Further extending the work to incorporate such performance cues may lead to greater performance increases.

Query-Classes for Web Video Collections: Zhang *et al.* [6] describe an application of query-class-dependency for

video search in a domain entirely different from TRECVID: large-scale web-based video clip search. Web video clips are almost entirely different in content than the broadcast news clips in the TRECVID datasets, and the types of users (typical consumers for the web and news analysts or producers for TRECVID) are also quite different. So, in this case, the types of useful classes and queries are also different. In the proposed system, they predefine a set of five types of videos: "news," "finance," "movie," "music," and "funny" and classify each video in the search set as belonging to one or more of these classes based on associated textual information from the web pages that the clips appear on, meta-data included with the clip, and visual color and texture content features of keyframes from the clips. Incoming queries, which are simply strings of text keywords, are then also classified into one or more of these classes, based on the frequency of the search terms within documents of the various classes as well as the user's typical interest in each of the classes, given his or her history of search types and clicked results. The results returned from the search are then constrained to only the clips that occur inside the categories that correspond to the query. An evaluation of the precision of the system using this joint-categorization of video clips and queries shows significant improvement when compared to the case with no categorization at all.

Discussion: Clearly, there is a great deal of improvement that can be attained from applying query-class-dependent models for multimodal video search; however, there are still many remaining challenges. In particular, there is still a very much open issue of how, exactly, to arrive at an appropriate set of query classes for a given application. The intuitions of the system designers can be out-of-tune with the realities of the desires of the users. And the acquisition of training data for tuning search methods for various types of queries can be very expensive. For example, it is highly unlikely that the range of query classes is really on the order of five or ten. In reality, this small number of classes is more likely an artifact of the sparse training data available to discover the rich patterns in video search behavior. One direction forward may be to eschew the practice of using query classes in favor of directly predicting the search method without any classes at all, which we will discuss in more depth in the following section.

C. Difficulty Prediction for Adaptive Meta-Search in Text Collections

One of the key issues with the query-adaptive approaches to search that we have discussed so far is that the mechanisms for assigning incoming queries into one of the query classes are dependent exclusively on the queries input from the users. In experimental and benchmark tests, the entered queries can be several complete sentences and many example images; however, in many real web and video search systems, it is repeatedly found that typical users are willing to only enter a few keywords, on average about three or four, per query. Such input makes much of the natural language processing and named entity extraction necessary for many of the query-classification approaches next to impossible. In the text retrieval community, there has been some recent focus on the prediction of the difficulty or performance of a search engine on a given query, which is not dependent upon the lexical or semantic properties of the individual keywords in the query, but rather, on the statistics of the documents that are returned by the search engine on the full query and several atomic subqueries extracted from the full query [27]. Such methods have the promise to be robust against the problems incurred by trying to extract semantic intentions and meaning from the sparse input from search users. Indeed, such extractions of query semantics are in many aspects simply a proxy for mapping from query input to the expected performance of various search components, which can in turn be mapped into an appropriate combined search strategy. A direct prediction of search performance based on the statistics of the search results can circumvent some of the problems that might arise.

Difficulty Prediction Framework: In [27], Yom-Tov *et al.* propose a method for predicting the performance of text search engines directly from the results that are returned from the engines. Given an incoming query, a series of metrics is derived by issuing the query and various subqueries to the search engine. An overview of the system architecture is shown in Fig. 4. The main mechanism for measuring performance is to compare the overlap in returned documents between the results returned by the query and subqueries. The subqueries are determined by simply taking each individual keyword in the entered query and issuing it to the search engine as an atomic query. The top documents returned by each atomic query can then be compared to the documents returned by the full query, which includes all keywords combined together. The size of the intersection of the top ten documents returned from each method is then taken as a metric of the amount of agreement between the two sets. Another proven metric for predicting query difficulty is the document frequency of the keywords provided in the query, which is the number of documents where the terms appear. Query terms with high document frequency can be indicative of imprecise or ambiguous queries, which can be difficult to satisfy. For each of the query terms, the document frequency is calculated and the overlap of the top-ten returned documents from the atomic query with the results from the full query is also calculated. The overlap and document frequency histograms are derived for a set of training queries, which all have ground-truth relevance labels available, and a decision tree is learned to predict the relative performance of the search system for each query. Given new, unseen queries, the same overlap

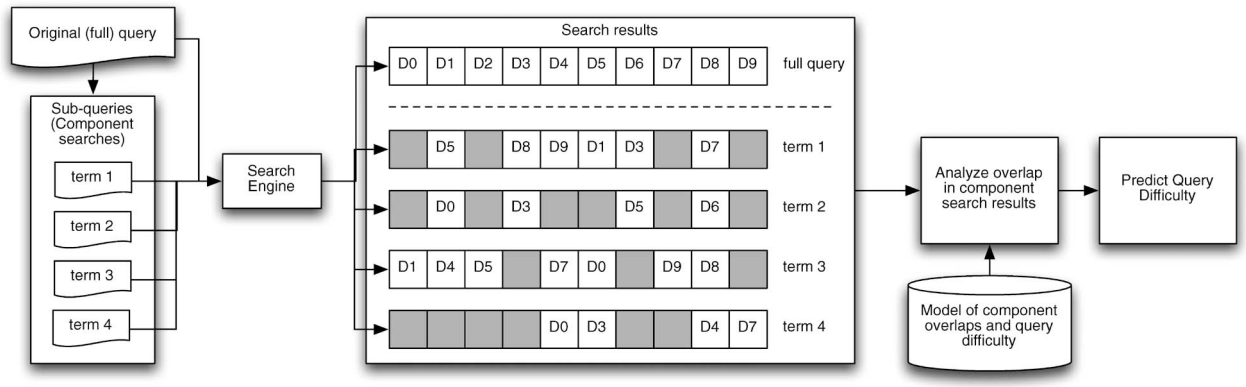

Fig. 4. ***Framework for predicting query difficulty in text search queries [27].***

and document frequency histogram can be derived and the performance of the system can be predicted. The authors find that this difficulty prediction framework can predict a ranking of query performance that is significantly correlated with the actual query performance measured in terms of precision at 10 (or the number of documents in the returned top ten that are actually relevant).

Adapting With Difficulty: Beyond simply predicting whether the engines work, the predictions can also be used to adapt and enhance the engine performance in a number of ways. For example, the authors show that automatic query expansion (where frequent terms in the top-returned documents are added to the query) is more likely to be successful for easier queries, so query difficulty prediction can lead to higher search performance through selective application of query expansion. Likewise, given a set of two possible queries (like the relatively short "title" and the longer "body" in TREC query topics), the system can selectively apply the most appropriate query given the estimated difficulty of each. And finally, the system can tune internal parameters in the search system based on the difficulty prediction. For example, the search engine used in the authors' experiments incorporates both keywords and lexical affinities for document ranking; however, regular keywords have a weighting that is fixed to be higher than lexical affinities. Further examination, though, shows that lexical affinities are more useful for difficult queries (which makes sense, since the given keywords are not performing satisfactorily) and so these lexical relationships should be afforded more weight for difficult queries and even less weight for easier ones. Actually, we can treat this approach as another case of query type classification. The approaches reviewed earlier group queries based on predefined semantic classes or unsupervised clustering. Here, the grouping is based on predicted difficulty.

