Experiment for Using Web Information to do Query and Document Expansion

Yih-Chen Chang and Hsin-Hsi Chen*

Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
E-mail: ycchang@nlg.csie.ntu.edu.tw; hhchen@csie.ntu.edu.tw

Abstract. ImageCLEF photo task of this year is a little different from those of previous years. The caption field in image annotations and the narrative field in the text queries are removed, and the visual queries (example images) are also removed from the image collection too. In the new definition, the information that can be employed for queries and images is less than before, so that it becomes harder to match query words and annotations directly. To deal with this issue, we explore the web to expand queries and documents. Many images and text information can be found in the web, but we should face the noise embedded. The experiment shows the query expansion improves performance about 16.11%. The document expansion brings too much noise and the performance decrease 28.24% after expansion. The media mapping method that we proposed in previous years is used for query expansion too. The results of formal runs show this method is still very useful in the new task definitions.

ACM Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval--- Information filtering, Relevance feedback

Keywords: Cross language image retrieval, cross-media translation, Query expansion, Document expansion

1. INTRODUCTION

Image retrieval is very important recently since more and more image data are available in the web. In imageCLEF photo tasks, many teams evaluate their image retrieval systems and share their experiences. In this task, each topic has a text query and a visual query, which simulate the information needs of users. In the previous years, a text query includes topic and narrative fields in several different languages and visual query include two or three example images in the image data set that are relevant to the topic. The image collection used in this task is about twenty thousand images, and each image is annotated with title, location, date, notes and a detail caption. This task has important changes in this year. First, the caption field in image annotations and narrative field in queries are removed. This change reflects the real world environment that image annotations and queries are usually short and rough. Second, the visual queries are not images in the image collection anymore. This change reflects that users may use their own photos as example images. It is not practical to find relevant images in image collection before using content-based search.

When the caption field in image annotation is removed, matching query words and image annotations becomes challenging than before. We will explore some other resources to expand query or document. The web provides a huge collection of data. Through web search engine (e.g., Google), we retrieve relevant images and text, and use them for expansion. Compared with query expansion of using pseudo relevant feedback in the corpus, outside resource like the web may bring in information that the target corpus may not have. At the same time, the information retrieved from the web may contain a lot of noises. How to filter out the noises is indispensable. In our work, we restrict the resource that we access to comparably pure web sites (e.g., Wikipedia).

Our method media mapping proposed in previous years achieves very good performance. Media mapping can be regarded as a kind of pseudo relevant feedback across different media for query expansion. We will employ media mapping method in this year, examine its performance in the new task definitions, and analyze if the web resource can bring in new information.

* Corresponding author

The paper is organized as follows. Section 2 introduces the methods we use in this year. Section 3 specifies each run we submitted. Section 4 shows and discusses the experiment results of this year. Finally, we conclude the remarks.

2. METHOD

2.1 Query Expansion

The query and images annotations are both short in this year. In this case we may want to expand our query and get more information. Many teams in last year have showed that query expansion using pseudo relevant feedback is very useful in this task. In this year we want to use out resource like web to do expansion and analysis if it can bring different information. When we want to use information of web pages the first problem is how to access them. The most easily way is using the web search engine like Google. When using a web search engine we need to submit a query to access the web pages. This query in most trivial way will be the textual query in the task. The language of textual query (e.g., Traditional Chinese) may different with the language of image annotations (e.g., English). If we submit the textual query directly, the language of web pages we access will different with the language of image annotations and can't use directly. There are at least two ways to solve this problem. Submit the textual query directly and do language translation to translate web pages we access into target language or do language translation before we submit textual query to the web search engine. In first way we need to translate all the web pages we get and may cost a lot of time so in this year we use the second way. But we will still let the first way a choice because second way have a problem that when there are name entity in textual query, if we translate before submit, we may have high chance to get a wrong translation of name entity. If the translation of name entity is wrong, we may get web pages that totally unrelated to the query.

