Multi-Modal Query Expansion for Web Video Search*

Bailan Feng[1,2], Juan Cao[1], Zhineng Chen[1,2], Yongdong Zhang[1], Shouxun Lin[1]
[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[2]Graduate School of the Chinese Academy of Sciences, Beijing 100039, China
{fengbailan, caojuan, chenzhineng, zhyd, sxlin}@ict.ac.cn

ABSTRACT

Query expansion is an effective method to improve the usability of multimedia search. Most existing multimedia search engines are able to automatically expand a list of textual query terms based on text search techniques, which can be called textual query expansion (TQE). However, the annotations (title and tag) around web videos are generally noisier for text-only query expansion and search matching. In this paper, we propose a novel multi-modal query expansion (MMQE) framework for web video search to solve the issue. Compared with traditional methods, MMQE provides a more intuitive query suggestion by transforming textual query to visual presentation based on visual clustering. Parallel to this, MMQE can enhance the process of search matching with strong pertinence of intent-specific query by joining textual, visual and social cues from both metadata and content of videos. Experimental results on real web videos from YouTube demonstrate the effectiveness of the proposed method.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process

General Terms

Design, Performance, Experimentation

1. INTRODUCTION

Most of existing popular multimedia search engines (such as Google, Yahoo! and Bing) allow users to represent their search intents by issuing the query as a list of keywords, and provide textual query expansion (TQE) strategy for disambiguation. Alternatively, the authors in [5] formulate a visual query suggestion (VQS) framework by providing image presentations to help users express their search intent for image search; nevertheless it is still mainly depending on textual clustering to overcome query ambiguity. As we all know, compared with images and audios, the annotations around web videos are usually much noisier, which could result in the unsatisfactory performance in video search application when query expansion is only based on textual cues.

Motivated by these analyses, actually, the noisiness of annotations around web videos is the main challenge which limits the effectiveness of web video search. To address this challenge, we propose a novel query expansion framework (see Figure 1), named multi-modal query expansion (MMQE) for web video search application.

The main contributions of this paper are twofold:

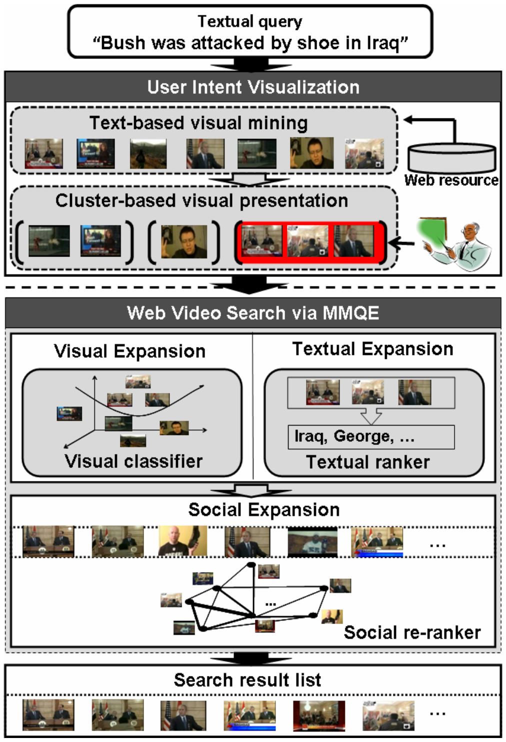

Figure1. Framework of Multi-Modal Query Expansion for Web Video Search

We propose a novel query expansion framework named MMQE for web video search. MMQE can assist users to formulate an intent-specific query by transforming ambiguous textual query to intuitive visual presentation based on visual clustering. Meanwhile, MMQE facilitates the process of human-computer interaction with a simply one-time operation.

We investigate different strategies based on textual, visual and social cues to enhance the process of search matching with strong pertinence of intent-specific query, which is able to help users specify and deliver their search intents in a more precise and efficient way, and can contribute to a significant search improvement.

2. MMQE FOR WEB VIDEO SEARCH

Given a textual query, MMQE for web video search is a two-stage process as illustrated in Figure 1. The first stage named user intent visualization is designed to formulate an intent-specific visual query for users. The second stage utilizes multi-modal cues of videos to provide a more effective expansion scheme for web video search. The challenge of noisy annotations around videos can be well addressed by both stages mentioned above.

2.1 User Intent Visualization

User intent visualization aims to avoid the interference of noisy annotations in search intent presentation. As introduced in [5], textual information is mainly used to suggest search intent by key-

*This work was supported by the National Basic Research Program of China (973Program, 2007CB311100), National High Technology and Research Development Program of China (863Program, 2007AA01Z416), National Nature Science Foundation of China (60873165, 60902090), Beijing New Star Project on Science & Technology (2007B071), Co-building Program of Beijing Municipal Education Commission.

SIGIR'10, July 19-23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

words clustering. Different from image annotations, video annotations are much noisier. Therefore, we depend on the more reliable and objective visual cues of videos to suggest search intent for users.

We first use annotations around videos to transform textual query to visual query set by word co-occurrence statistics. This process may bring lots of noisy videos due to the ambiguous textual query and the noisy annotations. For example, as illustrated in Figure 1, some noisy videos about somebody talking about "Bush attacked by shoe" are expanded into the visual query set. After that, we resort to affinity propagation [1] to automatically cluster the videos in the visual query set into different search intent categories based on SIFT feature. Then users can choose their intent-specific category intuitively to perform the following search process.

