SemRank: Semantic Rank Learning for Multimedia Retrieval

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

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

***Abstract*—Multimedia retrieval suffers from the lack of common feature representation between a text based query and the visual content of a video repository. One approach to bridging this representation gap is known as query-by-concept, where a query and video are mapped into a common semantic feature space. One of the challenges with using semantic concepts for multimedia retrieval, is that the available vocabulary size is generally not sufficient for representing the content of the query and video. In addition, the lack of training data and visual feature representation often leads to low precision models. In this work, we explore the use of a query-by-concept approach for the multimedia Known Item Search (KIS) problem. We propose a semantic rank learning model, called SemRank, to overcome the challenges of the vocabulary size and lack of training data. First, we construct a semantic fusion model to combine the output from many noisy classifiers. Next, we train a gradient boosted regression tree model, using a semantic feature space derived from the query, video, and query-video similarity. Our approach is evaluated over a large internet video repository, and the results show that query-by-concept can be an effective model for multimedia KIS.**

I. Introduction

The ability to collect and share video has shifted the information retrieval task from a text based problem to a multimedia problem. Users expect that a query entered into their favorite search engine will not only return documents and web pages, but also images and video. This presents a challenging research problem, since multimedia data can include text, speech, and visual modalities.

One of the goals of multimedia retrieval is to find a common representation between a text user query and the multiple modalities of multimedia data. The use of semantic concepts is one approach to overcoming this representation gap. Semantic concepts provides a high level feature representation for objects, such as Snow or Person. Low level features from image, speech, and text are then mapped into this common feature space. Table I provides an example of the semantic concepts used in the TRECVid evaluation [1].

The semantic concept representation has proven to be an effective approach for classification and indexing tasks [1]. Research in the retrieval community has included semantic concept for the search task, but generally as an additional feature to the metadata that continues to dominate retrieval performance.

One of the problems with using semantic concepts for retrieval is the size of the available concept vocabulary. Obtaining labeled data, extracting multimodal features, and training classifiers for each semantic concept is a difficult and time consuming task. Retrieval over an internet size repository requires a large number of concepts to accurately reflect the content and identify an item at query time.

TABLE I
Semantic Concepts

Name	Definition
Airplane	Shots of an airplane.
Person	Shots depicting a person (the face may or may not be visible).
Prisoner	Shots depicting a captive person, e.g., imprisoned, behind bars, in jail or in handcuffs, etc.
Snow	Snow falling or already accumulated on the ground.
Suburban	Shots depicting an urban or suburban setting.

The reliability of the labels produced by concept classifiers, presents a second problem for multimedia retrieval. Many of these models suffer from low precision due to a lack of training data and visual feature representation.

In this work, we explore a query-by-concept approach [2] for multimedia Known Item Search (KIS). KIS is a retrieval task where the user attempts to find a previously seen video. The goal of a KIS system is to return the known item at the top of a ranked result set. Both the number and quality of concept labels influences the KIS systems ability to identify the known item.

In order to overcome the vocabulary problem, we model semantic multimedia retrieval as a rank learning problem [3]. Our semantic rank learning model, called SemRank, is a supervised learning approach that builds on the traditional unsupervised information retrieval models by including features from both the queries and videos. Our approach derives a semantic feature space and trains a semantic ranking model using gradient boosted regression trees [4]. To improve the quality of concept labels, we propose a semantic fusion model to combine the labels from many weak classifiers. These classifiers represent different systems trained on varying data and modalities for a given concept.

We evaluate our semantic retrieval approach using a set of KIS queries over a large internet video collection. In the KIS retrieval scenario, the user creates a query based on what he remembers and the system returns a ranked list of videos most likely to match the request. Table II provides an example of KIS queries from the TRECVid evaluation [1]. This unique multimedia retrieval problem provides a good test environment for our semantic retrieval approach, since we are attempting to identify a single correct answer from a large collection.