After translating textual query into target language and submit it to the web search engine, we can access many web pages. The second problem is how to use these pages. We have two trivial direct to use these pages when doing query expansion. Using some methods to choose some words from web pages to expand query or using content of top ranked pages (or snippets) to expand query directly. If we use some method to choose words, we can filter noise information but we may also loss the useful information if we don't choose the right words. In this year, we just want to check if we can find useful information in web pages we access and analysis the difference with information that feedback method find, therefore how to choice words isn't so important in this year. We just use the top ranked snippets to expand our query. We expect after doing expansion the precision will decrease since we may bring in many noises. But for some query the recall may improve if we expand information that related to query but not mentioned in query.

In query expansion we also try a simple method to filter noise. We limit the website we access comes from cyclopedia-based website (e.g., Wikipedia). This experiment has two goals. First we expect this kind of web pages aren't have so much noise and compared with pages that come from general website may improve the precision. Second by comparing the recall we can know these kinds of website can bring more or less information than general web site. The method we limit the web site is simple. We just add web site name as a query word when we submit query to the web search engine. And we still use the top ranked snipes to expand our query.

2.2 Document Expansion

If we expand word that related to the query, but the documents don't mention it, it won't work. There are at lease two ways to solve this problem. First, we can only expand words that in the document. Query expansion that using relevant feedback is this kind. Second we may expand the documents. In this year we do the experiment to try the second way. The method we used in document expansion is very alike the method used in query expansion. We use title field in image annotation as query and submit it to web search engine to get top ranked snipes to expand document. Unlike query expansion, in document expansion we don't need to do language translation, since it's already in target language. Therefore it won't have translation error, but in our expecting the document expansion may be harder than query expansion. In document expansion we need to expand about twenty thousand documents, each time of expansion may bring in some noises. We don't know if this may bring in too much noise into our corpus after expanding twenty thousand documents. In document expansion we add one more limitation than query expansion. We only expand the words that nearby the words in document. We set a window size (in our experiment is 5), and only expanding words in the window. Beside of noise problem, document expansion has some special problems that different with query expansion when doing

image retrieval. In query expansion, expanding hyponym words is alright, but in document expansion it may cause problem. For example, in query expansion, if the query word is “animal” and we expand hyponym words “tiger”, ”cat”, and ”dog”, it’s alright since tiger, cat, and dog are some kinds of animal and are relate to query. But for document expansion, this may cause problem if we don’t know which animal are actually in that image. If the image is actually about “rabbit”, and we expand words “tiger”, ”cat”, and ”dog”, the words we expand are totally not related to the document and may cause problem for some query.

2.3 Media Mapping Method

Media mapping method was proposed in last year. The main idea of this method is using intermedia to translate visual query into textual query or vise versa. The intermedia is resource that links two kinds of media (e.g., text and visual). For example, an annotated image corpus can be seen as an intermedia that links text information and visual information. When we use image collection as intermedia, media mapping method can be seen as relevant feedback that cross the different media and can be used in query expansion. In last year, we don’t use media mapping to do query expansion but create a new query and merge the results of new query and original textual query to get the final result. In this year, since we want to analysis if query expansion using web resource can bring different information with relevant feedback, we redo the media mapping method and use it to do query expansion. And since the task definition is different with last year, we can also exam media mapping method in the new environment. In the new environment there are two new challenges when using media mapping method. First, mapping visual query into related images in the intermedia is become harder. In last year the visual queries are images in the image collection, when we use image collection as intermedia, we can always map visual query into itself and therefore find related image. Second, the caption field in the image annotation has been removed. The textual information we can get from visual information is become less than last year. We are very interested if media mapping method can still work in the new environment.

We use media mapping to do query expansion as following step. First we submit visual query to CBIR system and submit textual query to text retrieval system to retrieve images in intermedia (in this experiment intermedia is image collection in this task), than we rerank top ten images in visual query result using textual query result. After reranking, we map visual query to the top n images in the reranked result (in this year the n= 1) and expand image annotations of these images to the textual query. Now we have an expanded textual query and we can submit it to the text retrieval system to get final result. There are two differences with the steps we use in last year. First, in last year we don’t use textual query result to rerank visual query result. But since mapping visual query to the related image is harder in this year, we may use textual information to help visual query mapping to the related images that in the intermedia. Second, after mapping visual query into intermedia and translating visual query to textual words we don’t create a new query but expand these words to the textual query.

