Known-Item Search by MCG-ICT-CAS*

Juan Cao, Yong-Dong Zhang, Lin Pang, Bai-Lan Feng and Jin-Tao Li

Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China

{caojuan, zhyd, panglin, fengbailan, jtli }@ict.ac.cn

ABSTRACT

This paper describes the highlights of known-item search system for TRECVID 2010. We first propose that there lies Understanding Gap between a video's author and user, which gap has been represented in the author labeled semantic text(sText) description and user generated visual text(vText) query. To bridge this gap, we explore the structured online knowledge from Wikipedia and the data-driven statistics from Google search engine to build map between sText and vText. The experiment results in this KIS task is promising. Meanwhile, by exploring all kinds of visual based methods, we conclude that the great diversity of the web video's content has made it difficult to find effective visual reference materials from Web, which leads to pool results for visual based methods.

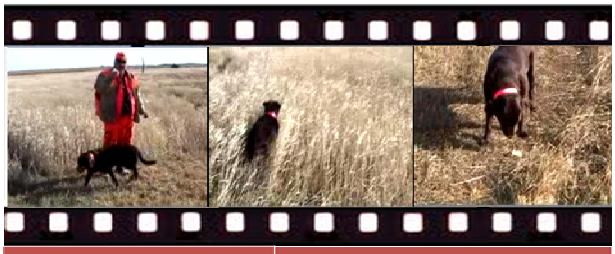
Keywords

known-item search, Understanding Gap, vText, sText, Wikipedia, context-oriented expansion

1. Introduction

When the users have lost in the tremendous scale web videos, the exact retrieval such as Known-item search technology can give them an effective solution to find what they really want. Focus on the details of this task, the query is a text description given by a user who has viewed this video, and several visual-related keywords called visual cues. The query text by user usually describe the visual attributes of the video such as object, person, and location visible in the target video. However, the author of this video has background knowledge of this video's particular situation, then the metadata generated by him usually includes the semantic description for the video content such as what event happened. Here we define the author's text metadata as ***Semantic-text(sText)*** description, and the user's text query as ***Visual-text(vText)*** description. Moreover, there is important difference between the two text descriptions, we called ***Understanding Gap.*** How to bridge the Understanding Gap between two texts is a key problem for known-item search.

Figure 1 is an example of this gap. The author has edited a video of hunting trip coming from cable TV, so he tagged the video with "hunting, cable TV …". But for a user, he hasn't this semantic knowledge, and only can give the vText to describe what he has saw from the video, such as "orange outfit" ,"black dog", "apple" etc.



Author Tags(sText):	User Query(vText):
hunting; hunters; hunting trip; gun; dog; comedy; cable TV; MTN	a man in orange outfit throwing an apple for a black dog with red collar to retrieve and the dog retrieves but eats the apple

Figure 1 An example of the Understanding Gap between sText and vText.

Moreover, the great innovation of TRECVID[1] dataset from static video dataset to the open web video collection has given a challenge for this task. Firstly, most of the videos in the Internet are edited by the amatuers，which leads that the quality of web video is diverse. So the effectiveness of visual feature need to be verified. On the other hand, the web video has textual tags labeled by users, but compared with the professional edit, they are sparse and noisy. So the enrichment and refinement of textual feature is needed.

According the above analysis, we apply two text enriching algorithms separately based on the online knowledge collection Wikipedia and on the online search engine Google. To verity the performance of these algorithms, we have designed following four runs for the known-item search task of TRECVID 2010, and have got the promising results shown in Table 1.

Run1: Search by combination of both wikipedia-based and context-oriented expansion for metadata.

Run2: Search by wikipedia-based expansion for metadata.

* This work was supported by the National Basic Research Program of China (973 Program, 2007CB311100), National Nature Science Foundation of China (60902090, 60873165, 60802028), Co-building Program of Beijing Municipal Education Commission.

Run3: Text baseline search using metadata.

Run4: Visual baseline.

