Semi-Supervised Rank Learning for Multimedia Known-Item Search

David Etter
Department of Computer Science
George Mason University
etterd@gmail.com

Carlotta Domeniconi
Department of Computer Science
George Mason University
carlotta@cs.gmu.edu

ABSTRACT

Known Item Search (KIS) is a specialized task of the general multimedia search problem. It describes the scenario where a user has previously seen a video and wants to find it again in a large collection using a text description. While there exists only one correct answer to a query (or topic), the goal is to return a ranked list of videos most likely to satisfy the request. This search problem includes content from speech, visual, and meta-data, and it is not clear how the individual modalities should be combined in the final result. Reranking models have been shown to be effective in problems such as image search, but the single ground truth video for a topic presents a challenge for building a model. In this paper, we propose a semi-supervised rank learning approach to the multimedia problem. We use a large training set of topics and ground truth videos to learn a pairwise ranking model based on gradient boosted regression trees. We define a learning feature space that consists of features derived from topics, videos, and topic-video dependent results. To overcome the KIS class imbalance problem, a set of pseudo positive training examples are identified from each of the multimedia modalities. This semi-supervised approach uses a ground truth video to select similar videos in each of the individual modalities. We then model the similarities as a graph and use a K-Step Markov approach to estimate the importance of nodes in the graph relative to the truth root node.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: [Retrieval models, Search process]; H.2.4 [**Systems**]: [Multimedia databases]

1. INTRODUCTION

The convenience of smart phones with high quality video capture capability has led to an explosion in the size of personal and internet video archives. Consumers now use their phones to capture and share short clips of personal activities, news events, blogs, and how-to instructions. According to YouTube press [23] over 100 hours of video are uploaded each minute and over 6 billion hours of video are watched each month. As the volume of content in these repositories expands, there is an increased need for effective multimedia search which exploits the multiple modalities of a video.

Known Item Search (KIS) is a specialized task of the general multimedia search problem. KIS describes the scenario where a user has seen a video before, must formulate a text description based on what he remembers, and knows that there is only one correct answer. As an example, consider the TRECVid 2012 [29] KIS topic 1213, "Find the video of a round silver colored weather satellite, men in white hard hats, nose cone placed in rocket, and rocket lifting off". Table 1 displays the known item for the satellite topic and lists example content for its speech, meta-data, and visual modalities. The KIS problem takes as input a text only description and returns a ranked list of videos, most likely to match the known item. Results are measured using the inverted rank of the known item in the result list.

The KIS problem poses a number of challenges for multimedia search. The first challenge is the modality gap between the text only topic and the speech, meta-data, and visual content of the repository. Information retrieval approaches can be used on the text modalities such as meta-data and speech to text, but it is not clear how this approach could make use of visual features from a multimedia object. Also, the KIS problem is unique from other multimedia search tasks in that it has only one correct answer.

A second challenge is how to effectively model the search and ranking problem in the unique feature space of each multimedia modality. Meta-data can include author content fields such as FileName, Title, Description, Subject, and Keywords. This content is often incomplete or missing, varies in length, and include numerous misspellings. Automated Speech to Text (ASR) and Optical Character Recognition (OCR) are sound and vision features that can be mapped into a text feature space, but are often noisy and incomplete. Visual features, such as color, texture, and local keypoint can be extracted from the video content, but again it is not clear how to map the text topic request to these feature spaces.

To overcome these challenges, we propose a semi-supervised rank learning approach to the KIS problem. This approach uses a large training set of KIS topics and ground truth videos to learn a pairwise ranking model based on gradient boosted regression trees [2]. We derive a final learning feature space that combines the individual modality features

ICMR'14 Apr 01-04 2014, Glasgow, United Kingdom
ACM 978-1-4503-2782-4/14/04.
http://dx.doi.org/10.1145/2578726.2578759.

from topics, videos, and topic-video dependent results. As an example, we use information retrieval models to initially rank and score a topic against each of the individual derived text modalities. Each result is then modeled in the feature space where the classifier learns a weight vector that identifies the contribution of each modality. The KIS rank learning approach provides an effective solution to the problem of modeling the search and ranking feature space of each multimedia modality.

