PKU at ImageCLEF 2008: Experiments with Query Extension Techniques for Text-Based and Content-Based Image Retrieval

Zhi Zhou[1,2,3], Yonghong Tian[1], Yuanning Li[1,2,3], Ting Liu[1,2,3], Tiejun Huang[1], Wen Gao[1]
[1]The Institute of Digital Media, School of EE & CS, Peking University, Beijing 100871, China
[2]Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Science, Beijing 100190, China
[3]Graduate University of Chinese Academy of Sciences, Beijing 100080, China
{zzhou, ynli, tliu}@jdl.ac.cn, {yhtian,tjhuang,wgao}@pku.edu.cn

Abstract.
In this paper, we present our solutions for the WikipediaMM task at ImageCLEF 2008. The aim of this task is to investigate effective retrieval approaches in the context of a large-scale and heterogeneous collection of Wikipedia images that are searched by textual queries (and/or sample images and/or concepts) describing a user's information need. We first experimented with a text-based image retrieval approach with query extension, where the expansion terms are automatically selected from a knowledge base that is (semi-)automatically constructed from Wikipedia. We show how this open, constantly evolving encyclopedia can yield inexpensive knowledge structures that are specifically tailored to effectively enhance the semantics of queries. Encouragingly, the experimental results rank in the first place among all submitted runs. The second approach we experimented with is content-based image retrieval (CBIR), in which we first train *1-vs-all* classifiers for all query concepts by using the training images obtained by Yahoo! search, and then treat the retrieval task as visual concept detection in the given Wikipedia image set. By comparison, this approach performs better than other submitted CBIR runs. Finally, we experimented with a cross-media image retrieval approach by combining and re-ranking text-based and content-based retrieval results. Despite the final experimental results were not formally submitted before the deadline, this approach performs remarkably better than the text-based retrieval or CBIR approaches.

1. Introduction

ImageCLEF 2008 contains five different tasks (i.e., photo retrieval, medical retrieval, visual concept detection, medical annotation, and WikipediaMM). In this paper we present our efforts in the WikipediaMM task of ImageCLEF 2008. This is our first year at ImageCLEF and the WikipediaMM task we participated in is also offered for the first time. We participated in all steps of the task, including topic creation, retrieval experiments, and relevance assessment.

The aim of WikipediaMM 2008 task is to investigate effective retrieval approaches in the context of a large-scale and heterogeneous collection of Wikipedia images that are searched by textual queries (and/or sample images and/or concepts) describing a user's information need. Towards this end, WikipediaMM 2008 task needs to deal with searching 75 topics from approximately 150,000 images in the Wikipedia collection. Roughly speaking, there are three challenges that the task participants must deal with:

- ***Scalability***. A good retrieval approach should perform well on such a large image set.
- ***Robustness.*** A natural, and also maybe the most effective solution for this task is text-based image retrieval approach since each image in this dataset is associated with user-generated alphanumeric, unstructured metadata (e.g., a brief caption or description of the image). However, text-based retrieval methods should be robust to noisy textual description that many images may be annotated with.
- ***Multi-modal Fusion.*** Considering the fact that some images in this dataset have few or even no descriptive

texts, it may be more preferable to combine text-based and content-based image retrieval (CBIR) for better performance. However, a pure combination of traditional text-based and content-based approaches is not adequate for dealing with the problem of Wikipedia image retrieval [1]. Therefore, how to effectively exploit the correlation between different modalities of retrieval clues remains challenging.

In this participation, we carry out three kinds of experiments. Our key idea is to experiment with different query extension techniques to help the retrieval system get close to users' real intent. In general, the oft-used approach of query extension is to add terms to queries or modify preliminary queries. In our experiments, query extension techniques are used in different situations. In the case of text-based image retrieval, the expansion terms are automatically selected from a knowledge base that is (semi-)automatically constructed from Wikipedia. This open, constantly evolving encyclopedia can yield inexpensive knowledge structures that are specifically tailored to effectively enhance the semantics of queries. Encouragingly, the experimental results rank in the first place among all submitted runs. In the case of CBIR, *1-vs-all* classifiers are trained for all query concepts by using the training images obtained by Yahoo! search. Then the retrieval task can be implemented as visual concept detection in the given Wikipedia image set. Clearly, the training images obtained by Yahoo! search are used to enhance the image retrieval task. In the last case, we experimented with a cross-media image retrieval approach by combining and re-ranking text-based and content-based retrieval results. Here the CBIR results are used to modify preliminary text-based retrieval results. Despite the final experimental results were not formally submitted before the deadline, this approach performs remarkably better than the single text-based retrieval or CBIR approaches.

The rest of this paper is organized as follows. First of all, we present the architecture of our system for the WikipediaMM task in section 2. Then three different approaches are described respectively in the following three sections. The experimental results of our approaches are presented in section 6. Finally we draw a conclusion and propose our future work.