Difficulty Prediction for Meta-Search and Federation: In [34], Yom-Tov *et al.* apply their query difficulty prediction framework to meta-search and federated search applications. The principle of the application is that various search engines with indexes over the same or different collections will have varying degrees of success with retrieving relevant documents for the same query. To merge the results from these various engines and return them to the user, the system should be capable of predicting the relative quality of each of the engines and weight the results in the merging phase accordingly. To test this application for meta-search, they take a standard set of TREC text documents and index them with various freely available desktop search tools, such as those available from Google and Yahoo. They then apply a standard set of TREC queries over the search set via each of the available desktop search engines and measure the overlap statistics, predicting the performance of each of the engines for the query. The results from each of the engines are then merged according to the predicted performance of the engine for the query, where the ranking from each engine is weighted by the predicted performance. The application for federated search is similar. In this case, however, the set of TREC documents is decomposed into various subsets, each of which is indexed by the engines independently. So, here, the performance of the engine on each subset of documents can be predicted and queries where particular subsets of documents are found to have poor performance can be weighted lowly, while high-performance sets can be weighted highly.

Discussion: Predicting the difficulty of a query (or the likelihood that a search method has performed well on the query) can be a useful tactic for query-adaptation. In particular, it helps by eliminating some of the complications in query-class-dependent retrieval that arise from attempting to adequately discover and define the right set of classes for a given application and classifying incoming queries into the right class, given only the sparse information in the query. Indeed, query difficulty prediction is a more direct solution to the problem of weighting

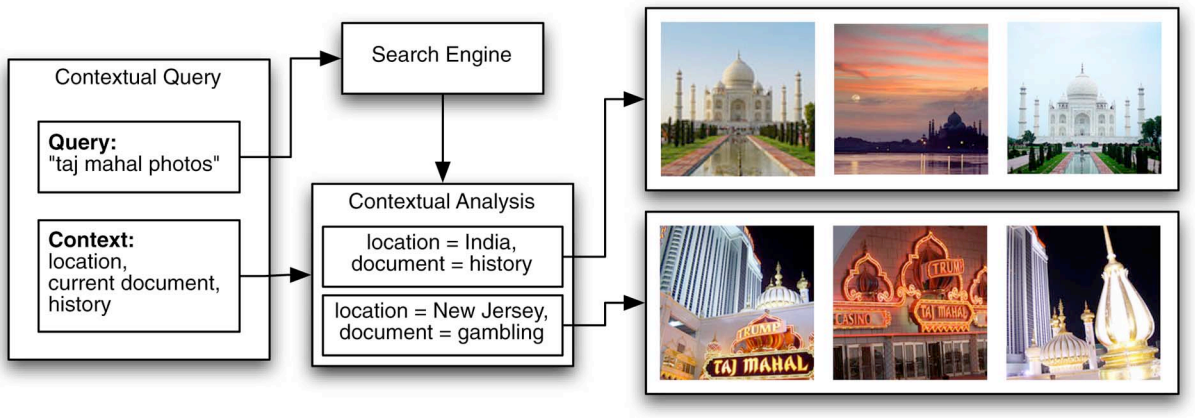

Fig. 5. ***Example of infusing photo queries with contextual awareness from location and task.***

various search methods by their effectiveness and bypasses the need to infer search intent. However, some room for investigation still remains on how exactly to apply this query difficulty prediction for query adaptation. Specifically, the proposed approaches only predict how well a given search method will perform on a certain query. In query adaptation, the real task is to select the right retrieval approach and adapted fusion strategy for each incoming query. So, there is an opportunity here to extend beyond query difficulty prediction and to specifically do retrieval model prediction or modality selection.

D. Context-Aware Query Adaptation in Text Collections and the Web

Another approach to compensating for the scarcity of information present in typical queries of only a few words is to incorporate cues about the searcher's context and leverage this information to enhance the query and refine the search results. An example application of contextual search is shown in Fig. 5. Context can come from any number of sources. In mobile applications, context can be the user's current physical location, so a search for "restaurant" could be restricted to restaurants that are geographically near the user. In web search applications, context can be the user's history of queries issued and the resulting documents that he or she has clicked, which can be a window into the user's identity and interests. So, a search for "columbia," could mean "Columbia University" for a prospective college student or "Columbia Sportswear" for an outdoors enthusiast or even "Columbia, South Carolina" for a fifth-grader researching state capitals for a class project. In this sense, any cues about the user's identity or current state can be great indicators of their search intentions and be used as predictors of search strategies. It has long been argued that such contextual cues will be at the core of the next generation of search engines [38]–[43], and many applications have made significant strides in this direction. We will review a few key contextual search applications in this section.

Current Document as Context: In [8] and [9], Kraft *et al.* formulate an approach to contextual search where the current state of the user is inferred from the document that he or she is browsing at the time of issuing the search. The intuition is that in many web browsing sessions, users conduct searches "at the point of inspiration," meaning that searches occur as part of a larger workflow, where each interaction with the search engine is a request for further information related to documents that the users are currently browsing.

The authors explore three separate mechanisms for infusing user queries with additional context, all of which can utilize context given as a weighted vector space term vector representing some snippet of text (such as a document, sentence, or paragraph) that the user is currently browsing. In the first and most basic case, called query rewriting, the user's query is augmented by extracting a few key terms from the current document context and appending those terms to the end of the entered query and then sending the resulting query to any standard search engine. This approach is found to work surprisingly well and actually mimics a behavior of users that has been observed from studies of query logs: if an initial query provides inadequate or unexpected results, users will frequently revise the query by appending a few additional terms to provide additional context [37]. The context-aware query rewriting approach does this somewhat automatically. The second case, called rank-biasing, is more complicated, since it requires having access to the internals of a search engine to be able to alter the actual mechanisms of the search. In this approach, the user's query is issued to the search engine to get the initial set of documents to be returned and then the context keywords extracted from the current document are weighted and

then used to re-order the documents from within the result set. The third approach is called iterative filtering meta-search, and in this case, the query terms provided by the user and the contextual terms extracted from the document are all combined in various manners to arrive at a set of many individual search results, which can be fused through any generic meta-search technique to give a fused, contextually aware result to the user. In this case, the key challenge is to derive a set of queries to issue, given the user's query and the context. The authors propose to do this using original query terms in each query and appending a few terms from the context term vector. Since the original query terms are included in every derived subquery, there is some assurance that the results from each query will be related and not completely disjoint, a condition which gives meta-search a better chance of succeeding. Each of the approaches, with various parameter configurations, is tested over a set of 200 real user queries and contexts. It is found that the simple query rewriting method works surprisingly well. The rank-biasing and iterative filtering meta-search methods improve significantly over the query rewriting method, though. It is also found that allowing users to manually add contextual keywords manually does not approach the automatic approaches that they have explored.