3. EXPERIMENT

This year we submit 27 official runs include 18 cross-lingual runs for eight different languages, 8 monolingual runs for three different languages, and 1 run for visual query only. All the queries with different source languages were translated into target language (e.g., English) using SYSTRAN system. We use Okapi with BM25 formula to do text retrieval.

In this year we consider about following issues. First we want to check web pages we access can bring new information when we use it to do query expansion. We will check the words we expand when the recall is improve. Second we will compare the results of query expansion runs that limit or not limit the web site. Third, we want to check what will happen when we do document expansion. The runs using both query expansion and document expansion will also be checked too. Fourth we will exam the performance of media mapping method, and we will do query expansion that using both media mapping and web pages and check if web pages can bring new information that media mapping doesn’t have. For our official runs some of the issues aren’t checked but checked in unofficial runs. Our official runs are described as follows:

(1) 8 cross-lingual runs that using textual query only and not doing query expansion:
NTU-ES-EN-AUTO-NOFB-TXT, NTU-FR-EN-AUTO-NOFB-TXT, NTU-RU-EN-AUTO-NOFB-TXT, NTU-PT-EN-AUTO-NOFB-TXT, NTU-JA-EN-AUTO-NOFB-TXT, NTU-IT-EN-AUTO-NOFB-TXT, NTU-ZHT-EN-AUTO-NOFB-TXT, NTU-ZHS-EN-AUTO-NOFB-TXT
They are regarded as baselines and are compared with runs using query expansion and document expansion.

(2) 3 monolingual runs that using textual query only and not doing query expansion:

NTU-EN-EN-AUTO-NOFB-TXT, NTU-ES-ES-AUTO-NOFB-TXT, NTU-DE-DE-AUTO-NOFB-TXT
These three runs serve as the baselines to compare with cross-lingual runs with textual query only, and to compare the runs using query expansion and document expansion.
(3) 3 monolingual runs, using media mapping method to do query expansion:
NTU-ES-ES-AUTO-FBQE-TXTIMG, NTU-EN-EN-AUTO-FBQE-TXTIMG,
NTU-DE-DE-AUTO-FBQE-TXTIMG
These runs using media mapping to do query expansion and will be compared with runs that doing query expansion using web resource and runs that not doing query expansion.
(4) 8 cross-lingual runs, using media mapping method to do query expansion:
NTU-PT-EN-AUTO-FBQE-TXTIMG, NTU-ES-EN-AUTO-FBQE-TXTIMG,
NTU-RU-EN-AUTO-FBQE-TXTIMG, NTU-IT-EN-AUTO-FBQE-TXTIMG,
NTU-ZHT-EN-AUTO-FBQE-TXTIMG, NTU-ZHS-EN-AUTO-FBQE-TXTIMG,
NTU-JA-EN-AUTO-FBQE-TXTIMG, NTU-FR-EN-AUTO-FBQE-TXTIMG
These runs using media mapping to do query expansion and will be compared with runs that doing query expansion with web resource and runs that not doing query expansion.
(5) 2 runs doing query expansion using web pages, not doing document expansion:
NTU-EN-EN-AUTO-QE-TXT-TOPIC, NTU-ZHT-EN-AUTO-QE-TXT-TOPIC
These two runs using web resource to do query expansion. They will compare with runs that not doing query expansion.
(6) 2 runs doing document expansion with web pages, not doing query expansion:
NTU-EN-EN-AUTO-DE-TXT-CAPTION, NTU-ZHT-EN-AUTO-DE-TXT-CAPTION
These two runs using web resource to do document expansion. They will compare with runs that not doing document expansion.
(7) 1 runs using visual query and media mapping method only:
NTU-IMG-EN-AUTO-FB-TXTIMG
This run using media mapping method to translate visual query into textual query and using text retrieval system to get the final result.

Our unofficial runs are described as follow:
(1) 2 runs doing both query expansion and document expansion:
NTU-EN-EN-AUTO-DE-QE-TXT, NTU-ZHT-EN-AUTO-DE-QE-TXT
These runs using both query expansion and document expansion. They will compare with runs that only use query expansion or document expansion or not use expansion.
(2) 2 runs doing query expansion using web pages and limiting web site that web pages come from:
NTU-EN-EN-AUTO-QE-WIKI-TXT, NTU-ZHT-EN-AUTO-QE-WIKI-TXT
These two runs dong query expansion using web resource, but the web site that web pages come from are limited to the cyclopedia-based website, Wikipedia. These runs will compare with runs do not have limit.
(3) 2 runs using both web resource and media mapping to do query expansion:
NTU-EN-EN-AUTO-QE-FBQE-TXTIMG, NTU-ZHT-EN-AUTO-QE-FBQE-TXTIMG
These two runs will be used to analysis if web resource can bring information that media mapping can't find. They will compare with runs using media mapping to do query expansion only.