2. System Architecture

In our system, we implemented a retrieve engine as a test environment for all methods. Fig.1 shows the architecture of our system, illustrating how different components works together to generate retrieval results. The components include:

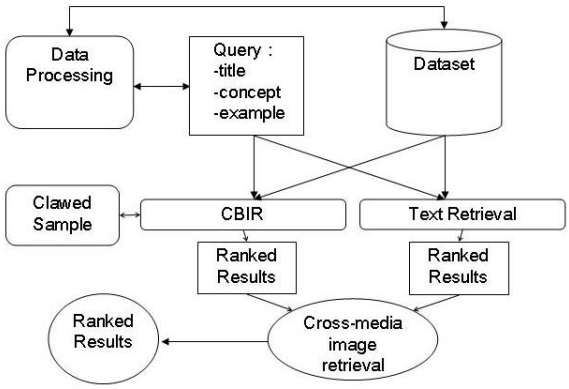


Fig.1. The architecture of our system for WikipediaMM 2008 task

(1) ***Data Processing Module***: Processing unit that performs several pre-processing tasks for the queries and the dataset.

(2) ***Text Retrieval Module***: Retrieval subsystem that searches the dataset with textual queries and returns relevant images.

(3) ***CBIR Module***: Content-based image retrieval subsystem that searches the dataset with visual features and

returns relevant images.

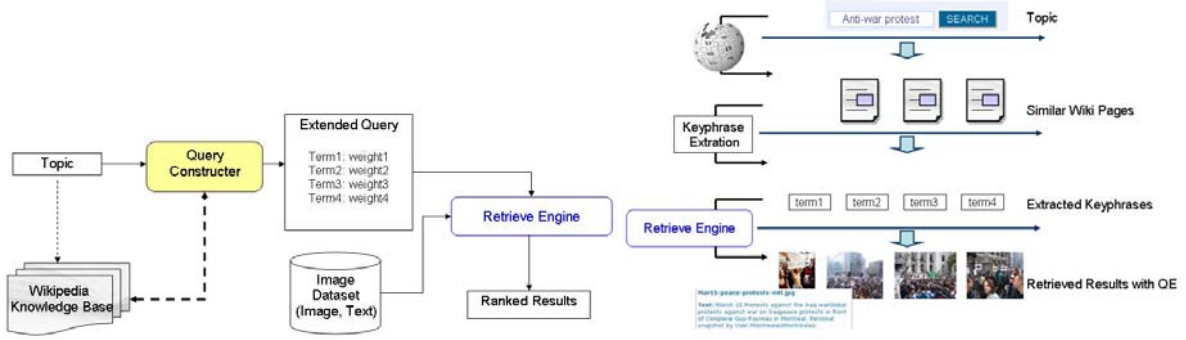
(4) ***Cross-Media Re-Ranking Module***: Processing unit that combines the sets of returned images from CBIR and text-based retrieval modules, and then performs cross-media re-ranking to obtain the final retrieval results.

3. Query Extension for Text Retrieval using Knowledge from Wikipedia

A natural solution for WikipediaMM 2008 task is text-based image retrieval method. To help the retrieval system get close to users' real intent, query extension 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 knowledge base that is (semi-)automatically constructed from Wikipedia. Organized with concept identified by URL (one user defined concept refers a single page) 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. Thereby, it is naturally attractive and promising that this open, and constantly evolving encyclopedia can yield inexpensive knowledge structures that are specifically tailored to effectively enhance the semantics of queries.

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

Fig. 2 shows the system framework and data flow of our text-based retrieval approach. As mentioned above, we first construct a knowledge base from Wikipedia pages for query extension. Specifically, each non-administrative Wikipedia 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 page. Thus given a textual query (and/or concept), the query constructer searches the knowledge base with the given query term and then extends the query with extracted terms. Finally, the extended query is given to the retrieval engine to generate the final search results.



(a) (b)

Fig.2. Text-based retrieval with query extension. (a) System framework, and (b) Data flow.

3.1 Knowledge Base Construction from Wikipedia for Query Extension

In our system, the construction of the knowledge base from Wikipedia includes the following steps:

(1) Near Pages Selection

We first index all titles of Wikipedia articles[1], and then retrieve the 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

[1] The Wikipedia articles and other related sources can be downloaded from http://download.wikipedia.org.

related pages of the input query.

(2) Page Keyphrase Extraction

Keyphrases play a key role in the Wikipedia knowledge base construction. In a Wikipeida page, keyphrases or keywords briefly describe the content of the concept. Thus they can be used to enhance the semantics of that concept. When facing roughly 2,480,000 Wikipedia pages, we are motivated to summarize concepts and measure the concept relatedness. Most existing keyphrase extraction algorithms, such as KEA [4], are supervised learning methods which require human labeled training sets. However, it's laborious to build such an appropriate training set. Moreover, Wikipedia has various lengths of pages with complicated structure. Therefore, unsupervised keyphrase extraction method is more preferable.