The class imbalance is one of the challenges with applying a rank learning approach to the KIS problem. Consider a KIS task where the top 100 ranked results are returned for a given topic. Given the nature of the KIS problem, the result set consists of 99 negative examples and 1 positive example. To overcome the class imbalance problem, a set of pseudo positive training examples are identified from each of the multimedia modalities. This semi-supervised approach uses a ground truth video to identify similar videos in each of the individual modalities. To identify pseudo examples we model the similarities as a graph and use a K-Step Markov approach [30] to estimate the importance of nodes in the graph relative to the truth root node. Each pseudo positive example is then assigned a decreasing graded relevance based on the distance from the truth video. This approach allows us to include both text and visual modalities when identifying pseudo positive examples.

Our contributions to the multimedia retrieval community include the following:

1. We construct a feature space consisting of topic specific, topic-video dependent, and video specific features, calculated from the meta-data, speech, and visual modalities of our text topics and video repository.

2. We introduce the concept of pseudo-positive KIS examples to offset the class imbalance of having a single known item. Pseudo-positive examples are identified in a similarity graph, using a K-Step Markov to estimate the importance of nodes relative to the truth root node.

3. We study a pairwise rank learning approach to the multimedia KIS problem using gradient boosted regression trees.

2. RELATED WORK

KIS in the context of multimedia search has been studied as part of a TRECVid task [29] in 2010, 2011, and 2012. The video collection for the TRECVid task consists of approximately 8,000 Internet Archive Videos and 300 topics and judgments for each task year. The task included both automatic and interactive systems, with results measured using the mean inverted rank.

An overview of the various KIS systems used during the 2011 TRECVid evaluation can be found in [4] and [6]. Task participants attempted to bridge the understanding gap between a topic text and the video collection. Text based approaches included enriching topics and meta-data using external knowledge such as Wikipedia, ontologies, or translations. An approach to bridging the visual modality gap was to identify examples images from a web image search engine. Most of the task participants concluded that the visual modalities provided little benefit to the final rankings. The top scoring team in the task created a classifier that transformed the original text topic into a set of shorter modality specific queries. All of the participating systems attempted to fuse multiple modality results. Our rank learning approach differs from all of the modality based fusion approaches used by the TRECVid participants. We attempt to learn pairwise ranking from a training set that includes both modality specific scores and visual features.

Table 1: Known Item from TRECVid 2012

<table>
<tr><td colspan="3">**ASR**: Satellite electric eyes will scan the Earth s cloud cover broadcasting another reporting of the weather stations...space vehicle one of the most technically sophisticated of the space rockets misfortune six times...</td></tr>
<tr><td colspan="3">**FileName**: $1959-02-19_{W}eather_{E}ye._{-}o-_{.}1959-02-19_{W}eather_{E}ye_{5}12kb.mp4$</td></tr>
<tr><td colspan="3">**Meta:Title**: Weather Eye Vanguard II Satellite Scans Sky From Space 1959 02 19</td></tr>
<tr><td colspan="3">**Meta:Description**: Vanguard II satellite placed in nose cone of rocket launched partial newsreel brief silence at start of story</td></tr>
<tr><td colspan="3">**OCR**: rwx EYE Vanguard Satellite Suns Sky From Space Hum</td></tr>
<tr><td>
</td><td> </td><td> </td></tr>
</table>

Research related to the KIS problem has occurred in multiple text based domains such as person document [15] , web [14], email [7], Twitter, and Facebook [15]. The results in [15] show that a mixture of language models which combine evidence from different representations is an effective approach for this type of document retrieval. Personal document search is studied in [14] over email, presentations, web pages, and pdfs. They investigated techniques for improving document type prediction in personal desktop search. Their model uses type specific meta-data to generate a field-based collection query likelihood. Type specific results are then merged into a final ranked list which improves overall retrieval performance.

Learning to Rank describes a machine learning approach for constructing ranking models over a set of training data [3] [17]. Approaches to rank learning can be categorized as pairwise, pointwise, or listwise learning [3]. This framework has been successfully applied to rank learning by the information retrieval community [11] [31]. LETOR [24] provides a number of large datasets along with queries, ground truth, and precomputed feature vectors for studying the learning to rank problem for information retrieval. Query features are studied in [18] to evaluate their effectiveness on learning to rank models. Query features are fixed values of the query itself, and do not change across documents. These features include values such as the number of unique tokens, the number of named entities, and the number of retrieved document categories for a unique query. Our approach to rank learning extends the query feature approach to the KIS multimedia retrieval domain as a set of topic features which are video independent. The topic features are combined with video and topic-video dependent features to form the KIS feature space.

