7-2010

Faceted topic retrieval of news video using joint topic modeling of visual features and speech transcripts

Kong-Wah WAN

Ah-hwee TAN
Singapore Management University, ahtan@smu.edu.sg

Joo-Hwee LIM

Liang-Tien CHIA

FACETED TOPIC RETRIEVAL OF NEWS VIDEO USING JOINT TOPIC MODELING OF VISUAL FEATURES AND SPEECH TRANSCRIPTS

Kong-Wah WAN [1], *Ah-Hwee TAN* [2]

[1] Institute for Infocomm Research
1 Fusionopolis Way, Singapore 138632
kongwah@i2r.a-star.edu.sg, asahtan@ntu.edu.sg

Joo-Hwee LIM [1], *Liang-Tien CHIA* [2]

[2] School of Computer Engineering,
Nanyang Technological University, Singapore
joohwee@i2r.a-star.edu.sg, asltchia@ntu.edu.sg

ABSTRACT

Because of the inherent ambiguity in user queries, an important task of modern retrieval systems is *faceted topic retrieval* (FTR), which relates to the goal of returning diverse or novel information elucidating the wide range of topics or *facets* of the query need. We introduce a generative model for hypothesizing facets in the (news) video domain by combining the complementary information in the visual keyframes and the speech transcripts. We evaluate the efficacy of our multimodal model on the standard TRECVID-2005 video corpus annotated with facets. We find that: (1) the joint modeling of the visual and text (speech transcripts) information can achieve significant F-score improvement over a text-alone system; (2) our model compares favorably with standard diverse ranking algorithms such as the MMR [1]. Our FTR model has been implemented on a news search prototype that is undergoing commercial trial.

Keywords— Faceted Topic Retrieval, Multimedia Topic Modeling, Latent Dirichlet Allocation

1. INTRODUCTION

A key challenge for modern Information Retrieval (IR) systems is dealing gracefully with ambiguous queries. Studies have shown that most users pose short and broad queries, expecting the IR system to elucidate their possible interpretations [2]. The query "apple" is a classic example of an ambiguous query. In the absence of any disambiguating information about the user, an IR system may best return a mix of documents discussing the fruit, the company and the product.

To account for both the breadth of available information and any ambiguity inherent in the query, several authors have recently proposed *novelty* and *diversity* retrieval tasks with new definitions of relevance and evaluation measures [3, 4, 5]. Of particular interest to this paper is the faceted topic retrieval (FTR) task of Carterette *et al.* [5]. The goal of FTR is to return a set of documents that cover the different *facets* of an information need. The notion of facets here is interpreted broadly to encompass any binary property of a document that represents a fact or a topic that is contained in the information need. For example, given the query "War on Terror in Iraq", the facets may include "War in Fallujah", "Military Strategy in Pentagon", "Bin Laden Tape", etc. The facets of a query can be contained in one or more documents, and a document can contain one or more facets. A document is deemed relevant to the query if it contains any of the facets.

In this paper, we present a FTR model for news *video*. There are two problems in FTR [5]: (1) develop a *Facet Document Model* (FDM) that given a set of documents D, hypothesizes the set of facets in D ; (2) develop a *Facet Retrieval Model* (FRM) that finds the smallest set of documents that cover all of the facets. We focus on the first problem, namely to hypothesize a set of facets in (video) documents. While there are many ways to do that (e.g. clustering, extract keyphrase), we base our FDM on the topic modeling method (Latent Dirichlet Allocation, LDA) of Blei and Jordan [6].

Despite provened capable of mining semantic topics (facets) in text collections, the use of topic model for FTR of news video poses some new challenges. For one, naively feeding the speech transcripts (usually from Automatic Speech Recognition, or ASR) of news video as textual input to LDA will likely not yield good results. This is because ASR transcripts are noisy (ASR word error rates are generally above 20%), whereas most successful application of LDA are reported on clean text (e.g. newswire and publications). Apart from the few sporadic work in [7, 8], the utility of LDA to noisy text source remains suspect.

