Large-Scale Cross-Media Retrieval of WikipediaMM Images with Textual and Visual Query Expansion

Zhi Zhou[1,2,3,*], Yonghong Tian[3,*], Yuanning Li[1,2,3], Tiejun Huang[3], and Wen Gao[3]

[1] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[2] Graduate University of Chinese Academy of Sciences, Beijing 100049, China
[3] Institute of Digital Media, School of EE & CS, Peking University, Beijing 100871, China
{zzhou,ynli}@jdl.ac.cn, {yhtian,tjhuang,wgao}@pku.edu.cn

Abstract. In this paper, we present our approaches for the WikipediaMM task at ImageCLEF 2008. We first experimented with a text-based image retrieval approach with query expansion, where the extension terms were automatically selected from a knowledge base that was semi-automatically constructed from Wikipedia. Encouragingly, the experimental results rank in the first place among all submitted runs. We also implemented a content-based image retrieval approach with query-dependent visual concept detection. Then cross-media retrieval was successfully carried out by independently applying the two meta-search tools and then combining the results through a weighted summation of scores. Though not submitted, this approach outperforms our text-based and content-based approaches remarkably.

Keywords: Image retrieval, textual query expansion, query-dependent visual concept detection, cross-media re-ranking.

1 Introduction

The WikipediaMM task at ImageCLEF 2008 aims to investigate effective retrieval approaches in a large-scale collection of Wikipedia images. In the task, participants need to deal with searching 75 topics from approximately 150,000 images. Search over such a large-scale image collection offers many challenges. Among them, the most glaring challenge is the so-called semantic gap [8]. Even in the situation where images are associated with some textual descriptions, this semantic gap is still present since they do not fully capture all the subtleties of the semantics of the images.

To address the semantic gap issue, we experimented with several image retrieval approaches on the WikipediaMM dataset. A retrieve engine was implemented in this participation, which consists of four components respectively for data pre-processing, text-based image retrieval (TBIR), content-based image retrieval (CBIR), and cross-media retrieval. In TBIR, textual query expansion technique is used where the extension terms are automatically selected from a knowledge base (KB) that is semi-automatically constructed from the online large-scale encyclopedia — Wikipedia.

* Corresponding author.

C. Peters et al. (Eds.): CLEF 2008, LNCS 5706, pp. 763–770, 2009.

Encouragingly, the experimental results rank in the first place among all submitted runs. For CBIR, visual query expansion is employed through query-dependent visual concept detection to semantically annotate images or augment their rough semantics gathered from related text. By comparison, this approach performs better than the other submitted CBIR runs. Then cross-media retrieval is performed by independently applying the two meta-search tools and then combining the results through a weighted summation of scores. Though not submitted, this approach outperforms our text-based or content-based approaches remarkably.

The rest of this paper is organized as follows. Textual and visual query expansion approaches for two meta-search tools are described respectively in Section 2 and 3. Then the cross-media re-ranking approach is presented in Section 4. The experimental results are shown in Section 5. Finally we draw a conclusion in Section 6.

2 Textual Query Expansion for TBIR

A natural solution for WikipediaMM 2008 task is to use TBIR method. To help the retrieval system get close to users' real intent, query expansion techniques are often used by adding terms to queries or modifying preliminary queries. In this participation, we focus on how to automatically extract the expansion terms from a KB that is semi-automatically constructed from Wikipedia. Organized with concepts identified by URLs and links between concepts and external nodes, Wikipedia is not only a Web collection but also an online knowledge center which assembles all users' intelligences. Therefore, it is naturally attractive and promising that this open, and constantly evolving encyclopedia can yield inexpensive knowledge structures that can be exploited to enhance the semantics of queries.

Recently, "Wikipedia mining" has been addressed as a new research topic. WikiRelate [2] used link-based path length for computing relatedness for given concepts; Nakayama *et al.* [3] proposed the PFIBF (Path Frequency – Inversed Backward link Frequency) algorithm for Web thesaurus construction. However, none of work is made on using Wikipedia as the KB in information retrieval.

In Wikipedia, each non-administrative page is used as a term/concept describing individuals (e.g., *Jingtao Hu*), concepts (e.g., *Emissions trading*), locations (e.g., *Big Ben*), events (e.g., *collapse of the World Trade Center*), and categories (e.g., *microbiology*). For a given term, the related terms can be easily extracted from the corresponding Wikipedia pages, and then used to extend the query when this term is used as the query input. Finally, the extended query is fed into the retrieval engine to generate the final search results. In our implementation, we use the TF-IDF paradigm for text retrieval which has been widely used in text mining and information retrieval.