Reranking has been studied in the context of both image search [32], video search, and semantic indexing. The work in [26] improved MAP on a semantic indexing and retrieval task by reranking an initial video shot score, using a model that considers the homogeneity of the video to which it belongs. Automatic video search reranking is studied by [13], where they identified an initial result set using text search, concept detection, and image query-by-example. The top and bottom ranked items were then used as pseudo positive and negative examples to train a test time model to discover co-occurrence patterns. Their use of pseudo examples differ from our approach, which is applied at training time and identifies similar videos using a graded relevance to a ground truth video. A graph reranking approach was used by [16] to improve the initial text search results for a video search task. In contrast, our graph approach is not used to rerank an initial result set, but instead to identify additional pseudo positive examples for training.

3. OUR APPROACH

We propose a semi-supervised rank learning approach to the multimedia KIS problem. First we define a feature space derived from topics, videos, and topic-video dependent results. Next we identify a set of pseudo positive training examples using a similarity graph constructed from the ground truth videos. The pseudo positive examples are used to assign a graded relevance to our topic-video training pairs. Finally, a gradient boosted regression tree algorithm [2] is used to learn a pairwise ranking model over the training set.

Given a set of known item topics $T = (t_1, \ldots, t_m)$ where m is the size of the topic set, each KIS topic is derived from a video repository $V = (v_1, \ldots, v_n)$ where n is the size of the collection. We define the KIS task as a mapping:

$$F(t_i, V) = (s_{v_1}, \ldots, s_{v_n}) \tag{1}$$

where t_i is a known item topic, V is the video repository, and s_{v_i} is the calculated score for video v_i.

3.1 Feature space

To model the KIS feature space, we identify three categories of features which are defined in Table 2. Topic Features are derived from the text of the known item topic and include term count, unique term count, and named entity counts [19]. The person, location, and organization named entities are identified using a sequence tagger [8]. The topic features are video independent and allow the tree-based ranking algorithm to identify groups of similar topics that can be model as sub-trees.

Topic-Video Features are derived from the output ranks and scores of various information retrieval models applied to the topic video pair. The scoring models used in this work include term frequency inverse document frequency (TFIDF), probabilistic best match (BM25) [28] [25], and language models (LM) [22]. These models are applied to each of the text based fields: ASR [12], OCR, FileName, Title, Description, Subject, and Keywords. This category of features also includes the number of term matches, percentage of term matches, calculated term frequency (TF), and the calculated inverse document frequency (IDF) for each text field.

The final category is video Features which are derived from the automated speech, meta-data, or visual components of the video. The text-based features for ASR and meta-data include term counts by field and identified person, location, and organization named entities. From the visual modalities, we derive both low-level and high-level image features.

Low-level features include edge [21] histograms, color histograms [27], and bag-of-visual words using SURF [1] keypoint features. The edge direction histogram provides a compact and computational efficient representation of the video. The descriptor divides the image into 16 sub-images, using a 4 × 4 grid, to allow for the calculation of localized edge distributions. The edge histogram consists of four directional edges and one non-direction edge for each of the sub-images and is represented by an 80-dimensional feature vector. Keypoint detectors [1] attempt to detect a small set of locally stable points and their surrounding region. Keypoints are clustered into a set of visual words to form the vocabulary for a bag-of-words. The size of the visual word vocabulary and the weight of each term is a parameter to the final representation.

High-level features are the semantic concepts that we use to describe objects, events, and activities in video. Table 3 shows example concepts and descriptions from the TRECVid [29] semantic indexing task. These concepts provide an approach to bridge the semantic gap between text descriptions and the low-level features of a video. A concept specific model is trained over the low-level features which can then be used to assign a confidence to a previously unseen video.

3.2 Pseudo positive examples

Training a machine learning algorithm over a feature vector, derived from a KIS topic-video result, presents a challenge in the number of positive examples. Consider the TRECVid KIS task, where a system is required to return a ranked list of the 100 top videos for each topic. The output from this task results in one positive and 99 negative examples per topic. This large class imbalance creates a challenge for any supervised learning algorithm.