Another challenge relates to the quality of LDA output: LDA often produces word distributions that are coarse, with no apparent meaning amongst high probability words. This can degrade the quality of detected facets and undermine FTR performance. The common approach to deal with this problem is to introduce side-information into the modeling [6, 9]. In this paper, we explore the use of visual information in the shot keyframes to constrain facet development. There are two motivating intuitions: First, video footages are often repeated for similar or related news stories, and hence are highly correlated with the spoken (ASR) words. Second, different facets of a query may use different sets of words, and the same set of recurring visual shots become a "bridge" between these words, allowing them to be learned as distinct facets. To compute recurring visual shots, we use the Near Duplicate Image (NDI) detection method of Chum *et al.* [10]. We define a multimodal FDM to jointly account for the NDI shots and the ASR words as distinct but correlated sets of observations.

978-1-4244-7493-6/10/$26.00

The rest of the paper is organized as follow. We first examine some related work in Section 2 and briefly review the FTR task and its evaluation metric in Section 3. We provide details of our multimodal FDM in Section 4, and in Section 5, evaluate its elucidation of query facets on a ground truth video dataset from TRECVID-2005. We conclude with some discussions and further work in Section 6.

2. RELATED WORK

Carbonell and Goldstein [1] describe an early attempt to resolve query ambiguity by *search result diversification*. They proposed a Maximal Marginal Relevance (MMR) ranking function to tradeoff between maximizing relevance while minimizing similarity amongst the retrieved documents. Zhai *et al.* [4] extend MMR to a general framework to score documents with probability of relevance and novelty. Similar work by Chen and Karger [11] proposes a greedy algorithm to penalize redundancy in the result set. Carterette *et al.* note that the possibly many interpretations for a given query are indicative of the facets of the query and results can be re-ordered so that each facet can be represented in the top ranks with some probability [5]. Clarke *et al.* argue that result re-ordering must account for the interest of the overall user population [3].

Topic models are first derived as multinomial distributions of unimodal text data, and the joint modeling of multiple data types such as visual and text is not straightforward. The authors in [12, 13] model annotated images using the visual features and text annotations, for automatic annotation and retrieval respectively. Two ways of combining the two modalities are explored: feature concatenation and hierarchical modeling. The former treats the two modalities equally, while the latter first models each individually and then fuse them at a later stage. Our work in this paper has two important differences with these previous works. First, we jointly model visual features and text in the *video* domain. Our modeling granularity is coarser: our visual features are not at the local patch-level but rather at the keyframe-shot-level. This most closely resembles Wu *et al.*'s video representation with visual shot duplicates [14]. Second, our modeling objective is FTR of news stories. Specifically, by adding recurring shot features, we constrain the topic modeling process to learn facets that better correlate with actual news event.

Because topic models are unsupervised methods, often best results are obtained by incorporating a priori knowledge about the desired output (e.g. *must-link* constrains in clustering). Adding observations from cross-media types as a way to constrain topic modeling is proposed by several authors. Blei and Jordan describe an image annotation model to learn the correspondence between an image region and a word in the caption [6]. Most resembling of our own approach, Jain *et al.* guide topic formation of news photo caption by correlating the names with a face recognizer [9].

3. FACETED TOPIC RETRIEVAL

In this section, we briefly review the FTR task in [5]. Recall the two main problems of FTR: to develop a *retrieval* model (FRM) and a *document* model (FDM) for hypothesizing facets. We first review the evaluation metric and then present a probabilistic model for FRM. We reserve our proposed FDM, which is our main novelty, to the next section.