As shown in Fig. 1, three steps are used to construct the KB from Wikipedia:

(1) ***Near Pages Selection.*** We first download and index all Wikipedia pages with TF-IDF model. Only pages with a similarity score higher than threshold θ (θ is set to be 0.9 in our experiments) are chosen as the related pages of the input query.

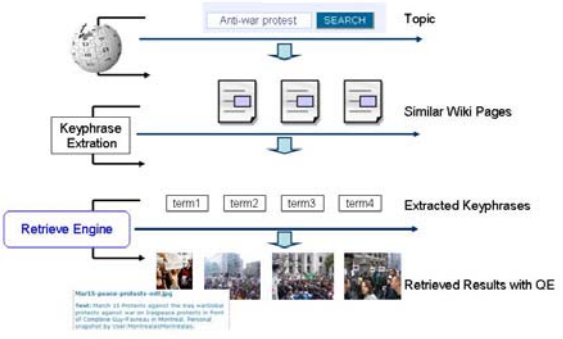


Fig. 1. Textual query expansion using the KB constructed from Wikipedia

(2) ***Page Keyphrase Extraction.*** In a Wikipeida page, keyphrases or keywords briefly describe the content of a concept. Thus they can be used to enhance the semantics of that concept. In our system, we employ an unsupervised keyphrase extraction algorithm presented in our previous work [4]. By treating text in a page as a semantic network, this algorithm computes several structure variables of Small-World Network (SWN) to select key nodes as keyphrases $K = \{(t_k, P(t_k))\}$, each with a probability score $P(t_k)$ indicating the importance of the extracted keyphrase t_k.

(3) ***Term Selection for Query Expansion.*** In practice, the top-ranked keyphrases cannot be directly used for query expansion. For instance, when searching "*saturn*", term "*moon*" is extracted as the keyphrase with a high score, but "*moon*" may appear on many pages and should be considered more *general*. To address this problem, a statistical feature *Inverse Backward link Frequency* (*ibf*) [3] is calculated as:

$$ibf = \log(\frac{N}{bf(t) + \beta}), \tag{1}$$

where $bf(t)$ is the number of backward links in which the link text contains term *t*, *N* denotes the total number of pages and β is a parameter in case $bf(t)$ is zero. Therefore, the final weight of a keyphrase can be computed as:

$$w_{t_k \in K} = P(t_k) \cdot ibf(t_k). \tag{2}$$

Then the keyphrases with their normalized weights are combined with the original query to construct an extended query to be fed into the retrieve engine.

3 Query-Dependent Visual Concept Detection for CBIR

In the WikipediaMM dataset, some images have few or even no descriptive texts. To address this problem, *query-dependent visual concept detection* can be used to semantically annotate images or augment their rough semantics gathered from related text. Given the pre-defined query concepts, *1-vs-all* visual concept detectors are trained for

all these concepts by using the training images obtained by Yahoo! search. Clearly, these training images can be used for visual query expansion to enhance the CBIR task. As shown in Fig 2, the training process includes the following three steps:

(1) ***Building the training set.*** For each query concept, top k (k=30 in our system) images are clawed from Yahoo! image search engine. Then some unrelated images with respect to the concept are manually filtered out whichforms a positive training set. Negative images for each concept are randomly selected from positive images of the other concepts.

(2) ***Building Bag of Words (BOW) representation.*** SIFT [5], Dense-SIFT [6] and Color-Dense-SIFT are extracted from the training sets of all concepts. Then k-means algorithm is employed to quantize different types of features and create a combined visual codebook. All images are represented by a set of tokens of the visual words.

(3) ***Supervised training for each topic.*** Unsupervised probabilistic latent semantic analysis (pLSA) [7] is utilized to infer the latent topic distribution of the training images based on the BOW representation. Then support vector machine (SVM) is used to train a *one-class classifier* for each concept in the latent topic space.