Further inspection of the ranked result list, reveals that a number of negative examples are similar to the ground truth item in one or more modalities. Topic 891 of the 2012 TRECVid KIS task states, “Find a video of yellow bus driving down winding road in front of building with flags on roof and driving past geysers”. Consider the three example videos in Table 4. The first video, titled “100 th Anniv Old Faithful Inn Yellow Busses Ride the Old Road”, is the ground truth video for topic 891 and the two additional videos are clearly similar in title, meta-data description, and video image. The identified similarities in multiple modalities show that while a single correct answer exists for a given KIS topic, the problem could be generalized to one of graded relevance rather than simple binary classification. Identifying additional pseudo positive KIS examples helps to lessen the class imbalance problem and results in boosting similar videos higher in the ranked result list.

We propose a semi-supervised learning approach to KIS where the training set for a given topic includes both the single truth example and a set of pseudo positive examples. The pseudo positive examples are identified by similarity to the truth video across all of its modalities. Each pseudo positive example is assigned a decreasing graded relevance based on the distance from the truth video.

Table 2: Topic, Topic-Video Dependent, and Video Features

Type	Feature	Description	Feature Count
Topic	Term	Count of terms	1
	Unique Term	Unique terms	1
	Person	Count of Person Named Entity	1
	Location	Count of Location Named Entity	1
	Organization	Count of Organization Named Entity	1
Topic-Video Dependent	TFIDF	TFIDF Weight Model	7
	BM25	BM25 Probabilistic Model	7
	LMIR	Language Model	7
	Percent Term	Percentage of Term Match	7
	IDF	Inverse Document Freq of Match	7
	TF	Term Freq of Match	7
Video	Term	Count of terms	7
	Unique Term	Unique terms	7
	Person	Count of Person Named Entity	7
	Location	Count of Location Named Entity	7
	Organization	Count of Organization Named Entity	7
	Edge	Edge direction histogram	80
	Color	Color histogram	64
	Keypoint	Visual Bag-of-words	50
	Concepts	Semantic concepts	50

Table 3: Semantic Concepts

Name	Definition
Airplane Flying	An airplane flying in the sky
Car	Shots of a car
Cityscape	View of a large urban setting, showing skylines and building tops.
Demonstration	One or more people protesting. May or may not have banners or signs
Road	Shots depicting a road

To identify pseudo examples we model the similarities as edges in a graph and estimate the importance of nodes in the graph relative to the truth root node.

For a given KIS topic t_i, we construct a directed graph $G_i = (V_i, E_i)$, where V_i is the set of video nodes in the topic specific graph and E_i is the set of edges. We define the ordered pair (u, v) as the directed edge connecting video node u to video node v.

The topic specific graph is iteratively constructed by initially selecting the truth video as the root node and performing a similarity search in the video collection using each modality. The result nodes from each iteration are used as search nodes for the next iteration. We define the iterative graph construction as follows: $\forall j = 1, \ldots, T$, video v_j is added to V_i and (u, v_i) is added to E_i if

$$F_i(u, v_j) > \alpha_i, \qquad (2)$$

where T is the size of the video collection, and $F_i(u, v_j)$ is a modality similarity score between the current root u and each video v_j in the collection. A modality specific threshold α_i is used to select the subset of videos. The α_i is empirically selected for each modality using a validation set. The graph is initially constructed as a directed multigraph where a video node pair (u, v) may be selected as a directed edge by more than one modality. To identify a single edge weight connecting u, v we traverse the graph, collapsing multiple edges using the maximum of the modality specific weights. The final graph is modeled as a directed acyclic graph (DAG), and therefore does not contain loops or parallel edges.

Given our topic specific similarity graph we would like to assign an importance measure $I(v|u)$ to each video node v in the graph with respect to the truth node u. To calculate I we use a K-Step Markov approach [30] which generates random walks of fixed length K, beginning at the root node u. The importance of node v to the root u is defined as:

$$I(v|u) = [Mp_u + M^2 p_u \ldots M^k p_u]_v \qquad (3)$$

where k is a fixed number of steps, M is the transition probability matrix, and p_u is the initial root probability set.

Table 4: Similar Videos

Title: 100 th Anniv Old Faithful Inn Yellow Busses Ride the Old Road Desc: As a part of the Old Faithful Inn 100 th Anniversary Celebration yellow National Park busses. . .	
Title: Yellowstone Porcelain Basin Desc: . . . warm spring morning to shoot video of the Porcelain Basin area within Norris Geyser Basin. . .	
Title: Yellowstone Snowloads Diminish and Lion Geyser Roars Desc: When the roads in Yellowstone are clear enough to safely allow cars. . .	