Location as Context: In the ZoneTag system [49], geographic and social context are used to assist users with labeling and sharing their personal photographs. When a user takes a photograph with their cameraphone, the current cell tower ID is used to capture the approximate geographic location of the user. Given the tagging history of other users in this location, a set of geographically relevant tags can be suggested for the user, which he or she can choose to add to the photo before posting online. The personal social network of the user, given by his or her self-defined friendships with other users, is also leveraged to give greater weight to tags used by socially connected users over other random users. Such location context can also be used to expand and/or refine mobile photo search, giving preference to search results that are geographically close to the user's current location. Indeed, in a complementary system, Zurfer [54], the cell tower ID is used to give location context for browsing and exploration of publicly shared photographs. This context can be used to explore photos that were taken in and around the location where the user is currently standing.

Generalized Context Framework: In [47], Wen *et al.* propose a general probabilistic model for contextual retrieval. The authors observe that in many contextual retrieval scenarios there is an issue of informational incompatibility between the context available and the types of queries that can be issued, while much of the prior literature had been focused on cases where there is, in fact, a straightforward compatibility between the context and the queries. For example, in Kraft *et al.*, which we described above, the context is simply a string of keywords and is therefore compatible with query system: the contextual cues can be simply appended to or fused with regular queries. However, in many applications, such as a geographically aware search, the contextual information is in an incompatible format; the context is latitude–longitude coordinates, but the query interface only allows for textual queries. Wen *et al.* propose to utilize logs of issued queries and their associated contexts along with the documents that users have clicked in order to build models of the correlations between contexts, queries, and documents. In the log, each query session consists of a query, a user context, and the clicked, retrieved documents. So, regardless of the specific application, the mutual information between the user's context states and the found clicked documents can be calculated over time. Likewise, the mutual information between the context states and query terms can be learned. These correlations can be used to constrain the users' queries to only relevant subsections of the search set, or to expand the users' queries to be more accurate and to reach more relevant documents.

The authors test this framework in a context-aware PC troubleshooting scenario. Here, users are searching over a database of technical help documents to help assess and correct problems with their personal computers. The queries are issued as text keywords and the context provided to the system is a set of many of the registry settings on the searcher's system. The primary assumption is that when users issue a query and see the resulting documents, the documents that they click on can be considered as being somewhat relevant to their query and context. The logs can then be mined to learn the correlations (via mutual information) between the user's registry settings and text terms occurring in queries or documents. The resulting framework can then append additional related terms to a user's query, based upon learned correlations with the given contextual state. An evaluation of the method is conducted by training over a large corpus of technical help documents and tens of thousands of real user queries and contexts and testing on a set of a few dozen held-out queries and contexts. It is found that contextual models give huge improvements over text search alone in ranking relevant documents.

Discussion: Contextual cues about the user, such as his or her identity, location, or state of mind (along with many other factors), can be powerful instruments for mitigating the problems incurred by the scarcity of information typically provided in user queries. This contextual information can be noisy, imprecise, or meaningless, however. For example, a user in the middle of a web browsing session may not be interested in documents related to the document that he or she is currently reading and may be starting off on a completely different tangent.

So, contextual cues can be detrimental (as well as beneficial) and mechanisms should be in place to predict when such cues are important and when they are irrelevant.

IV. ANALYSIS AND IMPLICATIONS

In the previous section, we have seen a variety of query-adaptive applications in many diverse media types and domains. There are many threads that run through these disparate systems, however, and many lessons for the future of adaptive multimodal video search can be drawn through these examinations.

A. Understanding Search Intent

At the core of all of the query-adaptive search systems that we have discussed is the objective of optimizing the internal mechanisms of the search algorithm in order to best serve the unique needs of each incoming query. This is a challenging proposition, since search systems rarely have a one-size-fits-all solution, so the right way to handle one query might give disastrous results for another. This sensitivity to the query is further complicated by the fact that queries themselves are incredibly low on information, typically having only a few keywords, so distinguishing one sort of query from another can be difficult. Even worse, the same query can have different meaning for different users (or even different meanings for the same user in different contexts) and the accompanying optimal search approach may vary as well.

Many of the systems that we have studied accomplish this prediction of optimal search models through the use of some proxy method. In particular, many systems propose to partition queries into a set of predefined classes where queries within each class take on the same optimal strategy. In addition to having the same search strategy, however, queries within the same class ought to have some qualities that are measurable from the query (which, again, is only a few keywords) that make selection of the correct class (and, in turn, the optimal query processing strategy) for an incoming query feasible.

In Voorhees *et al.* [11], the query classes are represented in term vector space and incoming queries are mapped into classes based on their similarity in that space. Benitez *et al.* [13] use a similar approach for query-by-image-example queries: the classes of queries are represented as centroids in image feature space, and incoming query images are matched to a previous set of queries based on the distance in feature space. Kang and Kim [5], on the other hand, build language models for the various types of documents that are relevant to the classes of queries in their system and classify incoming queries based on their likelihood of being generated by each of the language models. The trend in much of the recent work in multimodal video search has been to perform some lightweight natural language processing on the incoming query text to extract counts of various parts of speech as well as named entities along with the appearance of predefined lists of keywords related to specific topics [3], [7], [14], [17]. This proposal, which is heavily reliant on reasonable part-of-speech and named entity detection, is somewhat of an artifact of the TRECVID test domain, where the provided text queries are complete sentences, which is generally not the rule in many consumer applications. The net effect of many of these approaches is that queries within the same class will necessarily have similar semantics, which are typically interpretable by a human observer; however, it is not necessarily true that semantically similar queries should really adopt the same search strategy. Similarly, it is not necessarily true that a query will neatly fit into just one of the available query classes. Indeed, many systems are beginning to address this by allowing soft membership across various query classes, allowing the various fusion methods to be combined for a single query [3], [32].

Many of these complications necessitate a second look at the problem of query classification with a renewed understanding that the real objective is predicting the right search model for a query and that inferring search intent from the query is a stepping stone towards this goal. However, this self-imposed step creates many challenges and limitations of its own.

B. Beyond Classes

We can learn from recent developments in textual retrieval applications that inference of the semantic content of a query via natural language techniques may not be the only way to proceed. The query difficulty prediction approaches proposed by Yom-Tov *et al.* [27] are particularly interesting in this respect, since they are agnostic of the linguistic content of the query or of any approximation of the intent of the user. Instead, these methods focus directly on the statistics of terms and their respective appearances in the search set along with metrics for directly predicting the retrieval performance of the search system. The authors have already demonstrated the successful application of the system in prediction of fusion weights in meta-search and federated search applications; however, the prediction mechanism is designed solely for use with textual queries. One might imagine that prediction methods could similarly be developed and applied for multimodal video retrieval applications with the key challenge being the implementation of similar prediction mechanisms for other search methods that are applied over audio and visual modalities. For example, in query-by-image-example searches with multiple examples, the coherence of results from component single image searches with the results from a multi-image search may be a predictive measure that could be used. Perhaps similar prediction strategies exist for other common approaches in multimodal video search, such as text search over concept detection models.