Our contributions to the semantic computing community include the following:

1) We introduce a query-by-concept approach for multimedia Known Item Search (KIS)
2) We construct a semantic fusion model, to fuse labels from noisy concept classifiers and improve retrieval performance.
3) Finally, we propose a semantic rank learning model, called SemRank, and show improved ranking over traditional information retrieval models.

II. Related Work

Multimedia retrieval is influenced by a number of research areas including information retrieval [5] , rank learning [3] [6], semantic indexing, multiview learning [7], and known item search [1].

Rank learning is a supervised learning approach introduced by the information retrieval community [8] [3] that constructs ranking models using features derived from queries [9], documents, and their similarities. This approach incorporates similarity scores from unsupervised information retrieval models, such as language models (LM) [10] and probabilistic models [5] [11]. LETOR [12] is a large collection of publicly available datasets for evaluating learning to rank algorithms. Our approach extends the rank learning approach to semantic multimedia retrieval. We construct a rank learning model using gradient boosted regression trees [13] [14] [4] trained over a semantic concept feature space.

The Large Scale Concept Ontology for Multimedia (LSCOM) [15] was a collaborative effort among researchers to develop a standard set of semantic concepts. The work produced a set of 1000 semantic concepts that were used to describe a large collection of broadcast news video. A collaborative annotation on the TRECVid [1] collection is described in [16] [17]. The work of [16] used an active learning approach to filter the candidate set of video frames for annotation. A semi-automatic annotation approach is used in [17], where the system suggests concepts for a video frame based on concept dependencies.

Both [18] and [19] study the type of visual features and learning algorithms that optimize a semantic classifier. The work of [18] studies the size of the visual vocabulary [20] feature size required for a semantic classifier. Using a supervised learning approach, they found that a vocabulary size of 1024 to 4096 performed the best on a large video repository.

An overview of video search techniques is given in [2]. They divide approaches to video search into three categories. The first category is query-by-keyword, where traditional information retrieval models are used to match a text query with the video metadata. This approach generally ignores the image content of the video. The second category is query-by-example, which includes sample images as part of the query. This approach uses both text and visual modalities, but forces a user to provide sample images. The final category is query-by-concept, where both the video and query are mapped into a semantic concept feature space. Semantic concept query expansion is studied in [21]. They propose both a rule-based and statistical query expansion approach to identify concepts in a text query. The identified concepts are used to rerank the initial result set. Semantic search is used in the context of event detection in the work of [22]. Their approach maps text queries to concepts using a text language model constructed for each concept. The language model uses a set of documents, retrieved from a web search, to identify words related to the semantic concept.

Known item search is a multimedia retrieval task where there exists a single correct result. The KIS task models a scenario where a user has previously seen a video and would like to find it again in a large video collection. This task has most notably been studied as part of the TRECVid evaluations [1]. The work of [23] studied the KIS problem in the context of semi-supervised learning. Their approach derived additional training instances using a graph based algorithm to identify videos that were similar to the known item. A survey of current KIS work is found in [24]. They found that most approaches attempted to use traditional information retrieval approaches in combination with multimodal fusion. We evaluate our system using this large internet video collection and set of queries and truth videos.

III. Our Approach

We propose a rank learning approach [3] to semantic concept based multimedia retrieval. Given a set of weak concept labels from multiple classifiers, a semantic fusion approach is used to derive a final semantic labeling. Queries to the retrieval system are provided in the form of a concept set and matched to the video semantic labels using a probabilistic model [11]. Rank learning improves on traditional information retrieval models by considering features of both the video and query, in addition to their feature similarity.