3.3 Multimedia Rank Learning

To learn a ranking model we follow a pairwise ranking approach which considers the relative order between pairs of videos. Pairwise ranking algorithms use binary classification in order to minimize the number of misclassified pairs.

Our approach to multimedia rank learning uses a framework based on gradient boosted regression trees [9] [10]. This framework has been successfully applied to rank learning by the information retrieval community [11] [31] and was the base approach for all of the winning teams at the Yahoo! Learning to Rank Challenge [5] [2].

The gradient boosting framework uses a stage-wise approach to generate an ensemble of weak models, each a simple regression tree, that when combined produce a strong rank learning classifier. The algorithm uses regression trees to perform gradient descent in function space and can be trained to minimize a general differentiable loss function. The final ranking score is a linear combination of the output scores from each of the simple regression tree models.

Consider a feature vector $x \in R^d$, derived from the topic, video, and topic-video dependent feature set. The boosted regression tree model maps the input feature vector to a ranking score $f(x) \in \Re$:

$$f(x) = \sum_{i=1}^{N} \beta_i \times f_i(x) \tag{4}$$

where $f_i(x)$ is the learned model for a single regression tree, β_i is the learned weight associated with that tree, and N the number of trees. The function $f_i(x)$ produces a ranking score by traversing the regression tree and evaluating a particular feature x_i with the weight at the given node. The final output score for $f_i(x)$ is the fixed value from the leaf node selected by the best path traversal.

A high level review of our approach is described in Algorithm 1. Given a set of KIS training topics Q and a video collection T, we begin by building a feature vector x_q consisting of the topic features, video features, and topic-video dependent features for each topic-video pair. Next, a video similarity graph is constructed, where a truth video Q'_q is the root node of the graph. An importance score is given to each node of the graph using a K-Step Markov approach. The graph node score is used to determine the relevance weight assigned to each topic-video pair. The KIS rank learning model is generated using gradient boosted regression trees trained over the relevance weighted feature vector. Given a previously unseen topic, we construct a feature vector for each new topic-video pair and apply the KIS rank learning model to determine a final ranking score.

Algorithm 1: KIS Learn to Rank

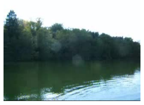
```
**Input**: $Q$ KIS topic set, $T$ video collection, $Q'$ truth set
**Output**: $f(x)$ regression tree model
**for** *each q in Q* **do**
    $x_q$ = build-FeatureVector($q$,$T$);
    $G_q$ = build-SimGraph($Q'_q$,$T$);
    $I_q$ = run-KStepMarkov($G_q$, $Q'_q$);
    $x'_q$ = assign-Relevance($x_q$, $I_q$);
**return** $f(x)$ = train-KISRankLearn($x'$);
```

4. EXPERIMENTS

Experiments are conducted using the known item topics and video repository from the TRECVid 2012 evaluation. This video collection consists of approximately 8,000 Internet Archive Videos distributed in MPEG-4/H.264 format and released under the Creative Commons license. These videos total about 200 hours and have a duration between 10 seconds and 3.5 minutes.

Table 5 provides a sample of the topics and ground truth known item images from the repository. Topics are provided to the system as a text only description of both the audio and visual components of the video. The video repository includes the MPEG-4/H.264 video, the original collected author meta-data, and speech to text. The system returns a ranked list of the top 100 videos most likely to match the topics. The system is evaluated using the mean inverted rank (IR) for the 361 known item topics and ground truth results of the TRECVid task. Table 6 shows an example of how the best field match for a topic can be found in different meta-data fields.

Table 5: Example topics from TRECVid 2012 KIS

Topic	Description	Video
893	Find the video of man speaking German with long hair and green jacket and soccer ball in a parking lot.	
909	Find the video of woman pouring black oil from milk carton.	
968	Find the video of three men, one with spiked white hair and black and red vest.	
1035	Find the video with a lake and its shores.	

4.1 Analysis

Baseline experiments are conducted using a text only information retrieval approach [20], where the text modalities are merged. The meta-data, ASR, and OCR from the repository are used to generate a video document that can be indexed and retrieved using state-of-the-art retrieval algorithms. The results provide both a baseline comparison and a set of topic-video features used by our rank learning algorithm. In this experiment, we merge all of the video text fields into a single document for indexing and retrieval. Both the topics and video documents are preprocessed for stop word removal, word stemming, and spell correction. The results in Table 7 show the IR for the 361 topics using three different retrieval models. The IR is calculated at five different ranking points, starting at the top returned document and ending with document 100. Results show that the BM25 model produces the top IR scores at each of the ranking points for the combined video documents.