Infusing queries with the context of the user similarly has been shown to aid in retrieval applications, though these have yet to see application in the multimodal video search domain, though there seems to be a great deal of promise in such applications. For example, a query for "Iraq war" can return a huge number of relevant stories in a broadcast news video database; however, the Iraq war, as a topic, is very large and complicated and is covered in the news from a range of perspectives, including stories about on-the-ground battles in Iraq, the policies being debated by national politicians, or even human interest stories about the soldiers serving in the war and their families at home. So, if such a query, with wide-ranging possibilities, is issued in the context of a browsing and searching session by a single user, then the model of infusing the query with context based on the specifics of the current news story being watched may be applicable. If the visual stream shows many military scenes, with gunfights and armored vehicles, then perhaps stories with a similar on-the-ground focus might be most relevant. On the other hand, if the visual stream shows politicians, addresses, or the inside of the Senate chamber, then maybe stories with a political focus are more relevant. It is easy to see how the inclusion of snippets from the current document, as proposed by Kraft *et al.* [8], can be applied to the text of stories in news videos but also extended to the detected visual contents of those stories.

An interesting aspect of these context-based query adaptive systems is that it is likely that the reality of these systems is such that context is helpful in certain scenarios and detrimental in others. For example, in the case of including contextual keywords from the document that the user is currently viewing when issuing a search, the reality of the situation may be that the user is entirely switching gears and searching for a topic completely unrelated to the current document. In this case, the context infused from the current document is mostly noise and may only serve to make the query more ambiguous. Similarly, in a PC troubleshooting application, the user may be using a replacement PC, since the problem PC is not functioning correctly. Here, the registry states of the machine being used to search provide incorrect contextual information about the query at hand. So, clearly, contextual search systems can be beneficial, but can also serve to adapt the user's query in unintended and counterproductive ways. The context-based query adaptive systems that we have observed do not have mechanisms in place for predicting whether the context will help or not; however, these sorts of applications may require the peculiar property of being able to adapt themselves based on predictions on whether their context-based adaptation scheme will be successful.

C. Fusion Models

The fusion approaches across many of the systems that we have discussed are fairly uniform. The goal, in general, in these frameworks is to combine cues from various sources, whether it is a meta-search over various search engines or the combination of single-modal search methods for a video search application. By and large, the fusion approaches can all be reduced to a simple weighted linear fusion across the results from the individual search methods, where, typically, each individual search method provides a score or ranking for each document in the collection and the systems weight the score assigned to each document by each method according to the anticipated utility or reliability of the method for the type of query being processed. While there is some variability across approaches (i.e., fusion may be done using absolute scores or ranks, or one method might be used to gather the candidate set of results and another method might be used to just rerank those results), the core of the fusion method is fairly constant.

While this approach has an elegant simplicity and has been used with great success across a broad range of applications, its lack of subtlety begs for a more examined and better-considered approach. In video domains, for example, the modalities to be considered for fusion are not independent at all and the cues derived from one modality can have implications on the proper interpretation of some other modality. In some recent work [50], specific types of multimodal video search approaches which consider the interrelationships between the available modalities have been explored for a few important types of queries. In particular, much interest has been invested in searches for specific named persons in news videos, where searches over the text transcript can be used to surface a few candidate shots, which can then be promoted, given their response to face detection algorithms, or demoted, given the detected presence of anchor persons or other visual concepts not related to person queries.

Another fusion method that we have explored in recent work [51], [52] is a reranking approach, where the results of one search method are taken as a hypothesis of the actual desired ranking of the documents in the search set. This initial list of results is then mined in another modality to infer recurrent characteristics of relevant documents which can be used to reorder and refine the results. Both [51] and [52] are applied in the TRECVID broadcast news video domain. In [51], the initial search is taken from a text search over the speech recognition transcripts and the results are reranked by taking the top-returned documents to be pseudo-positive and pseudo-negative examples are sampled from elsewhere in the list. The documents are then reranked according to recurrent patterns in their low-level visual features. Fig. 6 shows the overall architecture of the reranking method. In [52], the initial search is taken from either text or query-by-visual-concept searches and the results are reordered according to the detection scores of a large collection of 374 high-level semantic visual concept detectors, using a similar approach to sampling pseudo-positives and pseudo-negatives. Such reranking approaches are promising avenues for exploiting cues from

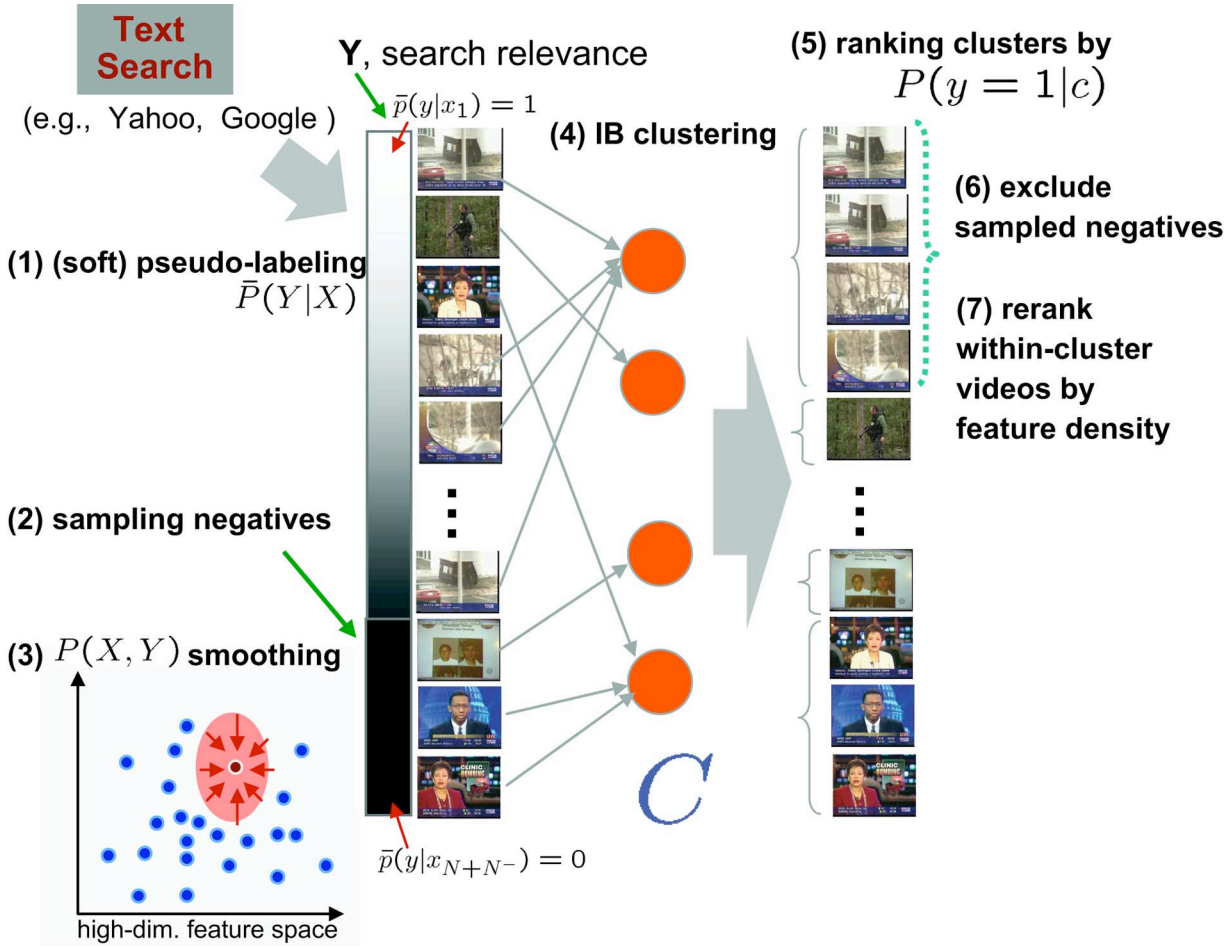

Fig. 6. ***System architecture of reranking method based on information bottleneck [51].***

modalities that are difficult to use in a straight-forward manner for certain queries. For instance, both approaches succeed in leveraging visual cues to enhance text search for named person searches despite the fact that query-by-example or query-by-concept search methods typically offer little help in these scenarios. Furthermore, in each of these applications, the effectiveness seems to vary across different queries. The decision to apply reranking and the order in which to apply the search methods is still and open research issue in query adaptive retrieval.