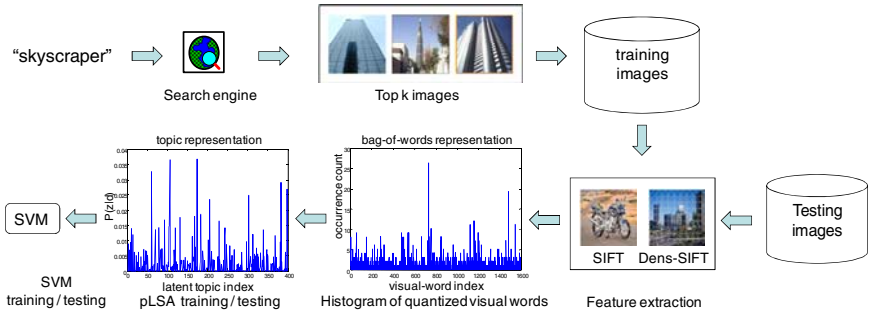


Fig. 2. Query-dependent visual concept detection for CBIR

Given the trained *1-vs-all* visual concept detectors for all query topics, we can perform the concept detection for each test image by firstly representing it with the visual words from the trained codebook, inferring its latent topic distribution based on the trained pLSA model, and finally computing the responds of the trained SVMs for different concepts. Concept is detected only when the corresponding respond is above a given threshold. For CBIR, test images are finally ranked according to their responds with respect to the query concept.

4 Query-Independent Cross-Media Re-ranking

For better retrieval performance, we study cross-media image retrieval by combining both TBIR and CBIR methodologies. In our implementation, cross-media retrieval is performed by independently applying the two meta-search tools and then combining

the results through a weighted summation of scores. Here the weights are query-independent, say, identical for all queries. Then the re-ranking score is computed as:

$$WeightedScore(q,d) = w_1 * Score(q_{text}, d_{text}) + w_2 * Score(q_{visual}, d_{visual}) \quad (3)$$

A key point here is to compare the overlap of the results returned by different retrieve engines. Let $\mathbf{R}_1$ and $\mathbf{R}_2$ respectively denote the result sets of TBIR-based and CBIR-based retrieval engines, and M_1 and M_2 be their sizes. Let image $d_i^1 \in \mathbf{R}_1, i < M_1$, and $d_j^2 \in \mathbf{R}_2, j < M_2$, then an overlap set $\mathbf{G}$ can be obtained:

$$\mathbf{G} = \{(d_i^1, d_j^2) : d_i^1 = d_j^2, i < M_1, j < M_2\}, \quad (4)$$

where (d_i^1, d_j^2) stands for an image both returned by the two engines. Let H_1 and H_2 be the numbers of overlap images in Top N ranked images, $H_1 = \#\{d_i^1 : d_i^1 \in \mathbf{G}, i < N\}, H_2 = \#\{d_j^2 : d_j^2 \in \mathbf{G}, j < N\}$, then the weight of each engine can be calculated as:

$$w_l = \frac{\sigma/2 + H_l/N}{\sigma + \sum_l H_l/N}, \quad (5)$$

where l is the engine identifier and σ (we set $\sigma = 0.1$) is an adjusting parameter.

5 Experiments

This section describes our experiments for the WikipediaMM task. Note that some of the experimental results reported here were not submitted before the deadline.

The experiments are evaluated by *MAP* (Mean Average Precision), *P@N* (precision of top *N* images), and *R-precision*. The ground-truth results are given in the evaluation phase of the WikipediaMM task.

5.1 Experiments with TBIR

The first set of experiments is to evaluate the performance of TBIR approach with different query expansion methods.

Query expansion by using the automatically-constructed KB. Different methods are used to automatically construct KB from Wikipedia for query expansion, by using different text sources (e.g., titles, links or fulltext of Wikipedia articles) and different term selection algorithms (e.g., TFIDF-based, Small-World (SW)-based, SWIBF-based). Therefore, four *automatic query expansion* methods were evaluated in our experiments, respectively denoted by *QE-Title-TFIDF*, *QE-Link-TFIDF*, *QE-Fulltext-SW*, and *QE-Fulltext-SWIBF*. We also use *NO-QE* to denote TBIR without query expansion. In all experiments, only top 20 terms are used.

Surprisingly, all these automatic query expansion methods can not significantly improve the TBIR performance, compared with *NO-QE* (See Table 1). Thus we should consider how to improve the quality of the constructed KB.