Table 6: Example topics from TRECVid 2012 KIS

Topic	Description	Best Found in
895	Find the video titled "Sunday Quickie" of a man who is wearing glasses and a blue shirt standing by the window and watching the rain outside and discussing his trip to Home Depot and Harveys Hamburger Kiosk.	Meta Title
1002	Find the video of man demonstrating use of children's laptop.	OCR
1051	Find the video of a shirtless boy playing with toy helicopter, gun and soldiers.	Meta Desc
1115	Find the video of a close-up face shot of a man wearing dark glasses and a white shirt who is giving a satire X-lawyer advertisement.	FileName
1167	Find the video titled "Welcome to Best Bible Study on Earth" which starts with a picture showing the mountains, lake, and sky and then a map of the United States where the narrator solicits you to go to their website.	ASR

Table 7: Mean Inverted Rank by Model

Model	IR@1	IR@3	IR@5	IR@10	IR@100
TFIDF	0.249	0.296	0.303	0.310	0.316
BM25	0.288	0.315	0.327	0.333	0.339
LM	0.282	0.314	0.323	0.327	0.334

Table 8 provides further analysis of our baseline text only retrieval models. This table shows a breakdown of the total documents found at each of the ranking points. The results show that BM25 outperformed the TFIDF model at rank 1 (IR@1) by 14 KIS videos. It is also interesting to note that while 82 videos are identified after rank 5, they increase the final IR by only .012.

Table 8: Count of Topics Found by Model

Model	Ct@1	Ct@3	Ct@5	Ct@10	Ct@100
TFIDF	90	129	140	158	218
BM25	104	126	143	160	225
LM	102	129	143	154	221

The next set of experiments follow the information retrieval approach, but are performed on each of the text modalities. These experiments provided modality specific topic-video features for the rank learning approach and help to identify the contribution of each of the text modalities. Tables 9 and 10 show the IR and count found for the 361 topics, using a BM25 retrieval model for the seven text modalities of our video. The IR and count are calculated at five different ranking points, starting at the top returned document and ending with document 100.

The meta-data description provided the highest IR and count found. This modality is provide by the author and contains the least noise and most detailed description of the video. The results also show that the meta-data Title and FileName identify a large number of relevant videos. Authors often include key terms in these fields that summarize the content of the video. ASR and OCR results suffer from noise generated during the translation from speech and video into text. However, Figure 1 shows that the ASR is comparable to the meta-data Description for the number of unique ground truth videos identified.

Table 9: Mean Inverted Rank by Field Type

Field	IR@1	IR@3	IR@5	IR@10	IR@100
ASR	0.085	0.098	0.099	0.103	0.105
OCR	0.074	0.087	0.088	0.091	0.093
File	0.105	0.128	0.136	0.141	0.146
Title	0.149	0.173	0.180	0.183	0.188
Desc	0.171	0.191	0.198	0.204	0.210
Keyword	0.002	0.004	0.005	0.005	0.006
Subject	0.080	0.102	0.106	0.110	0.113

Table 10: Count of Topics Found by Field Type

Field	Ct@1	Ct@3	Ct@5	Ct@10	Ct@100
ASR	31	42	43	53	78
OCR	27	37	38	47	66
File	38	57	69	83	122
Title	54	74	85	94	142
Desc	62	79	91	109	162
Keyword	1	2	4	5	7
Subject	29	46	54	65	92

4.2 Rank Learning

The Rank Learning experiments use a 10-fold cross validation of the KIS topic set where each fold is divided into train, validate, and test. A learning feature vector is constructed for each video and consists of the topic, topic-video, and video features. The scores for the topic-video features are derived from the models and fields described in the information retrieval analysis experiments. A keyframe is extracted every two seconds from each video in the collection to derive the set of visual features. OCR text for a video consists of the concatenated text extracted from each image frame. Edge, color, and local keypoint features are also extracted at every frame.

The gradient boosted regression trees used for ranking are trained using a cross-entropy cost function. To avoid model over-fitting, the number of trees is controlled by monitoring the prediction error on the validation set. Models were iteratively trained to a maximum of 1000 trees and results show that 100 tree provided good accuracy on the validation set. The maximum number of leaves per tree and learning rate were also used to control over-fitting by monitoring the validation set. The reported results use a maximum of 5 leaves per tree and learning rate of .05.