The semantic concept space is defined as $C = (c_1, \ldots, c_d)$, where d is the dimensionality of the concept space. Given a query set $Q = (q_1, \ldots, q_n)$ of size n, a mapping to the semantic concept space is defined as $F_q(q_i, C) = (q^i_{c_1}, \ldots, q^i_{c_d})$, where $q^i_{c_x}$ is the score for concept c_x in query q_i. A similar mapping is defined for the video repository $V = (v_1, \ldots, v_m)$ of size m, such that $F_v(v_y, C) = (v^y_{c_1}, \ldots, v^y_{c_d})$, where $v^y_{c_x}$ is a score representing concept c_x in video v_y. The multimedia retrieval problem is then defined as a ranking $R(q_i, V) = (r^i_{v_1}, \ldots, r^i_{v_m})$, where V is the video repository, q_i is a video query, and $r^i_{v_t}$ is the rank of video v_t for the given query.

TABLE II
KNOWN ITEM SEARCH QUERIES

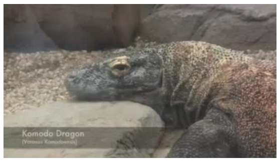
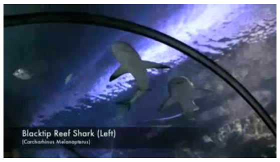
Query ID	Description	Video Shots
912	Find the video of bathroom with brown walls, checked curtains and picture of camel on wall.	
961	Find the video set in a book shop. A dark-haired woman reads aloud from a book, which has a red cover with a heart on it.	
1042	Find the video showing a Komodo Dragon and a Lionfish.	 

Figure 1 provides an overview of the SemRank retrieval model. The video repository consists of a collection of internet videos and their associated metadata [1]. The metadata can include multiple modalities such as speech to text, optical character recognition, image features, and descriptive text provided by the video author. The repository is used as input to systems which provide semantic concept classifiers. Each system is trained using one or more modalities using both in-domain and out-of-domain data [1]. Given a video from the repository, a set of concept labels and scores are provided from each of the systems. Next, we train a semantic fusion classifier using all of the system output labels and derive a final concept set for the semantic repository.

Given a set of training queries, we construct a rank learning feature space consisting of semantic features from the query, video, and their initial similarity ranking. The resulting semantic ranking model is applied to the test query set which generates a final ranking.

A. Semantic Concept Fusion

Multimedia retrieval in a semantic concept space, presents a challenge both in the quality of concept classifiers and in the number of concept labels required to provide coverage for a generic retrieval task. It is difficult to obtain labeled data for training a large number of classifiers and extracting features from a large video repository is resource intensive. Our approach fuses the output labels from different systems to provide a higher quality set of semantic concept labels.