Part of the attractiveness of the straightforward tactics of the linear fusion approach, of course, is that the model parameters can be trivially discovered through exhaustive search for fusion weights and that the varying approaches and necessary classes of queries can be systematically discovered from past queries and associated relevance data. In contrast, it is unclear how to discover more complex search models that consider broader and, perhaps, nonlinear interrelationships between modalities and search methods. In the text retrieval community there has been some interest in the development of supervised learning approaches to better combine information sources for ranking [55]–[58]. The general thrust of most of these applications is that many generic text search methods and meta-search fusion models are simply designed through heuristics, which while reasonable in their formulation and powerful in their application, are ultimately only informed guesses at proper objective functions and fusion weights as determined by human designers. The contention, here, is that, given a set of queries for which relevance labels are available, data-driven techniques (using machine learning algorithms) can be applied to arrive at higher-performing fusion approaches.

For example, basic vector space models for text retrieval [21] work by weighting query and document terms according to a number of factors, typically adding weight for high frequency within a document and subtracting weight for high frequency across the entire collection and perhaps adjusting for the length of the document [24], [26], though the mechanisms for combining these various factors are largely driven by intuition and are perhaps not particularly principled. In [57], the authors undertake a data-driven exploration of term weighting.

Using a large corpus of relevance-labeled training queries, an evolutionary learning technique is applied to discover axioms that should exist in term weighting schemes in order to ensure effectiveness. The finding is that many of the heuristically defined, classical retrieval approaches [24], [26], abide by many of these data-driven axioms, though some new axioms also emerge from the data. It is found that applying the discovered cues to retrieval can improve upon the previous approaches. In [55], Nallapati explores the use of discriminative models for search. Various combinations of term weighting factors are used as a feature space to learn a support vector machine classifier over a set of training queries with relevance labels. The resulting model, which is essentially a weighting on each of the possible term weighting factors, is then applied to test queries. It is found that this approach is comparable to standard retrieval approaches for basic text retrieval tasks. In multimodal tasks, however, other nontext features (for example in web page retrieval: counts of incoming links or URL depth) can also be placed in the input feature space. The resulting model also includes weights on these aspects and is shown to significantly improve retrieval.

Similarly, in many popular meta-search applications the ranks or scores resulting from multiple search engines are combined using simple operations on the resulting ranks. Typical approaches to finding an aggregated rank: include taking the mean, median, minimum, or maximum rank of the document across various engines; counting the total number of higher ranked documents across the various results; or even modeling a Markov Chain over the rank relationships between documents. In [56], Liu *et al.* propose to formalize the rank aggregation problem into a supervised optimization framework, wherein they minimize the disagreements between the aggregation method and the ground truth over the training data. In practice, the optimization scheme results in a weighting scheme over each of the component search engines, such that each engine's results are not trusted equally. It is found that this learning-based approach to rank aggregation significantly outperforms other (nonlearning) methods.

The guiding principle behind these supervised approaches is that many ranking cues (be they term frequency, document frequency, link structures, or the results of an entire engine) have different levels of reliability and therefore ought to be accorded variable levels of trust. This reliability level (and the resulting weight accorded to the information source) can be learned through optimizing some criterion (such as retrieval performance) over a set of training queries with relevance labels. This is very much in tune with the weighted linear fusion method that we have discussed earlier, with the distinction that an exhaustive search for weighting parameters is perhaps much cruder than these proposed approaches and that linear fusion does not encompass all of the possible combinations that a more robust learning technique might discover.

Many of these investigations in text retrieval have been primarily in the direction of developing the fusion models themselves and have not yet looked towards the implications of query adaptation and how it can affect performance. In principle, though, such models can be learned independently for each query class, resulting in a unique supervised fusion model for each class. Learning models that can adapt in a classless adaptation framework (such as the difficulty prediction approaches discussed earlier) is still very much an open issue. The application of the findings from this research towards problems in query-adaptive multimedia search tasks may be a fruitful area for further investigation.

D. Learning Relevance

In almost any learning application, the existence of hard ground truth data is a key limiting factor in building robust and well-tested solutions. This problem is especially dire in multimedia retrieval applications, where practical example queries for training are scarce and actual human-supplied relevance labels are even scarcer. A case in point is the TRECVID benchmark data, where 24 queries and their associated relevance labels are generated each year, meaning that over the course of six years, only about 150 queries have been defined and evaluated. Given this sparseness of training data, it should come as no surprise that all of the class-dependent video search systems that have been proposed use no more than ten query classes: it is impossible to discover any deeper trends in query behavior. In reality, however, there are quite literally infinitely many possible keyword combinations or search intentions, and therefore there are likely a multitude of undiscovered classes of queries. So, while search relevance labels are a key component for designing and implementing all of the proposed query-adaptive systems, the generation of reliable and deep data for search relevance is an expensive and time-consuming proposition, sometimes prohibitively so.

In the MetaSeek system [13], we saw the use of relevance feedback (where users mark the top returned results as relevant or not) as a method for gaining approximate relevance labels for past queries. While relevance feedback has been an often-investigated technique in information retrieval, it has not been shown to be popular among users in any widespread application. Furthermore, many of the retrieval systems where relevance feedback can make a tangible difference are research-oriented multimedia search systems, where feedback is especially useful in the application of content-based visual features. Such systems typically have scalability issues which prevent them from reaching the widespread audience that could provide a considerable amount of feedback information.

Web-scale search engines, on the other hand, seem to sit in some other section of the universe of data availability, but are similarly impoverished. Major web search engines, and even those focused on images and

videos often handle thousands of real queries per second; however, it is hard to figure out which lessons, exactly, can be learned from such large volumes of queries. A typical metric that we have seen several times is a record of which documents have been clicked after the users have seen the results page from their query, though the meaning of these clicks is up for considerable interpretation. So, while it is not entirely unreasonable to interpret these clicks as assurances of the relevance of the clicked documents, the real sources of the motivations for clicking (or not clicking) a search result can range from any number of justifications: from poor summaries provided by the engine to losing interest in the search results before delving any deeper. So, while feedback remains an important tool for designing adaptive search methods, the available data ranges from deep evaluation over a few queries in the TRECVID community to scant evaluation over massive collection of queries in the web domain, whereas the target is really a mix of the best of each: a deep evaluation over massive sets of queries.