In our system, we employ an unsupervised keyphrase extraction algorithm presented in our previous work [5]. This algorithm uses several set-independent feature weights, treating text in a page as a semantic network. Several structure variables of Small-World Network (SWN) are used to select key nodes from the network 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 Extension

However, the top-ranked keyphrases can not be directly added for query extension. For instance, when searching "saturn", term "moon" may be extracted as keyphrase with a high score, but "moon" can appear on many pages and should be considered more *general*. To address this problem, a statistical feature *Inverse Backward link Frequency* (ibf) is calculated as:

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

where $bf(t)$ is the number of backward links in which the link text contains term *t*, *N* denotes the total number of articles and β is a parameter in case $bf(t)$ is zero.

Therefore, the final weight of a keyphrase can be computed as:

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

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.

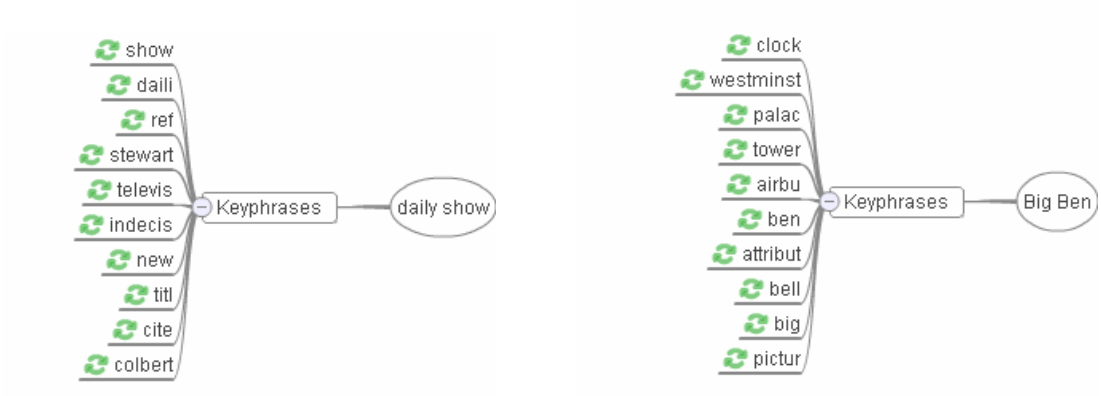


Fig. 3 Examples for topic "daily show" and topic "Big Ben" in the knowledge base

For each topic, we can extract a knowledge tree which consists of the given topic, and keyphrases extracted from Wikipedia pages. Moreover, this knowledge tree can be pre-constructed or online constructed. By treating each query as a topic, the text-based retrieval system finds the topic in the knowledge base, and uses the weighted keyphrases as query extension to enhance the system performance. It should be noted that if no relevant keyphrases can be found in the base with respect to the query, it is very easy to perform the above steps online.

3.2 TF-IDF Model for Text Retrieval

For the retrieve engine, we use the TF-IDF paradigm which is widely used in text mining and information retrieval. It is defined as follows:

$$w(t,d) = TF(t,d) \times \ln \frac{|d|}{DF(t)}$$

where *TF* stands for the frequency of term *t* in document *d*, *DF* for the document frequency of term *t* in the dataset. This traditional method represents the documents by vectors with the TF-IDF value as terms' weights.

In our text retrieval system, the score of query *q* and document *d* correlates to the dot-product between document and query vectors in VSM. Documents with larger scores are ranked higher in result sets. Then the scoring function is defined as follows:

$$Score(q,d) = \sum_{t \text{ in } q} w(t,q) * TF(t,d) * IDF(t) \qquad (3)$$

4. Content-Based Image Classification and Retrieval

In many real-world image datasets such as Wikipedia image collection, there are some images have few or even no descriptive texts. To address this problem, content-based image retrieval (CBIR) is used in our system. Given the pre-defined query concepts, *1-vs-all* classifiers are trained for all these concepts by using the training images obtained by Yahoo! search. Then the retrieval task can be treated as visual concept detection in the given Wikipedia image set. Fig is a summary of our approach. Clearly, the training images obtained by Yahoo! search are used for enhancing the image retrieval task.