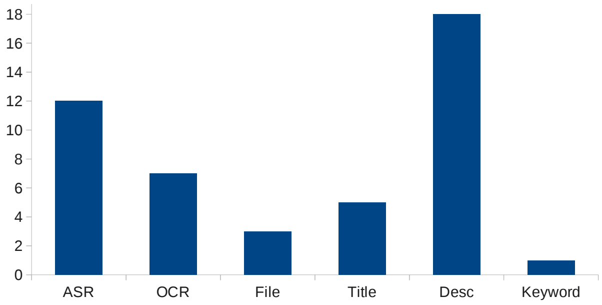


Figure 1: Unique Count of Topics Found by Field

Tables 11 and 12 provide a comparison of the baseline and rank learning models. The results are reported at five ranking points for mean inverted rank and total count found. The four baseline models provides a comparison of text only retrieval models. The baseline meta-data model combines the video text fields of filename, title, description, subject, and keywords. The top baseline model is the BM25 using the combined text document.

The first rank learning model (RankLearn1) uses the derived feature space and gradient boosted regression trees to learn a model. This model does not use pseudo-positive examples and includes a single truth video and 99 negative videos for each training topic. We see from the results that the lack of positive examples causes the mean inverted rank and count found to drop below the top baseline run.

The next two runs (RankLearn1 and RankLearn2) extend the RankLearn1 model with pseudo positive examples, derived from the K-Step Markov graph ranking approach. The graph is iteratively constructed by selecting the KIS truth video as the root and performing a similarity search in the video collection using each of the seven text and the three visual modalities. RankLearn2 allows max path lengths of 2 on the graph and identifies approximately 5 pseudo positive examples per topic. RankLearn3 expands the max path length to three and identifies approximately 9 pseudo positive examples per topic. Pseudo positive examples are assigned a graded relevance from 1 to 4 using the importance measure I assigned by the K-Step Markov approach. Our reported results assign relevance 4 for ground truth, 3 for $I >= .05$, 2 for $.05 > I >= .01$, and 1 for $.01 > I >= .001$. These values were determined using the validation set.

The results for RankLearn3 show an increase of 21 topics found at rank position 1 over the TFIDF baseline and an increase of 7 topics found over the BM25 baseline. These results show that the rank learning models are able to boost additional positive KIS examples higher in the ranked result list. Our top mean inverted rank of 0.356 is competitive with the top TRECVid 2012 task participants where the top reported score was 0.419.

5. CONCLUSIONS

In this paper, we investigated a semi-supervised rank learning approach to the multimedia KIS problem. We constructed a feature space consisting of topic specific, topic-video dependent, and video specific features, calculated from the meta-data, speech, and visual modalities of our text topics and video repository. Pseudo-positive KIS examples are identified in a similarity graph, using a K-Step Markov to estimate the importance of nodes relative to the truth root.

Pairwise rank learning, using gradient boosted regression trees, are applied to the KIS problem to improve ranking results. Our results show that combining pseudo positive training example with the rank learning framework, improves Known Item Search ranking at all ranking points. Future work will examine approaches for enriching the topic-video dependent feature space with features derived from the visual modalities.

6. REFERENCES

[1] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008.

[2] C. Burges. From ranknet to lambdarank to lambdamart: An overview. *Microsoft Research Technical Report*, MSR-TR-2010-82, 2010.

[3] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136. ACM, 2007.

[4] L. Chaisorn, Y.-T. Zheng, and K. Sim. Known-item search (kis) in video: Survey, experience and trend. In *Information, Communications and Signal Processing (ICICS) 2011 8th International Conference on*, pages 1–4. IEEE, 2011.

[5] O. Chapelle and Y. Chang. Yahoo! learning to rank challenge overview. *Journal of Machine Learning Research-Proceedings Track*, 14:1–24, 2011.

[6] S. Chen, K. McGuinness, R. Aly, N. E. O'Connor, and F. de Jong. The axes-lite video search engine. In *Image Analysis for Multimedia Interactive Services (WIAMIS), 2012 13th International Workshop on*, pages 1–4. IEEE, 2012.

[7] N. Craswell, A. P. de Vries, and I. Soboroff. Overview of the trec-2005 enterprise track. In *TREC 2005 conference notebook*, pages 199–205, 2005.