There has been much discussion lately about the use of humans and communities for labeling and annotation by making these mundane tasks fun and engaging. In the specific domain of image annotation, there are a few such examples. One is the Flickr photo community, where users can easily add annotations, or "tags," to their photos and share them openly. While users could always add labels to their personal collections, which would be stored on their personal computers, it was never a common practice. With Flickr, however, labeling photos becomes an easy and fun activity since there is the added incentive that others will find and view your photos through the tags that you have assigned. Another example is the ESP Game [53], where two remote participants are paired up, shown an image, and prompted to enter text labels to be associated with the image. If the two agree on an entered term, then both receive a point. In this case, there is the added benefit that the agreed-upon terms are highly likely to be useful annotations for the image, so these can be retained by the system to enable search or learning over the image collection. Meanwhile, users are happy to submit these labels, since they find the game fun and engaging.

Having observed this reversal of the trend in the willingness of people to label images, it stands to reason that previously abandoned mechanisms for gathering feedback and evaluation, such as relevance feedback, might be revisited in a new light. For example, when certain users conduct web searches on a research topic and find that no Wikipedia article exists on the topic, they may gather the relevant documents that they have uncovered through search and amass them on a new Wikipedia article. Given similarly easy-to-use tools, users of online multimodal search engines may feel motivated to create similar trails for future users. If the user has spent time to gather a collection of images relevant to a query topic, he or she may like to leave the set for future users who are on a similar hunt. Here, the motivation for relevance feedback has flipped from the traditional sense, where the benefit is expected to be immediate and observable right away by the user, to a more modern, community-oriented sense, where the benefits to the immediate user are only a sense of contribution and, perhaps, stature, but the benefits to future users are long-standing. Such a framework has the obvious added benefit that iterations on top of this labeling of relevance provided by the users can enhance and improve many aspects of the multimodal query-adaptive search system, including the selection of query classes and the learning of appropriate multimodal search methods.

The problem of sparse training data can, of course, be addressed by moving in an opposing direction as well, by creating and promoting methods which can successfully learn models for dealing with many types of queries without the need for mountains of training data. The direction toward query difficulty prediction is one such approach. In this approach, a query-adaptive model can be learned with comparatively few example training queries. This is in stark contrast with an approach which tries to discover and learn all the various types and classes of queries, of which there are likely to be infinitely many. Such a process requires a never-ending set of training queries.

E. Future of Query Adaptation

We have seen many aspects of the state of the art in query-adaptive multimodal retrieval, but what are the specific areas of opportunity for us to explore to move forward? Here, we will highlight the space with room for growth, again, with an eye towards applications in multimedia domains.

Difficulty prediction: Among the most promising approaches that we have seen in the text search community is the prediction of query performance. We have already seen applications demonstrated in meta-search and federation in the text domain. We have a great deal of hope that such approaches can be refined and extended to multimedia applications in the form of model prediction or modality selection frameworks.

Context aware search: Another promising direction in text applications has been the inclusion of a degree of awareness of the searcher's current task, state, or personal history in search. There is much room for extending this in multimedia applications.

Fusion models: At the heart of the query-adaptive approach to retrieval is the premise that various individual search approaches can be combined using different fusion models, depending on the query being handled. However, the majority of current multimodal search engines use very simple and unimaginative fusion approaches: typically a weighted sum of scores or ranks. There would seem to be ample opportunity here to explore and exploit the interactions between modalities and search methods with greater subtlety. On the other hand, multimodal

fusion can really only be as strong as the individual components available, and there is certainly plenty of work left to be done on the individual search components used in multimedia search systems.

Training data: A query-adaptive system is only as good as the queries (and relevance labels) that it has already seen. A major sticking point of multimodal search systems across a variety of domains is that acquiring enough training queries and relevance labels (both in quantity and in quality) is immensely difficult. One proposal that we have put forth is to turn such annotation tasks inside out and to allow communities of contributors to collect, save, and share the results of their search queries.

V. CONCLUSION

We have observed that as information retrieval systems become more mature and expand to many diverse media types and application domains, the modalities available and the sources of information that can be mined for cues in search also grows. A limitation that emerges from this growing richness of information sources is a clear emergence of idiosyncrasies across many individual types of queries, such that the methods employed to best utilize all the available information sources for one query can be disastrously inadequate for another query. The search systems, then, need to be empowered with the flexibility to be able to adapt to each incoming unique query in order to predict and apply a search strategy that is likely to provide the most useful retrieval results. We have seen that many retrieval models across many diverse types of data can be framed in such a discussion of properly selecting the right combination of search methods and information sources to respond to a given query. In the text search domain, examples of such systems classically include meta-search and federated search applications, where results from individual search engines over overlapping or separate collections need to be weighted and fused. In the video search domain, a single video clip can be represented and searched using many cues that are embedded in it, from speech recognition transcripts, to low-level visual features, to semantic visual concepts, to video optical character recognition text and a successful system needs to intelligently utilize all of these cues when responding to a query.

We have shown that a feasible and successful solution to this query-adaptation problem is to subdivide the space of all possible queries into discrete classes or types of queries, where queries belonging to the same class have similar optimal search strategies. This approach is seen to be beneficial across a broad range of applications, from federated search over text and/or image collections, general web search, and in particular for video retrieval from broadcast news applications. However, we have noted that such systems have an added layer of complexity, since some notion of the user's query intent must be derived from the very sparse pieces of information (usually a few keywords) that have been provided. We have also seen that many emerging solutions to query adaptation in the text retrieval domain eschew the approach of imposing a class-dependency structure on top of the problem and attempt to directly adapt retrieval approaches through prediction of the individual success rate of the individual search methods available, or through the infusion of context about the user.

The *status quo* in multimodal video search is broadly dominated by class-dependent approaches. These approaches have dramatic and wide-ranging advantages over nonadaptive approaches but are also subject to many difficulties induced by the need for extracting meaning from sparse queries and the limited availability of relevance-labeled training data needed to adequately partition the query space into classes. Methods to predict the search method performance directly, via query difficulty prediction frameworks, have great promise for bridging some of these difficulties. In particular, class-dependent approaches effectually act as more of a stepping stone between the user query and the optimal search method to answer the query, and attempting to extract user intent through such sparse input can be detrimental. The difficulty prediction approaches are a more direct route between the query and the actual likelihood of success of the search method and are, therefore, the more desirable approach.

The approaches used to fuse information from disparate sources or component search methods are fairly simple and standard across almost all applications, typically using a weighted sum of available scores from search results. And while such methods are delivering strong results, future research in examining the subtle interactions between modalities and information sources in various types of media and retrieval documents may yield insight into fusion methods that are better suited for exploiting the structure and information leveraged by search methods. ■

REFERENCES

[1] K. Mc Donald and A. Smeaton, "A comparison of score, rank and probability-based fusion methods for video shot retrieval," in *Proc. CIVR 2005*, 2005.

[2] A. Natsev, M. Naphade, and J. Tesic, "Learning the semantics of multimedia queries and concepts from a small number of examples," in *Proc. 13th Annu. ACM Int. Conf. Multimedia*, 2005, pp. 598–607.