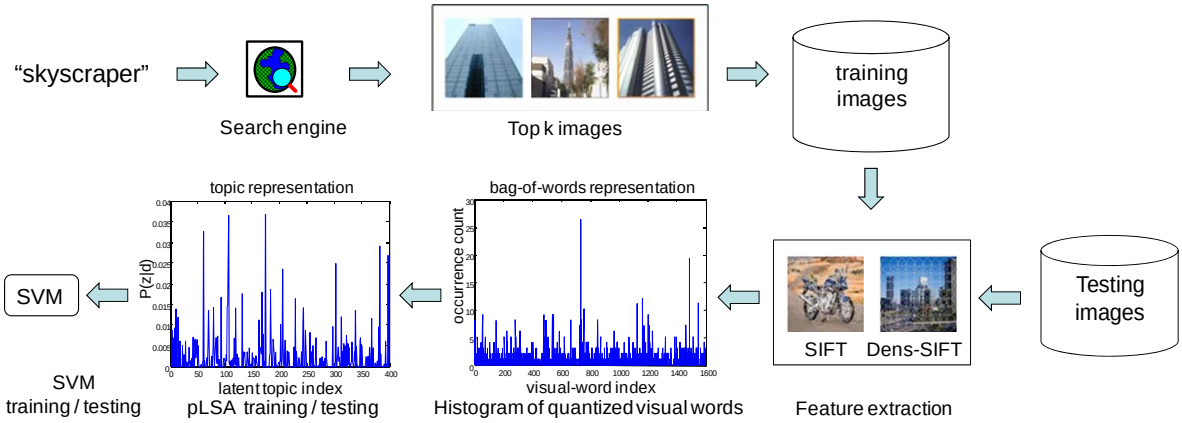


Fig. 4 A summary of our approach. Given the keyword for query concept "skyscraper", we first collect the top K Yahoo's search images to form the training set for concept learning. After feature extraction, quantization and pLSA modeling, a 1-vs-all SVM classifier is learned for the concept "skyscraper". During test stage, testing images from Wikipedia follows a similar process with the training stage, but uses the trained pLSA and SVM model for test. The results of the classifier are going to be used for image ranking for concept retrieval.

In our implementation, the training process includes the following three steps:

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

(2) ***Building Bag of Words (BOW) representation*****:** SIFT (or Scale-invariant feature transform [6]),

Dense-SIFT[9] and Color-Dense-SIFT are extracted from the training sets of all concepts. Then *k*-means algorithm is employed to quantized different types of features, forming a combined visual codebook for three types of features. 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) [10]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.

Given the trained *1-vs-all* classifiers for all query topics, the testing process includes the following three steps:

(1) ***Representing Wikipedia images with BOW***: After feature extraction and quantization, each test image is represented by the visual words from the trained codebook.

(2) ***Inferring latent topic distribution of test images***: Based on the trained pLSA model, we infer the latent topic distribution of the test imagse.

(3) ***Visual concept detection:*** For each test image, compute the responds of the trained SVMs for different concepts. Concept is detected only when the corresponding respond is above a given threshold. For a concept based retrieval, test images are finally ranked according to their responds with respect to the concept.

5. Cross-media Re-ranking with Text and Visual Content

For better retrieval performance, we study cross-media image retrieval by combining both text-based retrieval and CBIR methodologies. In our system, we use the re-ranking scheme to combine the retrieval results of the two engines. Given a query topic, either text-based retrieval or CBIR engine can return a set of result images. By combining the two result sets, the returned images are re-ranked with weighting factors

$$WeightedScore(q,d) = w_1 * Score(q_{text}, d_{text}) + w_2 * Score(q_{visual}, d_{visual}). \quad (4)$$

Here the key idea is to compare the overlap of the returned images between results returned by each engine. Let R^1 and R^2 respectively denote the returned result sets of text-based and CBIR-based retrieval engines, and M^1 and M^2 are their corresponding sizes. Let image $d_i^1 \in R^1, i < M^1$, and $d_j^2 \in R^2, j < M^2$, then we have a overlap set *G* of the two results:

$$G = \{(d_i^1, d_j^2) : d_i^1 = d_j^2, i < M^1, j < M^2\}, \quad (5)$$

where (d_i^1, d_j^2) stands for an image both returned by the two engines. The numbers of overlap images in Top *N* ranked images are computed as:

$$H_1 = \#\{d_i^1 : d_i^1 \in G, i < N\}, H_2 = \#\{d_j^2 : d_j^2 \in G, j < N\} \quad (6)$$

and the weight of each engine is calculated as:

$$w_l = \frac{\sigma / 2 + H_l / N}{\sigma + \sum_l H_l / N}, \quad (7)$$

where *l* is the engine identifier and σ is an adjusting parameter.

6. Experiments

In this section, we describe our experiments for WikipediaMM task at the ImageCLEF 2008. Note that some of experimental results reported in this section were not formally submitted before the deadline.

6.1 Experimental Methodology

Three sets of experiments were conducted. We first experimented with the text-based image retrieval approach with query extension. Text-based retrieval in different text source, with various query extension strategies were compared in the experiments. The second approach we experimented with is content-based image retrieval (CBIR). Finally, we experimented with the cross-media retrieval approach proposed in section 5.