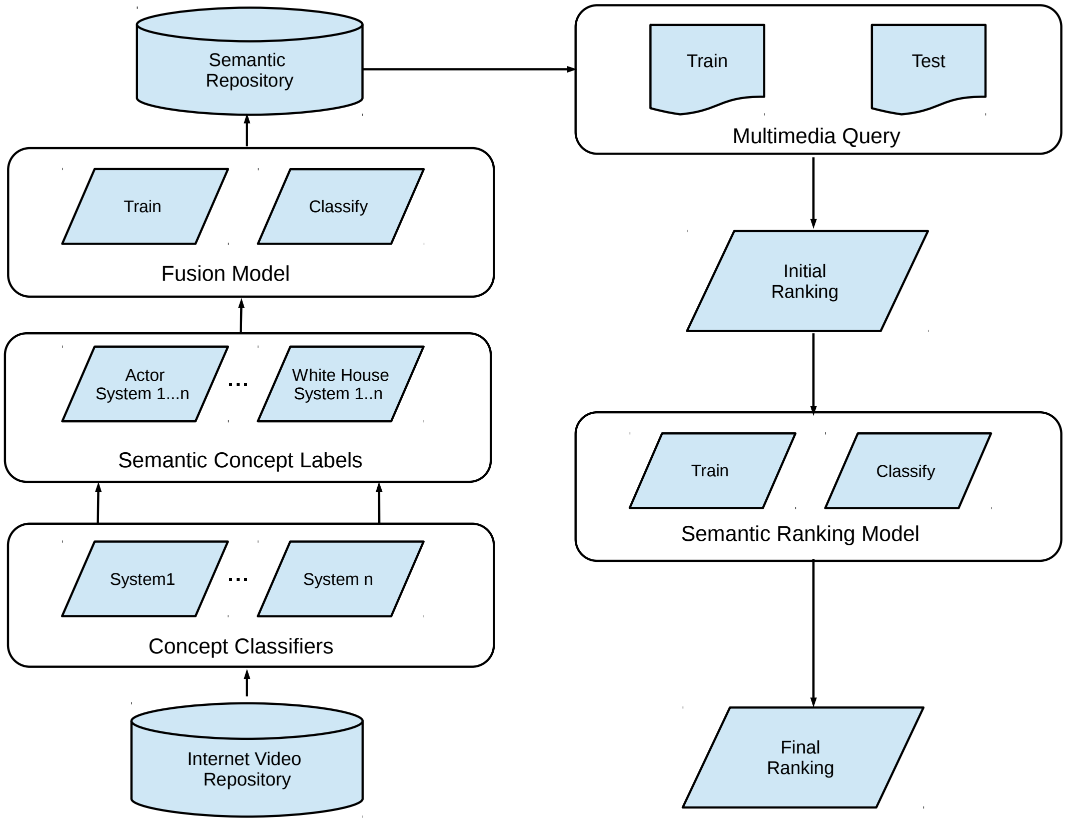
The video repository is populated with short internet videos and metadata. These multimedia objects includes modalities such as image, speech, and text, which are used to derive the input feature space for the concept classifiers. Table III provides an example video from our internet repository. The example includes speech to text, image frames, optical character recognition, and metadata provided by the video author.

We are given the semantic concept output from a number of different systems. These systems are trained with different modalities and a variety of supervised and semi-supervised learning algorithms [1]. Some of the classifiers were trained on text only while others attempted to combine features from all of the multimedia modalities. The training sets for the different systems also varied with some systems using only in-domain, while others attempted to include out-of-domain data. The output from each system is a ranked list of the videos most likely to contain the given concept. For any given video we are provided with a different set of semantic labels for each of the n systems. Table IV shows an example of the different system output for a given video.

A set of classifier systems is defined as $S = (s_1, \ldots, s_n)$, where n is the number of systems providing concept labels. For a given concept c, a system s produces a ranking $R(s_c, V) = (r^c_{v_1}, \ldots, r^c_{v_m})$, where V is the video repository, s_c is the system model, and $r^c_{v_t}$ is the rank of video v_t. We define the semantic mapping of video v_t and concept c_x for all systems S, as $F_s(v_t, c_x, S) = ((r^{c_x}_{v_t})_{s_1}, \ldots, (r^{c_x}_{v_t})_{s_n})$.

A Support Vector Machine [25] is used to construct a fusion model for each semantic concept in C. Our approach fuses the output from the different systems into a single semantic set. The input to the SVM for a concept c_x, video v_t, and label l, is of the form $(((r^{c_1}_{v_t}, \ldots, r^{c_d}_{v_t})_{s_1}, \ldots, (r^{c_1}_{v_t}, \ldots, r^{c_d}_{v_t})_{s_n}), l)$. This approach considers the output of all concept classifiers

Fig. 1. SemRank: Semantic Rank Learning



C and all systems S when constructing a fusion model for the concept c_x. This allows the fusion model to learn relations between concepts and consider their correlations during model creation. As an example, when training a fusion model for the concept Outdoor, input to the learning algorithm includes the features from the concepts Ocean, Lake, and Mountain. The output from the semantic fusion model is used as input to the semantic rank learning model.