[3] R. Yan and A. Hauptmann, "Probabilistic latent query analysis for combining multiple retrieval sources," in *Proc. 29th Annu. Int. ACM SIGIR Conf. Research Development Information Retrieval*, 2006, pp. 324–331.

[4] K. Cai, J. Bu, C. Chen, and P. Huang, "Automatic query type classification for web image retrieval," in *Proc. 2007 Int. Conf. Multimedia Ubiquitous Eng.*, 2007, pp. 1021–1026.

[5] I. Kang and G. Kim, "Query type classification for web document retrieval," in *Proc. 26th Annu. Int. ACM SIGIR Conf. Research Development Information Retrieval*, 2003, pp. 64–71.

[6] R. Zhang, R. Sarukkai, J.-H. Chow, W. Dai, and Z. Zhang, "Joint categorization of queries and clips for web-based video search," in *Proc. 8th ACM Int. Workshop Multimedia Information Retrieval*, New York, 2006, pp. 193–202.

[7] T. Chua, S. Neo, K. Li, G. Wang, R. Shi, M. Zhao, and H. Xu, "TRECVID 2004 search and feature extraction task by NUS PRIS," in *Proc. TRECVID 2004 Workshop*, 2004.

[8] R. Kraft, F. Maghoul, and C. C. Chang, "Y!q: Contextual search at the point of inspiration," in *Proc. 14th ACM Int. Conf. Inform. Knowledge Management*, New York, 2005, pp. 816–823.

[9] R. Kraft, C. C. Chang, F. Maghoul, and R. Kumar, "Searching with context," in *Proc. 15th Int. Conf. World Wide Web*, New York, 2006, pp. 477–486.

[10] J. Pitkow, H. Schutze, T. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar, and T. Breuel, "Personalized search," *Commun. ACM*, vol. 45, no. 9, pp. 50–55, 2002.

[11] E. M. Voorhees, N. K. Gupta, and B. Johnson-Laird, "Learning collection fusion strategies," in *Proc. 18th Annu. Int. ACM SIGIR Conf. Research Development Information Retrieval*, New York, 1995, pp. 172–179.

[12] M. Beigi, A. B. Benitez, and S.-F. Chang, "Metaseek: A content-based meta-search engine for images," in *Proc. SPIE Conf. Storage Retrieval Image and Video Databases VI (IST/SPIE-1998)*, San Jose, CA, Jan. 1998, vol. 3312.

[13] A. B. Benitez, M. Beigi, and S.-F. Chang, "Using relevance feedback in content-based image meta-search," *IEEE Internet Computing*, vol. 2, no. 4, pp. 59–69, Jul. 1998.

[14] L. Kennedy, A. Natsev, and S. Chang, "Automatic discovery of query-class-dependent models for multimodal search," in *Proc. 13th Annu. ACM Int. Conf. Multimedia*, 2005, pp. 882–891.

[15] S. Chang, W. Hsu, L. Kennedy, L. Xie, A. Yanagawa, E. Zavesky, and D. Zhang, "Columbia University TRECVID-2005 video search and high-level feature extraction," in *Proc. NIST TRECVID Workshop*, Gaithersburg, MD, Nov. 2005.

[16] S. Chang, W. Hsu, W. Jiang, L. Kennedy, D. Xu, A. Yanagawa, and E. Zavesky, "Columbia University TRECVID-2006 video search and high-level feature extraction," in *Proc. NIST TRECVID Workshop*, Nov. 2006.

[17] R. Yan, J. Yang, and A. Hauptmann, "Learning query-class dependent weights in automatic video retrieval," in *Proc. 12th Annu. ACM Int. Conf. Multimedia*, 2004, pp. 548–555.

[18] J. Smith and S. Chang, "VisualSEEk: A fully automated content-based image query system," in *Proc. Fourth ACM Int. Conf. Multimedia*, 1997, pp. 87–98.

[19] M. Flickner, H. Niblack, W. Ashley, J. Dom, B. Gorkani, M. Hafner, J. Lee, D. Petkovic, D. Steele, D. Yanker *et al.*, "Query by image and video content: The QBIC system," *Computer*, vol. 28, no. 9, pp. 23–32, 1995.

[20] S. Robertson and K. S. Jones, "Relevance weighting of search terms," *J. Amer. Soc. Information Science*, vol. 27, 1977.

[21] G. Salton, A. Wong, and C. S. Yang, "A vector space model for information retrieval," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, Nov. 1975.

[22] V. Castelli, L. Bergman, C.-S. Li, and J. Smith, "Search and progressive information retrieval from distributed image/video databases: The SPIRE project," in *Research and Advanced Technology for Digital Libraries*, C. Nikolaou and C. Stephanidis, Eds. New York: Springer, 1998.

[23] S. E. Robertson and S. Walker, "Some simple effective approximations to the 2poisson model for probabilistic weighted retrieval," in *Proc. ACM SIGIR*, 1994, pp. 232–241.

[24] A. Singhal, C. Buckley, and M. Mitra, "Pivoted document length normalization," in *Proc. ACM SIGIR*, Aug. 1996, pp. 21–29.

[25] J. L. Gauvain, L. Lamel, and G. Adda, "The LIMSI broadcast news transcription system," *Speech Commun.*, vol. 37, no. 1–2, pp. 89–102, 2002.

[26] S. E. Robertson, S. Walker, M. Hancock-Beaulieu, A. Gull, and M. Lau, "Okapi at TREC4," in *Text REtrieval Conf.*, 1992.

[27] E. Yom-Tov, S. Fine, D. Carmel, and A. Darlow, "Learning to estimate query difficulty: Including applications to missing content detection and distributed information retrieval," in *SIGIR*, 2005.

[28] NIST TREC Video Retrieval Evaluation. [Online]. Available: http://www-nlpir.nist.gov/projects/trecvid/

[29] A. Amir, J. Argillander, M. Campbell, A. Haubold, G. Iyengar, S. Ebadollahi, F. Kang, M. R. Naphade, A. Natsev, J. R. Smith, J. Tesic, and T. Volkmer, "IBM research TRECVID-2005 video retrieval system," in *Proc. NIST TRECVID Workshop*, Gaithersburg, MD, Nov. 2005.

[30] J. Adcock, A. Girgensohn, M. Cooper, T. Liu, L. Wilcox, and E. Rieffel, "FXPAL experiments for TRECVID 2004," in *Proc. TRECVID 2004 Workshop*, 2004.

[31] C. Snoek, J. van Gemert, J. Geusebroek, B. Huurnink, D. C. Koelma, G. P. Nguyen, O. deRooij, F. J. Seinstra, A. Smeulders, C. J. Veenman, and M. Worring, "The MediaMill TRECVID 2005 semantic video search engine," in *Proc. NIST TRECVID Workshop*, Gaithersburg, MD, Nov. 2005.

[32] M. Campbell, S. Ebadollahi, M. Naphade, A. P. Natsev, J. R. Smith, J. Tesic, L. Xie, and A. Haubold, "IBM research TRECVID-2006 video retrieval system," in *Proc. NIST TRECVID Workshop*, Gaithersburg, MD, Nov. 2006.