In WikipediaMM task, the dataset consists of approximately 150,000 images and their text descriptions. The pre-processing of the dataset includes stop-words elimination and stemming. We developed a case-sensitive word splitter and maintained a stop-word list which takes into account frequent words of Wikipedia, such as "wikitemplate", "GFDL", "category". English words are stemmed by Porter Stemmer [7] which is used in the Apache Lucene Engine [8]. The text descriptions of the collection are indexed with Lucene once they are pre-processed. The TF-IDF model has been implemented as a baseline in our experiments.

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

6.2 Experimental Results with Text-Based Retrieval

The goal of the first set of experiments is to check how well different text-based retrieval methods with query extension perform.

Text-based retrieval in different text sources. Each image in the dataset has two text sources: image filename and image description. Thus text-based retrieval system can search in the two sources or their joint set. Table 1 shows the testing results on the given collection, and Fig. 5 depicts the curves of precision at top *N* results.

Table 1: The testing results of text-based retrieval in different text sources.

Run ID	QE	Modality	MAP	P@5	P@10	R-Prec
text-filename	without	TXT	0.256498	0.442667	0.374667	0.292887
text	without	TXT	0.21003	0.405333	0.346667	0.254963
filename	without	TXT	0.155436	0.322667	0.244	0.223775

We can see that, by combining filename text and description to a single search field, text-based retrieval can achieve a MAP of 25.65%. This is much better than the retrieval models with only filename or text description. Clearly, it's reasonable to include additional textual clues for retrieval.

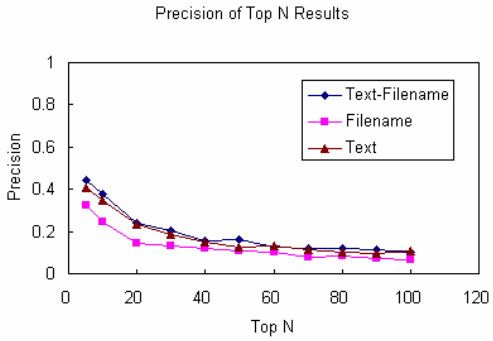


Fig. 5. Experimental Results of Precision at top N results of text-only retrieval in different text sources.

Text-based retrieval with query expansion. By using the approach described in section 3, we add further terms to a text query to effectively enhance its semantics (denoted by *QE-SW-20* and *QE-SWIBF-20*, respectively with different keyphrase extraction algorithms such as Small-World (SW) based algorithm and SW-IBF based algorithm). We also conducted experiments by using other query extension methods, such as query extension

with top *N* similar article titles in Wikipedia (denoted by QE-Title-*N*), and query extension with link text of *N* nearest pages in Wikipedia (denoted by QE-Link-*N*). For QE-Title-*N*, article titles are selected from Wikipedia with TF-IDF similarity measure. And for QE-Link-*N*, we implemented the lf-ibf algorithm (Link Frequency-Inverse Backward Link Frequency) proposed in [3]. Table 2 shows the testing results on the given collection, and Fig. 6 depicts the curves of precision at top N results.

Table 2: The testing results of text retrieval with query extension techniques.

Run ID	QE	Modality	MAP	P@5	P@10	R-Prec
QE-Title-20	with	TXT	0.256567	0.418667	0.362667	0.296697
QE-Title-10	with	TXT	0.255679	0.432	0.376	0.293633
QE-Link-20	with	TXT	0.227106	0.376	0.314667	0.253258
QE-Link-10	with	TXT	0.226688	0.368	0.318667	0.258306
QE-SW-20	with	TXT	0.236526	0.373333	0.336	0.261819
QE-SWIBF-20	with	TXT	0.260936	0.44	0.369333	0.285868

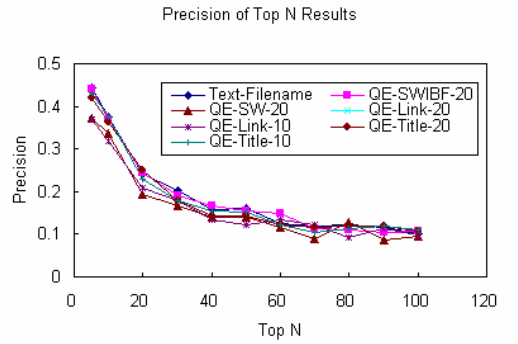


Fig. 6. Experimental Results of Precision at top N results of text-based retrieval with query expansion from the *automatically constructed* Wikipedia knowledge base.

In the experiments, we found that with a high threshold, the number of similar article titles is limited and there are few terms to be added to the query. From the experimental results shown in Table 2, we can see that using similar article titles in Wikipedia with respect to the queries directly does not help improving the performance of text-based retrieval system.