4. Results and Discussions

In the experiment, first we want to check if web resource can bring in new information when we use it to do query expansion. We use the top one snipe that web search engine returned to expand our query. Table 1 show the results of runs NTU-EN-EN-AUTO-QE-TXT-TOPIC, NTU-ZHT-EN-AUTO-QE-TXT-TOPIC
, NTU-EN-EN-AUTO-NOFB-TXT, and NTU-ZHT-EN-AUTO-NOFB-TXT.

Table 1. Comparing results that using or not using query expansion

Query Language	MAP/RECALL	Query Expansion (Using Web)	Without Expansion
Traditional Chinese (Cross-lingual)	MAP	0.1225 (+16.11 %)	0.1055
	RECALL	0.4461 (+18.14 %)	0.3776
English (Monolingual)	MAP	0.1577 (+7.57 %)	0.1466
	RECALL	0.5439 (+14.84%)	0.4736

In the experiment we find after expansion both recall and precision are improved. In our expect precision will decrease since we don't do anything to filter noise. We are very interested to analysis what words we expand let the recall and precision improved. Table 2 shows some examples. The **bold** words in Table 2 are the words that appear in relevant images but not in the original query. These words may be the words that make the performance improve.

Table 2. Examples for the expanded words

Topic 22	Recall Changes After Expansion	MAP Changes After Expansion
Type: Monoligual	0.0408 → 0.8265	0.0037 → 0.2147
Original Text Query	tennis player during rally	
Expanding Words	The purpose of this research is to examine and discuss the characteristics of the serves of elite male tennis players during **singles** competition. The top sixteen **single** players of the CPC Tennis Competition in the Taipei were observed	
Topic 37	Recall Changes After Expansion	MAP Changes After Expansion
Type: Monoligual	0.4105 → 0.7894	0.0887 → 0.1684
Original Text Query	sights along the Inka-Trail	
Expanding Words	**Machu Picchu**, Inca Pachacutis Sacred City Bad weather and frequent breaks from filming allowed time for additional investigation at **Machu Picchu**. The location is well down Huayna **Picchu** Mountain just above the Urubamba River. Two Inca routes approach the site	
Topic 44	Recall Changes After Expansion	MAP Changes After Expansion
Type: Cross-ligual	0.4162 → 0.6324	0.0632 → 0.1470
Translated Text Query	In Australian mainland mountain	
Expanding Words	Mount Kosciuszko - Wikipedia, the free encyclopedia Mount Kosciuszko, located in the Snowy Mountains, in Kosciuszko **National Park**, is the highest mountain in mainland **Australia** at 2228 m above sea level. It was named by the Polish explorer Count Paul Strzelecki in 1840 in honour of the	
Topic 55	Recall Changes After Expansion	MAP Changes After Expansion
Type: Monoligual	0.2716 → 0.3209	0.0176 → 0.1099
Original Text Query	drawings in Peruvian deserts	
Expanding Words	**Nazca Lines** and Culture - Crystalinks Giant Figures in **Peru** Desert Pre-date **Nazca Lines**. **Nazca** Fisherman. The Epoch Times, May 24, 2005. A group of about 50 drawings of giant figures recently discovered in the hills of **Peru**'s southern coastal desert near the city of Palpa	

In Table 2 we can find the name entities we expand from the web resource looks very useful. This may become one filter for selecting words in the future work. We expand some words that aren't relating to the topic but the precision doesn't decrease. We think this may because these words aren't appearing in the documents and therefore won't influence the performance. The second issue we want to check in the experiment is to compare the results of query expansion runs that limit the web site and not limit. Table 3 shows the result.