Table 1. the performance of four runs for known-item search task

Run_ID	meanInvertedRank
F_D_YES_MCG_ICT_CAS1_1	0.236
F_D_YES_MCG_ICT_CAS1_2	0.238
F_D_YES_MCG_ICT_CAS1_3	0.231
F_D_YES_MCG_ICT_CAS1_4	0.001

2. Visual baseline Search

The aim of visual run is to verify whether visual feature is effective in the known-item search process. This run includes the following three steps:

Visual data obtainment: The KIS task only supports the textual query. To collection the visual data, we first extract three to five important keywords from the query's visual cues. Then we search related web images for every selected cue from Flickr by its API, and expand the top 200 ones as the visual samples for this cue.

Visual feature representation: For the sample collection expanded from Flickr is diverse, in this step, we try to mine the main visual cluster of the collection and represent it. Firstly, we extract the SIFT feature and represent each image as a 5000-dimensional soft-visual-keywords vector according to the work of City University of Hong Kong [8]. Then, we cluster the 200 images in the sample collection of each visual cue by k-means clustering and utilize the center of the biggest cluster as the visual feature representation of this visual clue.

Visual similarity computation: Based on the above visual representation, we compute the cosine similarities between the visual cue and the test keyframes, and rank the retrieval list for each cue. Then the final search result for a query is generated by average fusing the retrieval lists of its all selected visual cues.

3. Text baseline Search

In the text baseline, we use the given metadata of videos, and propose an effective text pre-processing technology based on Wikipedia[3], called ***Longest Match Principle (LMP)***[6] .

Firstly, we extract text feature including description, subject, and title from the metadata of videos in test collections. Then, we apply LMP to refine the original text feature. At last, we use lucene[2] to build index and implement textual retrieval.

Among the video tags, Named Entity (NE) plays a key role in describing the video's semantic content. But many human annotated NEs are noisy and personalized, which include the abbreviations, nicknames and even misspells. We propose a Longest Match Principle (LMP) to validate each tag in Wikipedia. For example, a video tagged with "Lake, Superior", "Lake" and "Lake Superior" are all detected as Wikipedia concepts, LMP algorithm aims to select the "Lake Superior" as the standard representation for all these tags. In our experiment, the LMP has achieved important improvement for the retrieval performance.

4. Web Expansion based Search

As introduced above, there is Understanding Gap between the author's metadata and user's query, therefore, directly matching the text query to the metadata cannot find the target videos. To bridge the gap, we try to mine the mapping relationships between sText and vText from the external web knowledge resources, including the structural knowledge expansion from the online corpus Wikipedia, and the context-oriented expansion from the result collection of Google search engine. To compare the effectiveness of both resources, we submitted two runs: Run 2 is Wikipedia-based expansion, and Run 1 is the fusion of Wikipedia and context- oriented expansion.

4.1 Structural Knowledge Expansion on Wikipedia

Wikipedia[3], as one of the largest online encyclopedia has attracted extensive attention from many research areas such as natural language processing and information retrieval. Compared with static knowledge ontologies such as WordNet, Wikipedia is a live collaboration whose content being continually created and updated by the web users. It implies that Wikipedia keeps synchronized with the growth of Internet, and we can get the expansion results for most of the popular web video tags. In our method, we utilize the Wikipedia-similarity computation proposed in [5], which has promising definition for semantic relatedness between concepts.

Firstly, we extract subject and title text from the query video's metadata, and further refine them by LMP algorithm and the other basic pre-processing such as stemming and filtering to make sure only the noun and verb are reserved. Then for each selected keyword, we obtain the top K related concepts by computing the wiki-similarity through the API provided in [5]. Afterwards, we use lucene[2] to build index for the expanded text dataset and compute the retrieval ranking scores for each video. At last, the score from the expanded text and the original ranking score in the text-baseline are linearly fused for final ranking in Run 2. From the experimental results, we can see that this run is much better than the text baseline run. The results show that Wikipedia based expansion can enrich the metadata with more semantic related words and thus can relieve the zero-matching problem between query and video.

4.2 Context-Oriented Expansion on Google

In the huge WWW, due to the needs of different web users, a web event is usually covered in many websites and in multiple media forms such as video, image, and texts. For example, given a query of "Pope Christmas mass", the top 10 results from Google search includes News reports, videos, as well as user comments from online community. These resources build a rich and consistence context for the event. Therefore, for a web video, if we can collect the polymorphic web resources describing the same event with it, then a much more enriched tag list can be obtained.