Query expansion by using the semi-automatically-constructed KB. After the KB was automatically constructed from Wikipedia, we then performed some manual confirmations. Here we use *QE-Fulltext-Semi* to denote this query expansion method. Note that in this case, the query expansion method still automatically selects terms from the KB to semantically expand a given query term. From Table 1, we can see that this *QE-Fulltext-Semi* method performs much better than all other models.

Table 1. The experimental results of different textual query expansion methods

Run ID	QE	Modality	MAP	P@5	P@10	R-Prec
NO-QE	without	TXT	0.2565	**0.4427**	**0.3747**	0.2929
QE-Title-TFIDF	with	TXT	0.2566	0.4187	0.3627	**0.2967**
QE-Link-TFIDF	with	TXT	0.2271	0.376	0.3147	0.2533
QE-Fulltext-SW	with	TXT	0.2365	0.3733	0.336	0.2618
QE- Fulltext-SWIBF	with	TXT	**0.2609**	0.44	0.3693	0.2859
QE- Fulltext-SEMI	with	TXT	**0.3444**	**0.5733**	**0.476**	**0.3794**

5.2 Experiments with CBIR

Compared with TBIR, our CBIR obtained a comparable precision in the top-ranked images (P@5=0.5307 and P@10= 0.4507 of CBIR *vs.* P@5=0. 5733 and P@10=0.476 of TBIR), but much lower MAP (0.1928 of CBIR *vs.* 0.3444 of TBIR) and R-Prec (0.2295 of CBIR *vs.* 0.3794 of TBIR). Although visual content ambiguity reduces the overall performance (MAP) by returning images with similar low-level features, the experimental results show that learning visual models from Web images (e.g., from Yahoo! search) do help to rank the content-relevant images higher. It also should be noted that, our CBIR approach performs best among all submitted CBIR runs in WikipediaMM 2008 task.

Table 2. The experimental results of CBIR

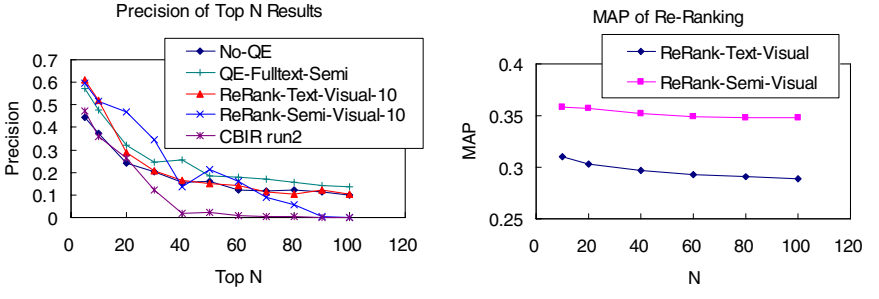
Run ID	QE	Modality	MAP	P@5	P@10	R-Prec
CBIR run1	with	IMG	0.1912	**0.5333**	0.4427	**0.2929**
CBIR run2	with	IMG	**0.1928**	0.5307	**0.4507**	0.2295

5.3 Experiments with Cross-Media Retrieval

In the last set of experiments, cross-media retrieval approach is used to achieve better performance by combining text-based and content-based retrieval results. In the experiments, we set M_2 smaller than M_1. This means that only the *top-ranked* images returned by CBIR are included in the re-ranking phase since the lower-ranked images may have much higher probabilities to be noises. Table 3 shows the experimental results, where *ReRank-Text-Visual-N* denotes the combination of CBIR and TBIR without query expansion, and *ReRank-Semi-Visual-N* denotes the combination of CBIR and TBIR with semi-automatic query expansion, and *N* denotes the corresponding parameter in Eq. (5).