B. SemRank

Recent work in information retrieval has shown that supervised rank learning [3] can improve ranking results over traditional unsupervised models. Supervised ranking models not only consider similarity scores, but also incorporates features derived from both the query and document. Given the limited vocabulary available from our semantic concept labels, we believe that a rank learning approach could improve initial ranking by incorporating the semantic features from our queries and videos. As an example, a ranking model may weight a Language Model similarity score higher for a video containing one set of semantic concepts, but choose a Probabilistic score for a different set. The ranking model also has the ability to use the semantic concepts of the query to identify query classes and produce different feature weights for each class.

We define a ranking feature space for the video v_t and query q_i using the triple $(s^{v_t}_{q_i}, o^{q_i}, o^{v_t})$, where $s^{v_t}_{q_i}$ are similarity features, o^{q_i} are query features, and o^{v_t} are video features. Table V shows our derived semantic feature space. Similarity features are the traditional similarity scores from unsupervised retrieval models such as Term Frequency Inverse Document Frequency (TF-IDF), Language Model [10], and Probabilistic Model [11]. These features consider the interaction between a query-video pair. In addition to the scores, this class of features includes the percentage of match concepts, the total concept frequency (CF) of all matches, and the total video frequency (VF) of all matches. The difference between CF and VF is that a concept classifier produces output for every video shot boundary. This means that a given video may include multiple positive labels.

Video features are derived from the given video and are associated with the semantic concept labels from our semantic fusion model. This feature set includes total concept count, frame count, and the unique concept count. The video class also includes the bag-of-concepts, which is the semantic concept equivalent of the bag-of-words used in natural language

TABLE III
VIDEO WITH METADATA

Video Shots	SNOW STORM DEC 15 Meteorological Alert WITH HEAVY SNOW AND BLOWING SNOW AS WELL AS SOME ICE PELLETS TONIGHT AND SUNDAY. THIS IS A WARNING THAT DANGEROUS WINTER WEATHER CONDITIONS ARE IMMINENT OR OCCURRING IN THESE REGIONS. Saturday Night Morning
FileName	MMMMMoon-WinterStormDec152007277-3.
ASR	And get them in and move around the best through it seems you know you're in the home and bring against them . Yeah it is out of work...
OCR	uogi ralo Alert HEAVY suaw AND amwmzs snow wan arm ICE...
Title	Winter Storm Dec. 15, 2007.
Description	The day that was. Winter storm Watch.... at Mobile Station 1, via our on site reporter.

TABLE IV
SEMANTIC CONCEPT LABELS

Shot	**Semantic Labels**
	Outdoor, Plant, Road, Sky, Vegetation, ...
	Airplane, Birds, Boat Ship, Car Racing, Daytime Outdoor, Dogs, Military, Sky ...
	Weather Security Checkpoint, Sun ...

TABLE V
SEMANTIC FEATURE SPACE

Type	**Feature**
Similarity	TFIDF Model Score Language Model Score Probabilistic Model Score Perc of concept match Total concept freq of match Total video freq of match
Query	Bag of concepts Concept count All match count
Video	Bag of concepts Concept count Shot count Shots with concepts Unique concept count Multi shot concepts

processing. Query features include bag-of-concepts, concept count, and the count of videos matching all query concepts.

Our semantic ranking approach uses gradient boosted regression trees [13] [14] [4] to model this unique semantic feature space. The algorithm follows a step-wise approach to construct a series of N weak regression tree models, $M(s^v_q, o^q, o^v)$, over the feature vector tuple (s^v_q, o^q, o^v). At each step, i, a weight value, β_i, is learned for the given model. The final ranking is generated by combining the output, $\sum_{i=1}^{N} \beta_i \times M_i(s^v_q, o^q, o^v)$, from each of the weak models.