Link structure mining of Wikipedia can be used in Web thesaurus construction or query expansion. However, due to the complexity of Wikipedia hyperlink network, there were too many noises in these links, including Wikipedia system-generated links, and links that are rarely referenced. Therefore, we obtained even worse results than text-based retrieval without query extension.

For query expansion with the automatically constructed knowledge base, we found that the performance of the text-based retrieval system using keyphrases with just high probabilities is worse than text-based retrieval without any query extension. This means that keyphrases with general meanings have negative effects on retrieval results.

In contrast, by using the Keyword-IBF method, general keyphrases referenced by many articles are granted a lower weight in the extended query. As a consequence, this query extension method tends to choose terms that are more specific. The average improvement of all queries in this case is around 0.5% over text-based retrieval methods without query extension.

Text-based retrieval with query extension from the semi-automatically constructed knowledge base. Text-based image retrieval approach with query extension can perform better than traditional text retrieval

approach by adding expansion terms automatically selected from a knowledge base that is automatically constructed from Wikipedia. However, the retrieval performance of this approach depends on the quality of the constructed knowledge base. Due to the current restriction of keyphrase extraction algorithms and knowledge base construction models, the automatically constructed knowledge base is not good enough to practical applications in most cases. Therefore, we performed some manual confirmations and modifications to the automatically extracted knowledge base. Here we use *QE-Semi* to denote this query extension approach.

Table 3: The testing results of text retrieval query extension.

Run ID	QE	Modality	MAP	P@5	P@10	R-Prec
text-filename	without	TXT	0.256498	0.442667	0.374667	0.292887
QE-SWIBF-20	with	TXT	0.260936	0.44	0.369333	0.285868
QE-Semi	with	TXT	0.3444	0.5733	0.476	0.3794

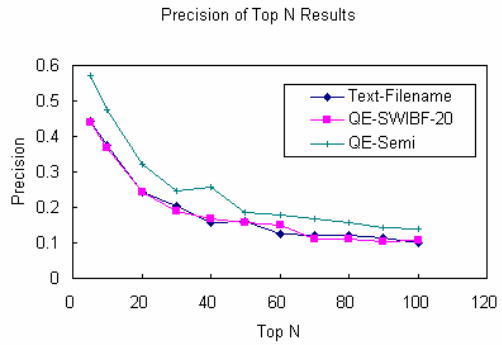


Fig. 7. Experimental Results of Precision at top N results of text-based retrieval with query extension from the *semi-automatically constructed* Wikipedia knowledge base.

By reasonably manual confirmation of the knowledge base, we obtained an obvious improvement of the retrieval results. From Table 3 and Fig. 7, we can see that this approach performs much better than all other models. This shows that query extension by using a knowledge construction procedure with a good keyphrase extraction algorithm and reasonably manual confirmation can remarkably improve the retrieval performance.

6.3 Experimental Results with CBIR

For content-based image retrieval, the experimental results show that our BOW methods perform better than other CBIR systems. Compared with the text-based systems, our CBIR obtained a comparable MAP (0.1912 of CBIR vs. 0.21003 of text-based retrieval), and higher precisions in the top-ranked images (P@5=0.5333 and P@10= 0.448 of CBIR vs. P@5=0.405333 and P@10=0.346667 of text-based retrieval). Although visual content ambiguity hampers 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 in higher places.

Table 4: The testing results of CBIR.

Run ID	QE	Modality	MAP	P@5	P@10	R-Prec
CBIR run1	with	IMG	0.1912	0.5333	0.442667	0.292887
CBIR run2	with	IMG	0.1928	0.5307	0.4507	0.2295

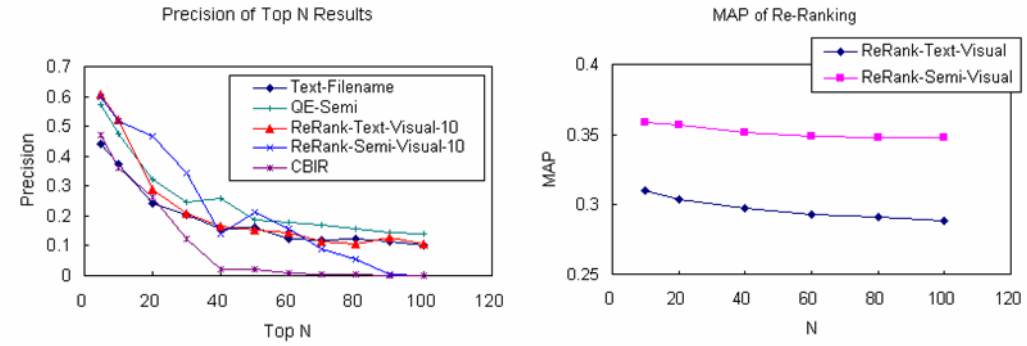
It also should be noted that, our CBIR approach performs best in all submitted CBIR runs in WikipediaMM 2008 task.