Table3. Comparing results that limit or not limit web site

Runs Name / Query Language	Web Site Limit	Recall	MAP
NTU-ZHT-EN-AUTO-QE-TXT-TOPIC Traditional Chinese (Cross-lingual)	No	0.4461	0.1225
NTU-ZHT-EN-AUTO-QE-WIKI-TXT Traditional Chinese (Cross-lingual)	Yes	0.4713	0.1290
NTU-EN-EN-AUTO-QE-TXT-TOPIC English (Monolingual)	No	0.5439	0.1577
NTU-EN-EN-AUTO-QE-WIKI-TXT English (Monolingual)	Yes	0.5102	0.1330

In Table 3 we can find performance doesn't change very much after limiting the web site we access and it even has little decrease in monolingual run. Comparing the pages we access, we find for some topics, limiting web site may let the pages we retrieved not so related with topic. For example, the English textual query for topic 20 is "close-up photograph of an animal", if we don't limit the web site the web resource we access is "Close-up pictures of wildlife, including elephant, lion and leopard …", it's related with the topic, but if we limit the web site, the web resource we access is "Xingyiquan is based on twelve distinct animal forms. Present in all regional and family styles …" it's actually unrelated with our topic. The Third issue we want to check is what will happen when we do document expansion. The results show in Table4.

Table 4. Comparing results that using or not using document expansion

Runs Name / Query Language	Document Expansion	Query Expansion	Recall	MAP
NTU-ZHT-EN-AUTO-NOFB-TXT Traditional Chinese (Cross-lingual)	No	No	0.3776	0.1055
NTU-ZHT-EN-AUTO-DE-TXT-CAPTION Traditional Chinese (Cross-lingual)	Yes	No	0.3050	0.0757
NTU-ZHT-EN-AUTO-QE-TXT-TOPIC Traditional Chinese (Cross-lingual)	No	Yes	0.4461	0.1225
NTU-ZHT-EN-AUTO-DE-QE-TXT Traditional Chinese (Cross-lingual)	Yes	Yes	0.3562	0.0799
NTU-EN-EN-AUTO-NOFB-TXT English (Monolingual)	No	No	0.4736	0.1466
NTU-EN-EN-AUTO-DE-TXT-CAPTION English (Monolingual)	Yes	No	0.3729	0.1154
NTU-EN-EN-AUTO-QE-TXT-TOPIC English (Monolingual)	No	Yes	0.5439	0.1577
NTU-EN-EN-AUTO-DE-QE-TXT English (Monolingual)	Yes	Yes	0.4156	0.1203

In Table 4 we can find document expansion looks doesn't useful. After using document expansion, all runs' performance is decreasing. We think the reason may because document expansion brings too much noise in our corpus.

The fourth issue in our experiment is to exam the performance of media mapping method in the new task definitions. The results show in Table 5.

Table 5. the result that using media mapping to do query expansion

Query Language	MAP/RECALL	Query Expansion (using media mapping)	Without Expansion
Traditional Chinese (Cross-lingual)	MAP	0.2565 (+143.12 %)	0.1055
	RECALL	0.6405 (+69.62 %)	0.3776
Simplified Chinese (Cross-lingual)	MAP	0.2565 (+143.12 %)	0.1055
	RECALL	0.6405 (+69.62 %)	0.3776
Portuguese (Cross-lingual)	MAP	0.2820 (+109.35 %)	0.1347
	RECALL	0.6733 (+51.81 %)	0.4435
Spanish (Cross-lingual)	MAP	0.2785 (+96.12 %)	0.1420
	RECALL	0.6718 (+50.89 %)	0.4452
Russian (Cross-lingual)	MAP	0.2731 (+100.66 %)	0.1361
	RECALL	0.6738 (+46.54 %)	0.4598
Italian (Cross-lingual)	MAP	0.2705 (+130.60 %)	0.1173
	RECALL	0.6481 (+67.59 %)	0.3867
French (Cross-lingual)	MAP	0.2669 (+95.96 %)	0.1362
	RECALL	0.6651 (+61.74 %)	0.4112
Japanese (Cross-lingual)	MAP	0.2551 (+117.29 %)	0.1174
	RECALL	0.6507 (+57.55 %)	0.4130
English (Monolingual)	MAP	0.2737 (+86.69 %)	0.1466
	RECALL	0.6812 (+43.83 %)	0.4736
Spanish (Monolingual)	MAP	0.2792 (+92.02 %)	0.1454
	RECALL	0.6282 (+32.30 %)	0.4748
German (Monolingual)	MAP	0.2449 (+128.87 %)	0.1070
	RECALL	0.5790 (+82.64 %)	0.3170