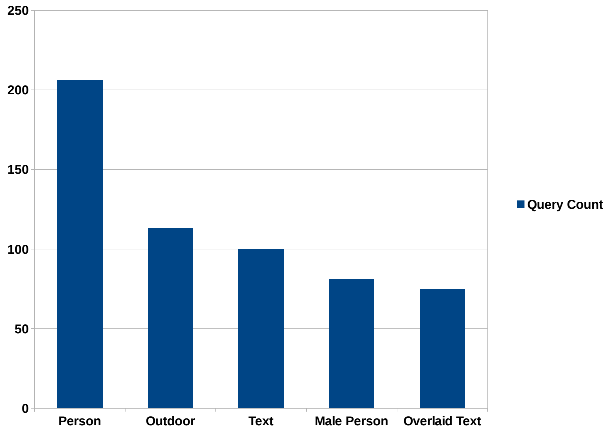
IV. EXPERIMENTS

We evaluate our approach using a large internet collection from the KIS and Semantic Indexing tasks of the 2012 TRECVid evaluation [1]. The IACC.1 multimedia test collection consists of approximately 8000 internet videos, available from the Internet Archive under a Creative Commons licenses. The collection includes a diverse set of content from both professional and home videos, with a duration between 10 seconds and 3.5 minutes. Many of the videos include metadata content in the form of title, keywords, and a description. Speech-to-text is also made available as part of the evaluation.

Queries are derived from the TRECVid KIS topic set, which is based on query-by-keyword [2], where the query is presented as a text description. Our experiments follow a query-by-concept model and use a set of semantic concepts to describe the known item query. The semantic labels for each query is derived from the collaborative truth annotation task described in [16] [17]. We drop any query which does not have a concept label and evaluate our system with the remaining 328 query-by-concept topics.

Figure 2 shows the top 5 semantic concepts found in our query-by-concept set. The query test set includes 306 unique semantic concepts, where the majority of the queries use 10 or fewer concepts. Our analysis of the set identified two categories of concepts. The first category consists of approximately 100 concepts, which occurred in 10 or more queries. This general set includes concepts such Person, OutDoor, and Overlaid Text. These concept provide a general filter to identify candidate videos. The second category consists of the approximately 100 concepts that were used in 3 or fewer queries, with about 50 occurring in only 1 query. This low frequency concept category allows the system to identify the unique properties of a known item video. Examples from the low frequency category include Tent, Stadium, Skier, John Kerry, and First Lady.

Fig. 2. Top Query Semantic Concepts



Our experiments use the semantic label output from approximately 50 different systems [1]. These systems vary in the domain of data used for training and the types of training algorithms. Many of the systems were trained with only IACC training data, while others include features from external sources such as internet search results. Each system produces labels for up to 346 semantic concepts and returns a ranked list of up to 2000 video shots most likely to match a given concept.

To measure the performance of the system, we calculate the inverted rank of the truth video for a given query. The count of total KIS found and the mean inverted rank (MIR) are used to measure the retrieval performance over the entire query set.

Our baseline experiment follows an unsupervised information retrieval approach. We construct the semantic concept set for a given video using the output from the 50 classifier systems. A classifier labels video at the shot level, which can result in multiple labels for a given video. We capture the ranking from each system and create an aggregate across systems using the count, best, worst, and average ranks. The semantic feature vector for the baseline approach is constructed by binning the average rank for each concept. The bins allow us to create a weighted bag-of-concept words for indexing and retrieval. The query feature space uses a similar bag-of-concept words approach. A probabilistic information retrieval model is used to compare and rank the query-video pairs.

The baseline results in table VI, show that this unsupervised approach to query-by-concept is able to identify 78 of the known items within the top 100 ranked videos. The results also show that 18 known items are ranked in the top 10 returned videos. These results are promising for our overall goal of showing that at query-by-concept approach can be successful for known item search. We were able to show that a simple retrieval model, using noisy concept labels, was able to retrieve approximately 80 of the known items.