3.1. Evaluation Metric

The object of FTR is to include in the early rankings, many documents that cover as many different facets as attested in the corpus. More formally, given the set of known facets f_i of query q, and a set of documents $\{d_1, d_2, \ldots, d_k\}$ retrieved up to rank k, FTR is evaluated by the *S-recall* of Zhai *et al.* [4]:

$$S\text{-}rec@k = \frac{1}{n_q}\sum_{i=1}^{n_q} I(f_i \in \{d_1, d_2, \ldots, d_k\}) \qquad (1)$$

where n_q is the number of facets of q, and $I(\cdot)$=1 if f_i appears in any of the documents ranked 1 to k, and 0 otherwise. At any fixed k, $S\text{-}rec@k$ rewards returning a list of k documents that contain as many of the n_q facets as possible.

3.2. A Probabilistic Facet Retrieval Model

Given a set of documents $D = \{d_1, d_2, \ldots, d_{|D|}\}$, suppose we have a set of facets $F = \{f_1, f_2, \ldots, f_{|F|}\}$. Denote the probability that all of the facets in F are contained in D as $P(F \in D)$. Assuming that facets occur in documents *independently*, the probability of a facet f_j in at least one document in D is: $P(f_j \in D) = 1 - \prod_{i=1}^{|D|}(1 - P(f_j \in d_i))$ and the probability of all facets in F being contained in D is:

$$P(F \in D) = \prod_{j=1}^{|F|}\left\{1 - \prod_{i=1}^{|D|}(1 - P(f_j \in d_i))\right\} \qquad (2)$$

It is obvious that maximizing $P(F \in D)$ over D and F will also maximize the S-recall metric in equation 1. However, this is generally NP-hard [5]. Hence, we may turn to a greedy heuristic: For each facet f_j, take the document d_i with maximum $P(f_j \in d_i)$, and rank it by its relative order in the original ad-hoc retrieval rank.

4. A FACET DOCUMENT MODEL OF NEWS VIDEO

In this section, we present a Facet Document Model (FDM) to hypothesize a set of k facets $\{f_1, f_2 \ldots f_k\}$ in a news video collection D. Our goal is to compute the $P(f_j \in D)$ in the previous section 3.2. Our overall approach is shown in figure 1. Two feature extraction tracks act on an input news story video simultaneously to compute text and visual features. These are then jointly combined using a generative model to compute facets. We discuss each of these key modules in the following.

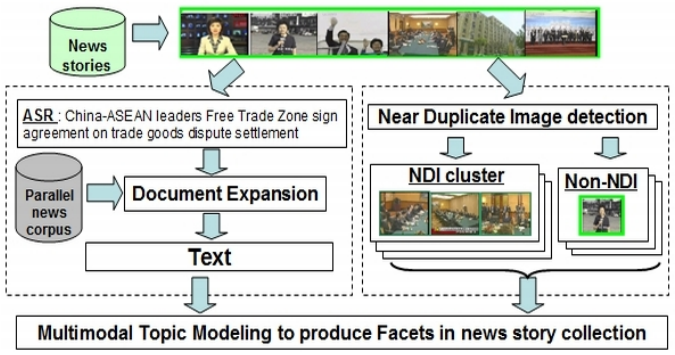

Fig. 1. FDM for hypothesizing facets in news video.

Table 1. Sample expansion words given the words on the left.

Condolezza	state, bush, secretary, security, stanford
Jintao	china, brazil, sino, economic, aids
Basketball	points, conference, group, match, nba

4.1. Text features

While important keywords are generally detected by ASR, the presence of the many misrecognized words can result in erroneous topic formation. Apart from stemming and discarding rare words, we use the document expansion approach of Wan to alleviate the problem [15]. The main idea is to introduce additional words to form an expanded text document vector. These words are selected based on their high mutual information in a parallel news corpus. Such a corpus is readily obtainable since news content is widely available over the internet. For the TRECVID-2005 dataset used in this paper, we build the parallel corpus by issuing to Google Archives News our query description and restricting retrieval time-range to the period when the dataset was collected (Nov-Dec 2004). Table 1 shows some examples of expansion words.