Based on the above analysis, we adopt the context-oriented tag recommendation (CtextR) approach in [6] to enrich the metadata for videos in the test collection. Given a web video, CtextR focuses on finding the resources describing the same video event from the WWW, and expanding and recommending tags under the context-consistent constraint.

First, we construct a query using the subject text from metadata of a video to capture the context of the video. Then, the query is submitted to Google search engine and the top returned resources are identified and collected. Next, significant keywords from the resources are discovered and ranked through a PageRank-like graph model. After that, we can obtain the top recommended words for each video. Similar to the Wikipedia based expansion, we then use lucene[2] to build index for this expanded text and compute the retrieval ranking scores for each video. At last, Run 1 is ranked by the linear fusion of scores from original metadata text retrieval, Wikipedia expansion and context-oriented expansion. From the experimental results, we can see that this fusion run is better than the text baseline run. Though the average results for all 300 queries are similar with Wikipedia-based run, the context-based expansion did improve the result of Wikipedia based run in many queries. These results encourage us to further consider more efficient fusion methods to take into account both the context-oriented expansion and Wikipedia based expansion.

5. Experiments and Analysis

In this year, we have submitted four runs for Known-item search task. One is based on the visual feature, and the others are based on the web expanded text features.

To verity whether the visual-based methods are effective to KIS search, besides the submitted visual baseline, we really have explored the other traditional visual feature based approaches[9] such as the concept-based search on this year's High Level Feature Detection results. Both have got very low precisions in the 130 topics for train. By analyzing the data, we summarized two reasons: **Firstly, the content of web video is too diverse to find useful reference materials.** Despite covering all kinds of general video categories such as politics, entertainment and travel etc., it is also personalize between different persons such as one of his habit or his interested thing. Most of these videos can not find another similar one from Internet. **Secondly, the quality of web video is diverse.** A great part of the web video is recorded and edited by amateurs with low quality, which lead to seriously miss-matching when we computing their visual similarities.

Table 2. the number of topics at different rank levels for run1(Wikipedia & Google), run2(Wikipedia) and run3(Text baseline).

Rank Level	1	<5	<10	<50	<100
Wikipedia & Google	54	87	96	140	**156**
Wikipedia	**56**	84	95	132	150
Text baseline	50	85	96	136	152

On the other hand, the three text-based runs have achieved stable good performances. Table 2 shows the detail ranking results of three runs, which means the target video has been ranked at top 1, top 5, top 10, top 50 and top 100. We can see that Wikipedia based expand has improved the hit rate of top 1 from 50 to 56, that is the aim for exact retrieval like KIS search. Meanwhile, the fusion of Wikipedia and Google has improved the hit rate of top 100 from 152 to 156. It implies that the rich reference materials can help us to get back some missed targets.

In the future work, we will continue to mine the social information from the web resources such as Wikipedia and Google.

Reference

[1] A. F. Smeaton, P. Over and W. Kraaij, Evaluation campaigns and TRECVid. In Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval(MIR '06), 2006

[2] J.Lucence,"Jakarta Lucene Text search engine in java", http://jakarta.apache.org/lucene/docs/index.html

[3] http://www.wikipedia.org

[4] K. Nakayama, T. Hara, and S. Nishio, "Wikipedia mining for an association web thesaurus construction," in Proc. of IEEE International Conference on Web In-formation Systems Engineering (WISE 2007), pp. 322–334, 2007.

[5] http://wikipedia-lab.org/en/index.php/Wikipedia_API

[6] Y.C. Song, J. Cao, Z.N. Chen, Y.D. Zhang, J.T. Li, Tag Transformer, ACM International Conference on Multimedia (ACM MM 2010), Florence, Italy, 2010. (Accepted)

[7] Z. Chen, J. Cao, Y. Song, J.Guo, Y.Zhang, J.Li, Context-oriented Web Video Tag Recommendation, the ACM World Wide Web 2010, pp.1079-1080, 2010.

[8] Y.G. Jiang, C.W. Ngo, J. Yang. Towards Optimal Bag-of-Features for Object Categorization and Semantic Video Retrieval. ACM International Conference on Image and Video Retrieval (CIVR), Amsterdam, 494-501 (2007)

[9] J. Cao, Y.D. Zhang, B.L. Feng, X.F Hua, L. Bao, and X. Zhang, MCG-ICT-CAS TRECVID2008 search task report, Proceedings of TRECVID, 2008.