6.4 Experimental Results with Cross-media Re-ranking

In the last set of experiments, text-based and content-based image retrieval approaches are combined so as to achieve better performance. In the experiments, we set M^2 smaller than M^1. This means that only the *top-ranked* images returned by the CBIR approach are included in the re-ranking phase since the lower-ranked images may have much higher probabilities to be noises in the CBIR system. Table 5 shows the testing results on the given collection, where *ReRank-Text-Visual-N* denotes the combination of CBIR and text-based retrieval without query extension, and *ReRank-Semi-Visual-N* denotes the combination of CBIR and text-based retrieval with query extension from the semi-automatically constructed knowledge base, and *N* denotes the corresponding parameter in formula (6) and (7).

Table 5: The testing results of Cross-media Re-ranking.

Run ID	QE	Modality	MAP	P@5	P@10	R-Prec
ReRank-Text-Visual-10	with	TXTIMG	0.309876	0.608	0.521333	0.338713
ReRank-Text-Visual-20	with	TXTIMG	0.303529	0.602667	0.512	0.342036
ReRank-Text-Visual-40	with	TXTIMG	0.297208	0.584	0.489333	0.33931
ReRank-Text-Visual-60	with	TXTIMG	0.292776	0.554667	0.473333	0.336636
ReRank-Text-Visual-80	with	TXTIMG	0.290972	0.538667	0.469333	0.334891
ReRank-Text-Visual-100	with	TXTIMG	0.288602	0.522667	0.462667	0.335717
ReRank-Semi-Visual-10	with	TXTIMG	0.358431	0.629333	0.514667	0.399319
ReRank-Semi-Visual-20	with	TXTIMG	0.356829	0.618667	0.514667	0.397366
ReRank-Semi-Visual-40	with	TXTIMG	0.351918	0.586667	0.501333	0.39878
ReRank-Semi-Visual-60	with	TXTIMG	0.348719	0.568	0.492	0.39878
ReRank-Semi-Visual-80	with	TXTIMG	0.348332	0.565333	0.490667	0.39878
ReRank-Semi-Visual-100	with	TXTIMG	0.348259	0.557333	0.488	0.39878



(a) (b)

Fig. 8 Performance of cross-media re-ranking: (a) Precision at top N results of cross-media re-ranking. (b) MAP of re-ranking with different values of *N* in formula (6) and (7).

We used 6 values of *N* (*N*=10, 20, 40, 60, 80, 100) in formula (6) and (7). It's interesting to find that when *N* increases, the number of overlap images in top-*N* results of the two systems tends to be closer. In this case, the parameters w_1 and w_2 tends to 0.5 and the preliminary results of the two systems are more likely to be equally treated. As a consequence, the retrieval performance decreases when *N* increases. Therefore, we used *N*=10 in the following experiments.

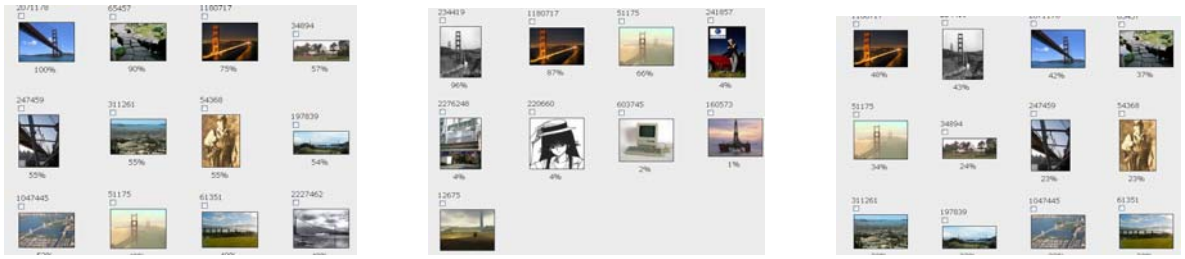
As shown in Table 5 and fig. 8, the re-ranking approach performs remarkably better than the single text-based or content-based retrieval approaches. For the combination of CBIR and text-based retrieval without query extension, the average improvement of all the queries in ReRank-Text-Visual-10 is around 5.34% over the single

text-based retrieval approach (25.6498% of MAP, as shown in Table 1). While for the combination of CBIR and text-based retrieval with query extension from the semi-automatically constructed knowledge base, the average improvement for all the queries in ReRank-Semi-Visual-10 is around 1.403% over the single text-based retrieval approach (34.44% of MAP, as shown in Table 3).