4.2. Visual features

To capture a higher level of visual information, we generalize the common practice of modeling images as bag-of-features to represent video as *bag-of-keyframes*. Following the approach of Wu *et al.* in [14], a keyframe is classified as whether it is a Near Duplicate Image (NDI) to other keyframe(s) or not. By assigning unique IDs to keyframes, they can be treated as visual words. All keyframes in a NDI-cluster are visually similar and are given the same ID. We can now generalize the TF-IDF weighting in text domain to these visual words. Figure 2 shows some examples of how visual similarity in videos is succinctly represented by the term frequencies (tf) and document frequencies (df) of visual words.

To compute near duplicates images, we use a color histogram combined with a spatial pyramid over the image to jointly encode global and local information. This approach is inspired by Chum *et al.* [10], who applied the method to efficiently handle NDI detection amidst jitter and noise. Figure 3 shows the spatial

Fig. 2. TFIDF weighting of keyframe-based visual words. Within the 4 videos, four NDI clusters are shown and colored differently. For the red NDI cluster, it appears in V_2, V_3, V_4, hence its df=3, tf=2 for V_3, tf=1 for V_2, V_4. All non-NDI in a video have df=1 and tf=1. Best viewed in color.

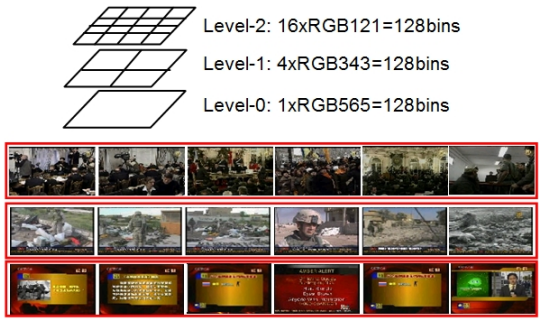

Fig. 3. Top shows the spatial division of the image at each level of the RGB pyramid. At each level, the RGB suffix refers to the number of bits in the color channel. E.g. at level-1, we have 4 divisions each with 3 bits red, 4 bits green, 3 bits blue, totaling 128 bins. Each bin count uses 2 bytes. Therefore, each image is represented by 768 bytes. Bottom shows sample NDI given the leftmost image as query.

pyramid configuration which is arranged so that an increasingly granular grid (i.e. from global to more localized) of color information is stored as we move up the level. The setup provides a highly compressed representation for each image that makes histogram comparison efficient. Given a query image, the NDIs are defined to be those within a specified Euclidean distance from the query. To compute a NDI cluster, we maintain a NDI list that initially only contains the query image. This list is then repeatedly populated with NDI of the new members in the list. The final NDI cluster is then given by the transitive closure of the NDI list.

4.3. Joint Modeling of Visual and Text

We now consider a corpus $\boldsymbol{D}$ of news video, each comprising of text $\boldsymbol{W}$ and visual words $\boldsymbol{V}$. Each video d is modeled as a mixture of latent topics, to *simultaneously* account for $\boldsymbol{W}$ and $\boldsymbol{V}$ as distinct set of observations. Our model is motivated by Blei and Jordan's Corr-LDA model for text and images [6], and also Jain *et al.*'s People-LDA model in [9]. We call our model

FDM-VT, denoting the use of both visual and text modality. Figure 4 shows the graphical representation of FDM-VT. The generative process of FDM-VT is as follow:

- Draw a multinomial ϕ over K topics: $\phi \sim \text{Dir}(\alpha)$
- For each topic $k = 1 \ldots K$,
  - draw multinomial $\beta_k \sim \text{Dir}(\eta_w)$ for text words
  - draw multinomial $\gamma_k \sim \text{Dir}(\eta_v)$ for visual words
- For each text word index n in d, n = 1 to N_d
  - Sample a topic z_n from ϕ: $z_n \sim$ Multinomial(ϕ)
  - Sample a text word w_n from β_{z_n}
- For each visual word index m in d, m = 1 to M_d
  - Sample a topic y_m from ϕ: $y_m \sim$ Multinomial(ϕ)
  - Sample a visual word v_m from γ_{y_m}