[8] D. Etter, F. Ferraro, R. Cotterell, O. Buzek, and B. Van Durme. Nerit: Named entity recognition for informal text. *Human Language Technology Center of Excellence, Johns Hopkins*, Technical Report 11, 2013.

[9] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.

[10] J. H. Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, 2002.

[11] Y. Ganjisaffar, R. Caruana, and C. V. Lopes. Bagging gradient-boosted trees for high precision, low variance ranking models. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 85–94. ACM, 2011.

[12] J. Gauvain, L. Lamel, and G. Adda. The limsi broadcast news transcription system. In *Speech Communication*, pages 37(1–2):89–108, 2002.

[13] L. S. Kennedy and S.-F. Chang. A reranking approach for context-based concept fusion in video indexing and retrieval. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 333–340. ACM, 2007.

Table 11: Rank Learning Mean Inverted Rank

Model	Desc	IR@1	IR@3	IR@5	IR@10	IR@100
Baseline1	ASR	0.085	0.098	0.099	0.103	0.105
Baseline2	Meta	0.241	0.276	0.285	0.292	0.299
Baseline3	TFIDF on all	0.249	0.296	0.303	0.310	0.316
Baseline4	BM25 on All	0.288	0.315	0.327	0.333	0.339
RankLean1	Rank Learn Binary	0.279	0.309	0.319	0.324	0.331
RankLean2	Rank Learn Pseudo1	0.296	0.322	0.332	0.338	0.345
RankLean3	Rank Learn Pseudo2	0.307	0.336	0.342	0.348	0.356

Table 12: Rank Learning Count Found

Model	Desc	Ct@1	Ct@3	Ct@5	Ct@10	Ct@100
Baseline1	ASR	31	42	43	53	78
Baseline2	Meta	87	115	130	150	203
Baseline3	TFIDF on all	90	129	140	158	218
Baseline4	BM25 on All	104	126	143	160	225
RankLean1	Rank Learn Binary	101	126	141	155	225
RankLean2	Rank Learn Pseudo1	107	128	143	161	225
RankLean3	Rank Learn Pseudo2	111	135	145	162	225

[14] J. Kim and W. B. Croft. Ranking using multiple document types in desktop search. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 2010.

[15] C.-J. Lee, W. B. Croft, and J. Y. Kim. Evaluating search in personal social media collections. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 683–692. ACM, 2012.

[16] J. Liu, W. Lai, X.-S. Hua, Y. Huang, and S. Li. Video search re-ranking via multi-graph propagation. In *Proceedings of the 15th international conference on Multimedia*, pages 208–217. ACM, 2007.

[17] T.-Y. Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.

[18] C. Macdonald, R. L. Santos, and I. Ounis. On the usefulness of query features for learning to rank. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2559–2562. ACM, 2012.

[19] D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[20] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*, 2006.

[21] D. K. Park, Y. S. Jeon, and C. S. Won. Efficient use of local edge histogram descriptor. In *Proceedings of the 2000 ACM workshops on Multimedia*, pages 51–54. ACM, 2000.

[22] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281. ACM, 1998.

[23] Y. Press. Youtube statistics. Accessed: 2013-08-01.

[24] T. Qin, T.-Y. Liu, J. Xu, and H. Li. Letor: A benchmark collection for research on learning to rank for information retrieval. *Information Retrieval*, 13(4):346–374, 2010.

[25] S. E. Robertson and K. S. Jones. Relevance weighting of search terms. *Journal of the American Society for Information science*, 27(3):129–146, 1976.

[26] B. Safadi and G. Quénot. Re-ranking by local re-scoring for video indexing and retrieval. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 2081–2084. ACM, 2011.

[27] P. Salembier. Overview of the mpeg-7 standard and of future challenges for visual information analysis. *EURASIP Journal on Applied Signal Processing*, 2002(1):343–353, 2002.

[28] G. Salton, A. Wong, and C.-S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[29] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[30] S. White and P. Smyth. Algorithms for estimating relative importance in networks. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 266–275. ACM, 2003.

[31] Q. Wu, C. J. Burges, K. M. Svore, and J. Gao. Ranking, boosting, and model adaptation. *Tecnical Report, MSR-TR-2008-109*, 2008.

[32] J. Yu, Y. Rui, and B. Chen. Exploiting click constraints and multi-view features for image re-ranking. *Multimedia, IEEE Transactions on*, 16(1):159–168, Jan 2014.