TABLE VI
BASELINE RESULTS

Rank	MIR	KIS Found
@001	0.0091	003
@003	0.0137	007
@005	0.0160	010
@010	0.0192	018
@020	0.0217	030
@050	0.0239	054
@100	0.0250	078

Our next experiment builds on the baseline results using our semantic rank learning approach. We construct a semantic feature space using query-video similarity scores, query concepts, and video concepts. The similarity scores are derived from the ranked results of our unsupervised information retrieval models.

We obtain the query-video similarity scores using the output ranks from tf-idf, language [10], and probabilistic [11] models. For each of these models we derive the percentage of concept matches, the total concept frequency of the matches, and the total video frequency of the matches. The query features are constructed using the bag-of-concepts, concept count, and the number of videos that matched all concepts in the query.

The video feature space includes: bag-of-concepts, concept count, number of shots in the video, count of shots with at least one concept, number of concepts occurring in single shot, and the number of concepts occurring in more than one shot. The weight for the bag-of-concepts in video feature space is derived from the average rank, as described in our baseline experiment. The gradient boosted regression tree model is trained using a 10-fold cross validation of the query set. The model uses 100 trees and is trained with a learning rate of .05. The maximum number of leaves for each tree is set to 5.

The results in VII, show an improvement in count found and mean average rank over the unsupervised retrieval model. The total number of known items found, increased to 96, with

20 found in the top 10 returned results. These results show that the semantic ranking model is able to learn from our rich semantic feature space.

TABLE VII
BASELINE SEMANTIC RANKING RESULTS

Rank	MIR	KIS Found
@001	0.0152	005
@003	0.0183	007
@005	0.0201	010
@010	0.0246	020
@020	0.0276	035
@050	0.0309	069
@100	0.0320	096

The baseline fusion experiment uses a supervised learning approach to derive labels for each video concept. A semantic fusion model is constructed for each semantic concept, using a Support Vector Machine [25]. The feature space is derived from the labels of all systems, for the given video. We train the models using a 10-fold cross validation of the video collection. The labels from the semantic models are used to construct the video bag-of-concepts, which are then used in a probabilistic retrieval model.

Our semantic fusion model, shown in table VIII, improves over the baseline information retrieval model with increases in total known items found and mean inverted rank. The results show that the semantic fusion model is able to fuse the outputs from our noisy classifiers and improve retrieval results.

TABLE VIII
SEMANTIC FUSION RESULTS

Rank	MIR	KIS Found
@001	0.0274	009
@003	0.0335	013
@005	0.0349	015
@010	0.0414	031
@020	0.0446	047
@050	0.0477	077
@100	0.0488	104

The final experiment applies our semantic rank learning model to the semantic fusion results. The semantic feature space is constructed in a similar manner to the baseline ranking model. The primary difference is that the similarity features are derived from the fusion results. We maintain the gradient boost regression tree parameters and perform a 10-fold cross validation.

The final results, shown in table IX, provide our best scores for known items retrieved and mean inverted rank. These results show that a query-by-concept model is an effective approach for multimedia KIS. The results also show that our semantic fusion model was able to improve the performance of the noisy concept labels. The semantic rank learning model improved the initial ranking results using the feature space derived from the query, video, and query-video similarity.

TABLE IX
SEMRANK RESULTS

Rank	MIR	KIS Found
@001	0.0366	012
@003	0.0478	021
@005	0.0540	030
@010	0.0595	043
@020	0.0639	064
@050	0.0663	088
@100	0.0675	118

V. CONCLUSIONS

We evaluated a query-by-concept approach for the multimedia KIS problem. Our semantic fusion model was able to fuse noisy concept labels and improve retrieval performance. We constructed a semantic rank learning model using gradient boosted regression trees. The model uses a semantic feature space derived from the query, video, and query-video similarity. The final results, evaluated over a large internet video repository, show that query-by-concept is an effective approach for multimedia KIS.