where N_d and M_d is respectively the number of text words and visual words in video d, and η_w and η_v are Dirichlet priors for the text and visual words distribution respectively. The above FDM-VT model results in the following joint distribution on text $\boldsymbol{W}$, visual $\boldsymbol{V}$ and the latent topics:

$$P(\boldsymbol{W}, \boldsymbol{V}, \phi, \mathbf{z}, \mathbf{y}) = P(\phi|\alpha)\Big(\prod_{n=1}^{N_d} P(z_n|\phi)P(w_n|z_n,\beta)\Big) \Big(\prod_{m=1}^{M_d} P(y_m|\phi)P(v_m|y_m,\gamma)\Big) \quad (3)$$

The main difference of our model from the Corr-LDA in Blei and Jordan [6] is that we use two multinomial distributions for the text and visual words. The sampling of visual words is essentially the same as the sampling text words. However, within a video, ϕ is a higher-level factor that is held fixed and it governs the ensemble of all text word and visual word observations. The topic-word multinomial β and γ now learns the combined co-occurrence of important text words across video documents and also the complementary visual words.

Several approaches to learning the FDM-VT parameters exist in the literature, such as Variational inference [6] and Gibbs Sampling [16]. We choose a simple extension of the latter by iterating over each text word, visual word and video document, each time resampling a single topic of the word (text or visual) based on the current topic assignment for the document and all other observed words (text and visual). A perplexity measure on a held-out set is used to determine learning convergence. On the 1028 TRECVID-2005 video documents comprising of 210K text words and 95K visual words, our implementation on a standard 3Ghz PC takes about 10 minutes to compute. We noticed that varying the Dirichlet priors η and α did not affect performance too much. We used the same value of 0.2 in the experiments below.

4.4. Re-ranking

After the FDM-VT learning has converged, the K latent topic distributions of both the text words and visual words are given by β_k and γ_k, k=1..K. In particular, the probability of a text-word w in the k^{th} latent topic is given by $P(w|\beta_k)$. Given $d_i \in \boldsymbol{D}$, we have $P(d_i|\beta_k) = \prod_{w \in d_i} P(w|\beta_k)$. We now assume that the multinomial β_k represents the k^{th} actual facet f_k in $\boldsymbol{D}$. The probability of d_i containing the k^{th} facet is now: $P(f_k \in d_i) = P(d_i|\beta_k) = \prod_{w \in d_i} P(w|\beta_k)$. Referring back to the greedy heuristic to maximize facet retrieval in section 3.2, for each topic $k = 1..K$, we find the document d_i that maximizes $P(d_i|\beta_k)$ and *re-rank* it by its relative order in the original ad-hoc retrieval rank.

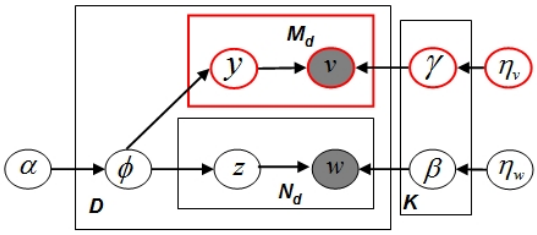

Fig. 4. Graphical representation of FDM-VT. The red box encloses the additional observations from the visual modality, which constrain the topic formation in the text modality.

5. EXPERIMENTAL EVALUATION

5.1. Dataset

A few corpora have been created to facilitate research in novelty and diversity tasks in text (TDT5 [5], TREC-2006 Q&A [3]) and images (ImageCLEFPhoto [17]). However, no corresponding dataset for video is available. To evaluate our methods, we turn to a subset of the TRECVID-2005 video dataset [18]. It comprises of about 127 hours of Chinese and English news broadcast from 5 different sources (e.g. CCTV4, CNN). The dataset also includes computed story boundaries from the CMU Informedia Lab.