In Table 5 we can find using media mapping to do query expansion get very well performance. The MAP improves about 86% ~ 143%. Comparing with results in last year, in last year performance improve about 71%~119% after using media mapping method. This result shows media mapping method isn't influence by the

new task definition and still very useful. The reason that the results of Traditional Chinese and Simplified Chinese runs are the same is because in this year the only different in these two queries are their encoding.

Query expansion using media mapping get very well performance, but we want to know if web resource can bring new information that media mapping don't have. We do query expansion using both media mapping and web resource to check if web pages can find some information that media mapping can't find. The result shows in Table6.

Table 6. The result that using both media mapping and web resource to do query expansion

Runs Name / Query Language	media mapping	web resource	Recall	MAP
NTU-ZHT-EN-AUTO-FBQE-TXTIMG Traditional Chinese (Cross-lingual)	Yes	No	0.6405	0.2565
NTU-ZHT-EN-AUTO-QE-FBQE-TXTIMG Traditional Chinese (Cross-lingual)	Yes	Yes	0.6533	0.2255
NTU-EN-EN-AUTO-FBQE-TXTIMG English (Monolingual)	Yes	No	0.6812	0.2737
NTU-EN-EN-AUTO-QE-FBQE-TXTIMG English (Monolingual)	Yes	Yes	0.6738	0.2442

The results in Table 6 show using both web resource and media mapping doesn't improve performance. The reason that MAP decreased after using web resource may because information from web pages has more noises than information from media mapping method. The recall doesn't improve shows we need other method to find other pages that have different information. The pages we access now can't find different information with media mapping method.

5. Conclusion

In this year, we try to use web resource to do query and document expansion. The experiment shows the name entities that expanded from web resource are useful. Limiting web site looks doesn't improve performance and may filter some related webs. The document expansion looks bring too much noise that performance decrease about 28%. Using media mapping to do query expansion improve performance very much, this shows media mapping is still very useful in this year's new task definition. Using both web resource and media mapping to do query expansion don't improve performance. We need other methods to access pages that have different information. We will use this year's experiment to investigate how to select words from web resource and how to filter noise and we will also try different method to access pages.

References

1. Besançon, R., Hède, P., Moellic, P.A., & Fluhr, C. (2005). Cross-media feedback strategies: Merging text and image information to improve image retrieval. In Peters, C.; Clough, P.; Gonzalo, J.; Jones, G.J.F.; Kluck, M.; Magnini, B. (Eds.), *Proceedings of 5th Workshop of the Cross-Language Evaluation Forum*, LNCS 3491, (pp. 709-717). Berlin: Springer.
2. Chang, Y.C. and Chen, H.H. (2007). Approaches of Using a Word-Image Ontology and an Annotated Image Corpus as Intermedia for Cross-Language Image Retrieval. *Working Notes for the CLEF 2006 Workshop*, September 20-22, 2006, Alicante.
3. Clough, P., Sanderson, M. & Müller, H. (2005). The CLEF 2004 cross language image retrieval track. In Peters, C.; Clough, P.; Gonzalo, J.; Jones, G.J.F.; Kluck, M.; Magnini, B. (Eds.), *Proceedings of 5th Workshop of the Cross-Language Evaluation Forum*, LNCS 3491, (pp. 597-613). Berlin: Springer.
4. Clough, P., Müller, H., Deselaers, T., Grubinger, M., Lehmann, T.M., Jensen, J., & Hersh, W. (2006). The CLEF 2005 cross-language image retrieval track, *Proceedings of 6th Workshop of the Cross Language Evaluation Forum*, LNCS 4022, (pp. 535-557). Berlin: Springer.
5. Lin, W.C., Chang, Y.C. and Chen, H.H. (2007). Integrating Textual and Visual Information for Cross-Language Image Retrieval: A Trans-Media Dictionary Approach. *Information Processing and Management*, Special Issue on Asia Information Retrieval Research, 43(2), (pp. 488-502).