This work focused on query-by-concept from the viewpoint of the video. Queries consisted of a set of semantic concepts, identified by a user, describing the known item. In the future, we plan to explore the problem of automatic query concept selection. This approach would allow the user to create a simple text query that is automatically mapped into a semantic concept feature space.

REFERENCES

[1] 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*. New York, NY, USA: ACM Press, 2006, pp. 321–330.

[2] C. G. Snoek and M. Worring, "Concept-based video retrieval," *Foundations and Trends in Information Retrieval*, vol. 2, no. 4, pp. 215–322, 2008.

[3] T.-Y. Liu, "Learning to rank for information retrieval," *Foundations and Trends in Information Retrieval*, vol. 3, no. 3, pp. 225–331, 2009.

[4] O. Chapelle and Y. Chang, "Yahoo! learning to rank challenge overview." in *Journal of Machine Learning Research, Workshop and Conference Proceedings*, 2011, pp. 14:1–24.

[5] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

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

[7] N. Chen, J. Zhu, and E. P. Xing, "Predictive subspace learning for multi-view data: a large margin approach," in *Advances in neural information processing systems*, 2010, pp. 361–369.

[8] 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*. ACM, 2007, pp. 129–136.

[9] 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*. ACM, 2012, pp. 2559–2562.

[10] 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*. ACM, 1998, pp. 275–281.

[11] S. E. Robertson and K. S. Jones, “Relevance weighting of search terms,” *Journal of the American Society for Information science*, vol. 27, no. 3, pp. 129–146, 1976.

[12] 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*, vol. 13, no. 4, pp. 346–374, 2010.

[13] J. H. Friedman, “Greedy function approximation: a gradient boosting machine,” *Annals of Statistics*, pp. 1189–1232, 2001.

[14] Y. Ganjisaffar, R. Caruana, and C. 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*, ser. SIGIR ’11. New York, NY, USA: ACM, 2011, pp. 85–94.

[15] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, “Large-scale concept ontology for multimedia,” *IEEE MultiMedia*, vol. 13, no. 3, pp. 86–91, 2006.

[16] S. Ayache and G. Quénot, “Video corpus annotation using active learning,” in *Advances in Information Retrieval*. Springer, 2008, pp. 187–198.

[17] M. Hradiš, M. Kolář, A. Láník, J. Král, P. Zemčík, and P. Smrž, “Annotating images with suggestionsuser study of a tagging system,” in *Advanced Concepts for Intelligent Vision Systems*. Springer, 2012, pp. 155–166.

[18] J. Guo, Z. Qiu, and C. Gurrin, “Exploring the optimal visual vocabulary sizes for semantic concept detection,” in *Content-Based Multimedia Indexing (CBMI), 2013 11th International Workshop on*. IEEE, 2013, pp. 109–114.

[19] Y.-G. Jiang, C.-W. Ngo, and J. Yang, “Towards optimal bag-of-features for object categorization and semantic video retrieval,” in *Proceedings of the 6th ACM international conference on Image and video retrieval*. ACM, 2007, pp. 494–501.

[20] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, “Speeded-up robust features (surf),” *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[21] A. P. Natsev, A. Haubold, J. Tešić, L. Xie, and R. Yan, “Semantic concept-based query expansion and re-ranking for multimedia retrieval,” in *Proceedings of the 15th international conference on Multimedia*. ACM, 2007, pp. 991–1000.

[22] J. Dalton, J. Allan, and P. Mirajkar, “Zero-shot video retrieval using content and concepts,” in *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM, 2013, pp. 1857–1860.

[23] D. Etter and C. Domeniconi, “Semi-supervised rank learning for multimedia known-item search,” in *Proceedings of International Conference on Multimedia Retrieval*. ACM, 2014, p. 257.

[24] 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*. IEEE, 2011, pp. 1–4.

[25] T. Joachims, T. Finley, and C.-N. J. Yu, “Cutting-plane training of structural svms,” *Machine Learning*, vol. 77, no. 1, pp. 27–59, 2009.