For queries, we used the topic labels provided by Wu *et al.*, who also used the same TRECVID dataset [14]. In their work, 33 story topics are defined on which each news story is manually labeled with one of the topic labels. Sample topic labels include "Bush visits Canada", "Mideast-Peace", "Arafat-health". For the full list of topics, please refer to their paper.

For our FTR evaluation, three human assessors further annotated facets on these video topics. For each of the 33 topics, the assessors were given the videos labeled as relevant and asked to name the facets and provide some keywords. Assessors were given free reign to name and describe facets, but they were advised to consult online resources such as the Wikipedia and news commentary sites. On average, there are 43.5 relevant videos per query topic, 5.7 facets per query topic, and 3.1 facets per video. There are a total of 70 unique facets. While the assessors generally agreed on the number of facets per video, they disagreed more on the facets in a query topic. We mostly resolve this by taking a liberal and all-inclusive approach to accept all facet suggestions.

5.2. Comparing Retrieval Engines

We compare FTR performance with a well-known diversity retrieval model of Carbonell and Goldstein [1], called Maximal

Marginal Relevance (MMR). This is a greedy method that selects the i^{th} document d_i according to a combination of its similarity to the query Q and its similarity to all higher-ranked documents at position 1 to k-1:

$$\textbf{MMR}(d_i, Q) = \alpha \text{Sim}_1(d_i, Q) - (1-\alpha) \max_{1 \le j < i} \text{Sim}_2(d_i, d_j)$$

where $\text{Sim}_1()$ is a standard language model-based document-query similarity function with Dirichlet smoothing [19], $\text{Sim}_2()$ is a Cosine similarity function between documents, and α is a trade-off parameter between 0 and 1. When $\alpha = 1$, MMR ranking falls back onto the standard relevance ranking based on query-document similarity. We set $\alpha = 0.5$.

For our FDM modeling of video, we have the following ways to hypothesize facets:

- Unimodal FDM modeling only on text:
  - **FDM-T-ASR:** on only the ASR speech transcripts.
  - **FDM-T-ASR+DE:** on the resulting text vector after document expansion on ASR transcripts.
- Joint FDM modeling on text and visual features:
  - **FDM-VT-ASR:** visual features and ASR
  - **FDM-VT-ASR+DE:** visual features and text from document expansion of ASR transcripts.

For each of the above, we also have the following options:

- **FDM on entire corpus:** hypothesize facets in the entire corpus (1028 videos). In our dataset, we have a total of 70 unique facets. So we set the number of latent topic in FDM modeling K to be 70.
- **FDM on top-D documents:** For each query, hypothesize facets in the top retrieved videos. In our dataset, we have on average 43.5 relevant videos per query. So we round off $D = 50$. We also have an average of 5.7 facets per query, and so we set K to be 6.

While the latter setup simulates the results obtainable in interactive retrieval, real-time response is severely limited by the huge CPU demand on visual computing. Hence, the first setup is generally the only permissible option for real-life deployment. In all 5 retrieval methods, re-ranking (Section 4.4) is performed on the top-100 videos. For fair comparison, the same $\text{Sim}_1()$ document scoring function is used.

5.3. Results and Analysis

For each comparative FTR method, we report S-recall at the 20^{th} rank. This assumes that users are interested in the first 2 pages of search results. While S-recall measures the retrieving of facets, it is nonetheless desirable to have a ranked list also populated by *relevant* documents. To capture both intent, we introduce an extra F1-measure that combines a standard precision metric with S-recall: F1 $= \frac{2 \cdot (Prec@20 \cdot S\text{-}rec@k)}{(Prec@20 + S\text{-}rec@k)}$. Table 2 reports both the S-recall and precision at rank 20.

All methods are able to retrieve nearly 50% of the facets, with FDM-VT-ASR+DE achieving the best overall F-score. Notably, this is obtained by modeling on the top-D documents, and the results from modeling on the entire corpus are marginally *worse* than the MMR baseline. This points to the general difficulty in topic modeling of noisy documents.

Table 2. FTR performance on TRECVID-2005 dataset. The best result in each FDM setting is in bold. An asterisk indicates significant statistical difference over the MMR baseline.