Table 3. Some experimental results of cross-media retrieval

Run ID	QE	Modality	MAP	P@5	P@10	R-Prec
NO-QE	without	TXT	0.2565	0.4427	0.3747	0.2929
CBIR run2	without	IMG	**0.1928**	0.5307	**0.4507**	0.2295
ReRank-Text-Visual-10	without	TXTIMG	**0.3099**	**0.608**	**0.5213**	0.3387
ReRank-Text-Visual-20	without	TXTIMG	0.3035	0.6027	0.512	**0.3420**
ReRank-Text-Visual-40	without	TXTIMG	0.2972	0.584	0.4893	0.3393
ReRank-Text-Visual-60	without	TXTIMG	0.2928	0.5547	0.4733	0.3366
ReRank-Text-Visual-80	without	TXTIMG	0.2910	0.5387	0.4693	0.3349
QE- Fulltext-SEMI	with	TXT	0.3444	0.5733	0.476	0.3794
CBIR run2	with	IMG	0.1928	0.5307	0.4507	0.2295
ReRank-Semi-Visual-10	with	TXTIMG	**0.3584**	**0.6293**	**0.5147**	**0.3993**
ReRank-Semi-Visual-20	with	TXTIMG	0.3568	0.6187	0.5147	0.3974
ReRank-Semi-Visual-40	with	TXTIMG	0.3519	0.5867	0.5013	0.3988
ReRank-Semi-Visual-60	with	TXTIMG	0.3487	0.568	0.492	0.3988
ReRank-Semi-Visual-80	with	TXTIMG	0.3483	0.5653	0.4907	0.3988



(a) (b)

Fig. 3. Performance of cross-media retrieval: (a) P@N and (b) MAP results with different values of *N* in Eq. (5)

From Table 3 and Fig. 3, it's interesting to find that when *N* increases, the preliminary result of each system is more likely to be equally treated and the overall performance decreases. For the combination of CBIR and text-based retrieval without query expansion, the average improvement of all the queries in *ReRank-Text-Visual-10* is around 5.34% over the single text-based retrieval approach (25.65% of MAP). While for the combination of CBIR and text-based retrieval with semi-automatic query expansion, the average improvement for all the queries in *ReRank-Semi-Visual-10* is around 1.4% over the single text-based retrieval approach (34.44% of MAP).

We also observed that the cross-media retrieval results have much higher precision of top-ranked images than both text-based retrieval or CBIR results. Generally speaking, text-based retrieval can return more relevant images by searching keywords with image descriptions, while CBIR can obtain high precision of top-ranked images but too many noises in lower-ranked images. Thus combining CBIR with text-based retrieval can help increase the precision of top-ranked images.

In conclusion, the cross-media retrieval approach performs remarkably well. This indicates that cross-media fusion is definitely a promising direction to investigate effective retrieval approaches in the context of a large-scale and heterogeneous collection of images.

6 Conclusion and Future Work

This paper reported our approaches for the WikipediaMM task at ImageCLEF 2008. We experimented with TBIR, CBIR and cross-media image retrieval approaches with query expansion. Encouragingly, the experimental results of our TBIR approach rank in the first place among all submitted runs. Despite not submitted, the cross-media approach performs much better than the single TBIR or CBIR approaches. Further experiments will be done by optimizing the KB construction procedure and taking better cross-media re-ranking strategies into account.

Acknowledgement

The work is supported by grants from Chinese NSF under contract No. 60605020 and No. 90820003, National Hi-Tech R&D Program (863) of China under contract No. 2006AA01Z320 and No. 2006AA010105, and National Basic Research Program of China under contract No. 2009CB320906.

References

1. Tian, Y.H., Huang, T.J., Gao, W.: Exploiting multi-context analysis in semantic image classification. J. Zhejiang Univ. SCI. 6A(11), 1268–1283 (2005)
2. Strube, M., Ponzetto, S.: WikiRelate! Computing semantic relatedness using Wikipedia. In: Proc. of National Conference on Artificial Intelligence (AAAI 2006), Boston, Mass, pp. 1419–1424 (2006)
3. Nakayama, K., Hara, T., Nishio, S.: A thesaurus construction method from large scale web dictionaries. In: Proc. of IEEE International Conference on Advanced Information Networking and Applications (AINA 2007), pp. 932–939 (2007)
4. Huang, C., Tian, Y.H., Zhou, Z., Ling, C.X., Huang, T.J.: Keyphrase extraction using Semantic Networks Structure Analysis. In: Proc. of the sixth IEEE Int'l. Conf. on Data Mining (ICDM 2006), pp. 275–284. IEEE press, Hong Kong (2006)
5. Lowe, D.: Object recognition from local scale-invariant feature. In: Proc. Int'l Conf. Computer Vision (ICCV 1999), pp. 1150–1157 (1999)
6. Lazebnik, S., Schmid, C., Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In: Proceedings of the IEEE CVPR 2006, pp. 2169–2178 (2006)
7. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Machine Learning 41, 177–196 (2001)
8. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach Intell. 22(12), 1349–1380 (2000)