2.2 Web Video Search via MMQE

With the selected intent-specific category, three expansion strategies are designed to boost the web search quality.

Firstly, a visual classifier is trained with the visual query videos as positive samples, and negative samples are randomly extracted from testing dataset. Actually, any classifier model is extensible in this framework. Considering the fact that the positive samples are of high quality and limited quantity, in this work, we choose SVM as the visual classifier model, and leverage the time efficiency and the performance effects acceptably.

Secondly, a textual ranker is designed to expand the original textual query with more relevant keywords. We calculate the tag frequencies in the intent-specific category and empirically extract the tags whose frequencies are greater than half of the maximum frequency as the expansion keywords. Both expansion strategies motioned above result in the improvement of search recall.

Thirdly, a social re-ranker is utilized to refine the average fusion list of visual classifier and textual ranker according to the social relationships among videos. Considering the fact that relevant videos are generally with similar social relationships, such as communities and categories, we employ a manifold ranking algorithm [2] to utilize these social relationships for search improvement. On the basic assumption that users only care about the top order relation of search results, we firstly construct a correlation graph on the top n fusion list. Each vertex denotes a video entity, and each edge denotes the social correlation between two video entities. Recommendation and category information are utilized to assign the weight of edges. If the corresponding two videos both have the same two, one or none of above cues, the weight of edges will be given high, middle or low values respectively. Then top k (k<<n) videos of fusion list are utilized as pseudo query points and ranking cues are reserved to trigger the propagation of social relationships on the graph. After convergence, the resultant ranking score of each video is in proportion to the probability that it is relevant to the intent-specific query. The process of social re-ranker mentioned above can adjust a number of relevant videos to a closer top position, and therefore contribute to the improvement of search precision. For example, as illustrated in Figure 1, the visual effect of the final search list is to some extent better than the fusion list after social re-ranking.

3. EXPERIMENTS

We conduct the MMQE framework on MCG-WEBV [4], a web video dataset containing 80,031 representative YouTube videos. The parameters n and k are empirically set to be 1000 and 10 to meet the requirement of near real-time search. We select ten topics with the most popular view from the existing studies [4] and extract no more than two keywords as the corresponding initial queries based on the statistical conclusion in [3]. The details of the hot topics are shown in Table 1.

Table1. Hot Topics

Topic	Keywords	Topic	Keywords
Bush was attacked by shoe in Iraq	bush shoe	HotForWords	hotforword
Miley Cyrus's show	miley cyrus	Highway hero dog	hero dog
Christmas blessings	christmas	Warcraft play scenario	warcraft
Barack Obama's inaugural speech	obama inaugural	Eye makeup	eye makeup
Madonna's concert in Brazil	madonna brazil	Iphone advertisement	iphone

We adopt NDCG@k as the evaluation metric and evaluate the performance of three search strategies:

Google_E: searching videos using the official textual expansion proposed by Google engine.
VQE+TQE: searching videos using the combined query expansion consisting of the visual classifier and the textual ranker.
VQE+TQE+SQE (MMQE): re-ranking the returned videos of VQE+TQE based on the social re-ranker.

The average performance over the ten queries is shown in Figure 2. From the figure, we can see that by specifying the search intent using visual and textual cues in the proposed expansion framework, VQE+TQE can overcome the noisiness issue and thus outperform the Google_E strategy. By further appending social cues to the expansion, the VQE+TQE+SQE (MMQE) strategy gets the best performance. Both observations above demonstrate the effectiveness of MMQE for web video search application.

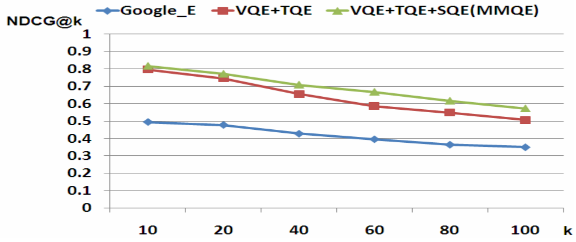

Figure2. Comparison of NDCG@k for MMQE and Google_E over Ten Queries

4. CONCLUSION

This paper has proposed a novel multi-modal query expansion (MMQE) framework for web video search. By formulating an intent-specific visual query and leveraging textual, visual and social cues of web videos, MMQE successfully addresses the noisiness challenge of annotations around web videos. Experimental results on web video search demonstrate the superiority of MMQE than the traditional web search engine.

5. REFERENCES

[1] B. Frey, et al. Clustering by passing messages between data points, *Science*, 319(5814):726, 2007.

[2] D.Y Zhou, et al. Ranking on data manifold, *Proc. NIPS*, 169-176, 2003.

[3] D. Tjondronegoro, et al. Multimedia web searching on a meta-search engine, *Proc. ADCS*, 80-83, 2007.

[4] J. Cao, et al. MCG-WEBV: A benchmark dataset for web video analysis, *Technical report, ICT-MCG-09-001*, 2009.

[5] Z.J. Zha, et al. Visual query suggestion, *Proc. MM*, 15-24, 2008.