	Prec@20	*S-rec*@20	F1
MMR	0.3276	0.4704	0.3862
modeling on entire corpus ($K = 70$)			
FDM-T-ASR	0.2410	0.4472	0.3132
FDM-T-ASR+DE	0.3187	0.4591	0.3762
FDM-VT-ASR	0.2772	0.4521	0.3437
FDM-VT-ASR+DE	0.3306	0.4612	**0.3851**
modeling on top-D documents ($D = 50, K = 6$)			
FDM-T-ASR	0.2810	0.4618	0.3494
FDM-T-ASR+DE	0.3487	0.4707	0.4006
FDM-VT-ASR	0.3272	0.4834	0.3903
FDM-VT-ASR+DE	0.3706	0.4992	**0.4254***

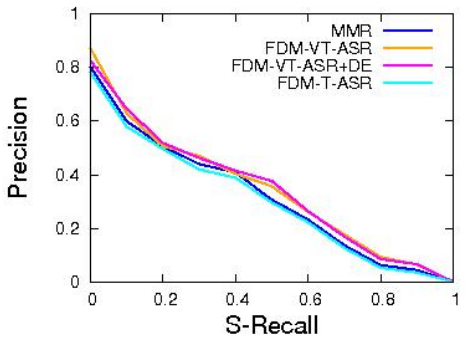

Fig. 5. S-recall versus Precision plot for three FDM models and MMR baseline. Best viewed in color.

Consistent with results obtained elsewhere [15], modeling benefits from augmenting text by document expansion. The query with the lowest F-score is "Bush second term plan" (0.0478). While the S-recall is decent, this query suffers from very low precision score as it has a lot of overlap with queries such as "War-on-Terror". The query with the highest F-score is "HIV aids" (0.7545). Again high precision plays a major part, since the topic is unique in the corpus.

Figure 5 compares the S-recall-vs-precision plots of three FDM models with the MMR baseline. While all four coincide closely, the text-only FDM-T-ASR model significantly underperforms compared to the rest, and the visual+text FDM-VT-ASR+DE model improves over the MMR baseline. In Table 3, we qualitatively show how the text-words multinomial β_k has benefited from the inclusion of visual features during topic learning. On each of two queries "War on Fallujah" and "Mideast peace", we introspectively pick a learned FDM topic that contains high probability words that are meaningful to the queries. Observe that words from multimodal model are more intuitive and correlate better to the query topic.

Table 3. Top-10 probability words of 2 learned FDM β_k topics, for "War on Fallujah" (top) and "Mideast peace" (bottom). The topics are picked introspectively. Meaningful words highlighted in bold.

FDM-T (text-only)	FDM-VT (Visual+text)
iraq peopl time good meet **unit**	**iraq iraqi** peopl **militari** govern **arm**
state baghdad chines dai	**kill forc attack baghdad**
arafat know **yasser** leader thing	**palestinian arafat peac israel** presid
peac minist peopl just **israel**	leader elect **yasser east middl**

6. CONCLUSIONS AND FUTURE WORK

While the past decade has seen tremendous progress in robust document ranking, a major challenge remains when users underspecify their true information needs. As a step towards addressing this problem, we propose a faceted topic retrieval (FTR) model to mine the wide range of facets in documents. Implicit in our methodology to resolve query ambiguity is that by returning a set of diverse or novel results that are pertinent to the query, the average user will find something useful.

Accurate elucidation of facets depends on learned topics that are highly correlated with actual queried topics or events. In the case of media-rich content such as video, we argue that this can best be achieved by joint modeling in the complementary visual and text modalities. We apply and test this approach to hypothesize facets in a news video search application. Using a variant of the LDA topic model to combine the visual keyframes and the speech transcripts, we show the efficacy of our method on the standard TRECVID-2005 news corpus annotated with facets. We find that the joint model can achieve significant F-score improvement over a text-alone system and that it compares favorably with standard diverse ranking algorithms such as the MMR. Motivated by our findings, we have implemented our FTR model on a live news search prototype that is undergoing commercial trial. As a further work, we will write a follow-up paper to provide more details and empirical findings in our real-world deployment.