[33] A. G. Hauptmann, M.-Y. Chen, M. Christel, W.-H. Lin, R. Yan, and J. Yang, "Multi-lingual broadcast news retrieval," in *Proc. NIST TRECVID Workshop*, Gaithersburg, MD, Nov. 2006.

[34] E. Yom-Tov, S. Fine, D. Carmel, and A. Darlow, "Meta-search and federation using query difficulty prediction," in *Proc. SIGIR 2005 Query Prediction Workshop*, Jul. 2005.

[35] J. R. Smith and S.-F. Chang, "Visually searching the web for content," *IEEE Multimedia*, vol. 4, no. 3, pp. 12–20, Jul. 1997.

[36] J. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. Jain, and C. Shu, "The virage image search engine: An open framework for image management," in *Proc. SPIE, Storage and Retrieval for Still Image and Video Databases IV*, 1996, pp. 76–87.

[37] R. Kraft and J. Zien, "Mining anchor text for query refinement," in *Proc. 13th Int. Conf. World Wide Web*, 2004, pp. 666–674.

[38] S. Lawrence, "Context in web search," *IEEE Data Eng. Bull.*, vol. 23, no. 3, pp. 25–32, 2000.

[39] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, "Rank aggregation methods for the web," in *Proc. 10th Int. Conf. World Wide Web*, 2001, pp. 613–622.

[40] E. Selberg and O. Etzioni, "The MetaCrawler architecture for resource aggregation on the web," *IEEE Expert*, vol. 12, no. 1, pp. 11–14, 1997.

[41] C. Silverstein, M. Henzinger, H. Marais, and M. Moricz, "Analysis of a very large web search engine query log," in *SIGIR Forum*, 1999, vol. 33, no. 1, pp. 6–12.

[42] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin, "Placing search in context: The concept revisited," in *Proc. 10th Int. Conf. World Wide Web*, 2001, pp. 406–414.

[43] E. Glover, S. Lawrence, W. Birmingham, and C. Giles, "Architecture of a meta-search engine that supports user information needs," *Ann Arbor*, vol. 1001, p. 48 109.

[44] L. Page, S. Brin, R. Motwani, and T. Winograd, "The Pagerank citation ranking: Bringing order to the Web, 1998, Tech. Rep., Stanford Digital Library Technologies Project.

[45] W. Croft, "Combining approaches to information retrieval," *Advances Inform. Retrieval*, pp. 1–36, 2000.

[46] E. Fox and J. Shaw, *Combination of multiple searches*, NIST Special Pub., no. 500215, pp. 243–252, 1994.

[47] J. Wen, N. Lao, and W. Ma, "Probabilistic model for contextual retrieval," in *Proc. 27th Annu. Int. Conf. Research Development Information Retrieval*, 2004, pp. 57–63.

[48] 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. TRECVID 2004 Workshop*, 2004.

[49] ZoneTag. [Online]. Available: http://zonetag.research.yahoo.com/

[50] A. Hauptmann, R. Baron, M. Chen, M. Christel, P. Duygulu, C. Huang, R. Jin, W. Lin, T. Ng, N. Moraveji *et al.*, "Informedia at TRECVID 2003: Analyzing and searching broadcast news video," in *Proc. TRECVID*, 2003.

[51] W. Hsu, L. Kennedy, and S.-F. Chang, "Video search reranking via information bottleneck principle," in *ACM Multimedia*, Santa Barbara, CA, 2006.

[52] L. Kennedy and S.-F. Chang, "A reranking approach for context-based concept fusion in video indexing and retrieval," in *Proc. Conf. Image Video Retrieval*, Amsterdam, The Netherlands, 2007.

[53] L. von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proc. SIGCHI Conf. Human Factors in Computing Systems*, 2004, pp. 319–326.

[54] A. Hwang, S. Ahern, S. King, M. Naaman, R. Nair, and J. Yang, "Zurfer: Mobile multimedia access in spatial, social and topical context," in *Proc. 15th Int. Conf. Multimedia*, 2007, pp. 557–560.

[55] R. Nallapati, "Discriminative models for information retrieval," in *Proc. 27th Annu. Int. Conf. Research Development Information Retrieval*, 2004, pp. 64–71.

[56] Y. Liu, T. Liu, T. Qin, Z. Ma, and H. Li, "Supervised rank aggregation," in *Proc. 16th Int. Conf. World Wide Web*, 2007, pp. 481–490.

[57] R. Cummins and C. O'Riordan, "An axiomatic study of learned term-weighting schemes," in *Proc. SIGIR'07 Workshop Learning to Rank for Information Retrieval (LR4IR-2007)*, Jul. 2007.

[58] C. Scheel, N. Neubauer, A. Lommatzsch, K. Obermayera, and S. Albayrak, "Efficient query delegation by detecting redundant retrieval strategies," in *Proc. SIGIR'07 Workshop Learning to Rank for Information Retrieval (LR4IR-2007)*, Jul. 2007.

[59] L. Xie, A. Natsev, and J. Tesic, "Dynamic multimodal fusion in video search," in *Proc. IEEE Int. Conf. Multimedia and Expo*, 2007, pp. 1499–1502.

ABOUT THE AUTHORS

Lyndon Kennedy (Student Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from Columbia University, New York, in 2003 and 2005, respectively. He is currently working toward the Ph.D. degree in the Department of Electrical Engineering at the same university.

His research interests include investigating techniques for automatically indexing and searching digital image and video collections as a Graduate Research Assistant at the Digital Video and Multimedia Lab, Columbia University.

Shih-Fu Chang (Fellow, IEEE) leads the Digital Video and Multimedia Lab, Department of Electrical Engineering, Columbia University, New York, conducting research in multimedia content analysis, image/video search, multimedia forgery detection, and biomolecular image informatics. Systems developed by his group have been widely used, including VisualSEEk, VideoQ, WebSEEk for visual search, TrustFoto for online image authentication, and WebClip for video editing. His group has also made significant contributions to the development of the MPEG-7 international multimedia standard. He worked in different capacities in several media technology companies.

Dr. Chang is the Editor-in-Chief of *IEEE Signal Processing Magazine* (2006 to present) and a recipient of a Navy ONR Young Investigator Award, IBM Faculty Development Award, and NSF CAREER Award. He served as a General Co-chair for ACM Multimedia Conference 2000 and IEEE ICME 2004.

Apostol (Paul) Natsev received the M.S. and Ph.D. degrees in computer science from Duke University, Durham, NC, in 1997 and 2001, respectively.

He is a Research Staff Member and Manager of the Multimedia Research Department, IBM T.J. Watson Research Center Hawthorne, NY. He joined IBM Research in 2001. His primary research interests are in the areas of multimedia semantic indexing and search, multimedia understanding and machine learning, as well as multimedia databases and query optimization. He is an active participant in the NIST TREC Video Retrieval (TRECVID) evaluation and leads the IBM video search team, which has achieved excellent performance in TRECVID several years in a row.

Dr. Natsev is a founding member of the IBM Research Multimedia Analysis and Retrieval project (MARVEL), which was awarded the Wall Street Journal Innovation Award (Multimedia category) in 2004, and he received an IBM Outstanding Technical Accomplishment Award in 2005.