Fig. 9 shows two examples of the re-ranking results. For query “golden gate bridge”, text-based retrieval returns more relevant images than visual-based information retrieval. By re-ranking the two results, we get a better performance and a higher precision of top ranked images than the previous two methods. For query “Singapore”, visual-based retrieval does not perform well, however, combining CBIR result with text-based retrieval can still improve the system performance.

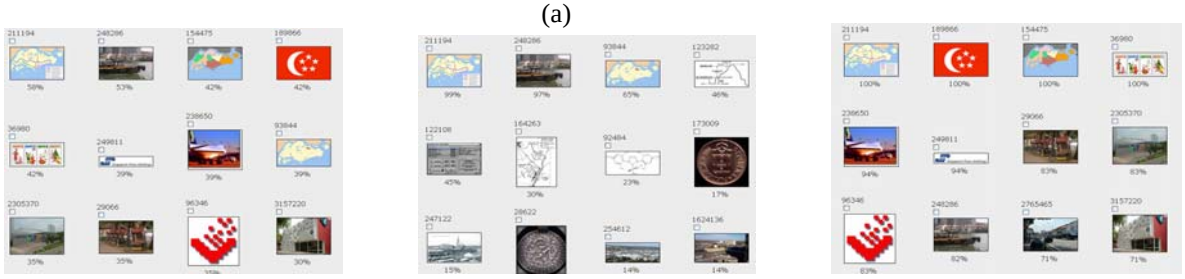
We also observed that the re-ranked results have higher precision of top-ranked images than results returned by text-based retrieval or CBIR. There are some possible reasons. Text-based retrieval can return more relevant images by searching keywords with image descriptions, while CBIR can obtain high 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 re-ranking 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.



Text-based Visual-based Cross-media Re-ranking

(a)



Text-based Visual-based Cross-media Re-ranking

(b)

Fig. 9 Search results for (a) topic “Golden Gate Bridge” and (b) topic “Singapore” using text-only (left), visual-only (middle) and their combination (right).

7. Conclusion and Future Work

In this paper, we reported our solutions for WikipediaMM task at ImageCLEF 2008. We experimented with text-based, CBIR and cross-media image retrieval approaches with query extension. Encouragingly, the experimental results of our text-based approach rank in the first place among all submitted runs. Moreover, our CBIR approach also performs better than other submitted CBIR runs. Despite the final experimental results were not formally submitted before the deadline, our cross-media approach performs remarkably better than the single

text-based or content-based retrieval approaches.

However, our experiments are just a first attempt towards effective cross-media retrieval in the context of a large-scale and heterogeneous collection of images. The query extension techniques can still be improved. Specifically, our knowledge base construction procedure strongly depends on the keyphrase extraction algorithm. Thus how to more effectively extract concepts and their relationships from Wikipedia is still a challenging future work.

Acknowledgement

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

Reference

[1] Y.H. Tian, T.J. Huang, and W. Gao, "Exploiting multi-context analysis in semantic image classification," *J. Zhejiang Univ. SCI.*, vol. 6A, no. 11, pp. 1268-1283, 2005.

[2] M. Strube and S. Ponzetto. "WikiRelate! Computing semantic relatedness using Wikipedia". *Proc. of National Conference on Artificial Intelligence* (*AAAI2006*), pages 1419-1424, Boston, Mass., July 2006.

[3] K. Nakayama, T. Hara, S. Nishio, "A thesaurus construction method from large scale web dictionaries", *Proc. of IEEE International Conference on Advanced Information Networking and Applications* (*AINA2007*), pp. 932-939, 2007

[4] I.H. Witten, G.W. Paynter, E. Frank, C. Gutwin, C.G. Nevill-Manning, "KEA: practical automatic keyphrase extraction". *Proc. of Fourth ACM Conference on Digital Libraries*, 1999.

[5] C. Huang, Y.H. Tian, Z. Zhou, Charles X. Ling, T.J. Huang, "Keyphrase extraction using Semantic Networks Structure Analysis", *Proc. of the sixth IEEE Int'l. Conf. on Data Mining* (*ICDM 2006*), Hong Kong, 2006, pp. 275-284, IEEE press.

[6] D. Lowe, "Object recognition from local scale-invariant feature," *Proc. Int'l Conf. Computer Vision* (*ICCV 1999*), pp. 1150-1157, Sep. 1999.

[7] M. Porter, The Porter Stemming Algorithm. (2005) http://www.tartarus.org/~martin/PorterStemmer

[8] Apache Lucene. Apache Software Foundation, http://lucene.apache.org/java/docs/index.html, 2005.

[9] S. Lazebnik, C. Schmid, J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories", Proceedings of the IEEE CVPR 2006, pp.2169 – 2178

[10] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis", Machine Learning 41(2001)177-196.