A few serious criticisms can be levied against some of our assumptions. The first is the assumption that facets occur in documents independently. In our retrieval model, we have also conveniently made use of privileged information about our dataset such as the average number of relevant documents and number of facets per query topic, to provide a number for D and K in section 5.2. It is not immediately clear how these parameters can be determined in a real-world setting. Having said this, specifying the number of latent topic K remains generally a problem in the topic modeling literature. Last but not least, it appears that further work can be explored along the direction of how best to exploit the use of other visual features for better joint modeling.

Despite these concerns, we believe we have made substantial progress towards our goal of elucidating facets in documents and providing diversity in video search results.

7. REFERENCES

[1] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," *Proc SIGIR*, pp. 335–336, 1998.

[2] J. Wen, J. Nie, and H. Zhang, "Query clustering using user logs," *ACM Transactions Information Systems, Vol. 20, no. 1*, pp. 59–81, 2002.

[3] C. Clarke, M. Kolla, G. Cormack, O. Vechtomova, A. Ashkan, S. Buttcher, and I. MacKinnon, "Novelty and diversity in information retrieval evaluation," *Proc SIGIR*, pp. 659–666, 2008.

[4] C. Zhai, W. Cohen, and J. Lafferty, "Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval," *Proc SIGIR*, pp. 10–17, 2003.

[5] B. Carterette and P. Chandar, "Probabilistic models of ranking novel documents for faceted topic retrieval," *Proc CIKM*, pp. 1287–1296, 2009.

[6] D. Blei and M. Jordan, "Modeling annotated data," *Proc SIGIR*, pp. 127–134, 2003.

[7] J. Cao, J. Li, Y. Zhang, and S. Tang, "LDA-based retrieval framework for semantic news video retrieval," *Proc ICSC*, pp. 155–160, 2007.

[8] M. Purver, K. Krding, T. Griffiths, and J. Tenenbaum, "Unsupervised topic modelling for multi-party spoken discourse," *Proc COLING/ACL*, pp. 17–24, 2006.

[9] V. Jain, E. Learned-Miller, and A. McCallum, "People-LDA: Anchoring topics to people using face recognition," *Proc ICCV*, 2007.

[10] O. Chum, J. Philbin, M. Isard, and A. Zisserman, "Scalable near identical image and shot detection," *Proc CIVR*, pp. 549–556, 2007.

[11] H. Chen and D. Karger, "Less is more: Probabilistic models for retrieving fewer relevant documents," *Proc SIGIR*, pp. 429–436, 2006.

[12] F. Monay and D. Gatica-perez, "Modeling semantic aspects for cross-media image retrieval," *IEEE PAMI*, vol. 29, pp. 1802–1817, 2007.

[13] R. Lienhart, S. Romberg, and E. Horster, "Multilayer plsa for multimodal image retrieval," *Proc CIVR*, pp. 1–8, 2009.

[14] X. Wu, A. Hauptmann, and C. Ngo, "Novelty detection for cross-lingual news stories with visual duplicates and speech transcripts," *Proc ACM Multimedia*, pp. 168–177, 2007.

[15] K. Wan, "Exploiting story-level context to improve video search," *Proc ICME*, 2008.

[16] T. Griffiths and M. Steyvers, "Finding scientific topics," *Proc Natl Acad Sci U S A*, vol. 101 Supp 1, pp. 5228–5235, 2004.

[17] M. Paramita, M. Sanderson, and P. Clough, "Diversity in photo retrieval: overview of the ImageCLEFphoto task 2009," *CLEF working notes*, 2009.

[18] *TRECVID 2005*, http://www-nlpir.nist.gov/projects/trecvid.

[19] *Lemur Toolkit*, http://www.lemurproject.org.