Document Expansion for Text-based Image Retrieval at CLEF 2009

Jinming Min, Peter Wilkins, Johannes Leveling, and Gareth Jones

Centre for Next Generation Localisation
School of Computing, Dublin City University
Dublin 9, Ireland
{jmin,pwilkins,jleveling,gjones}@computing.dcu.ie

Abstract. In this paper, we describe and analyze our participation in the WikipediaMM task at CLEF 2009. Our main efforts concern the expansion of the image metadata from the Wikipedia abstracts collection - DBpedia. In our experiments, we use the Okapi feedback algorithm for document expansion. Compared with our text retrieval baseline, our best document expansion RUN improves MAP by 17.89%. As one of our conclusions, document expansion from external resource can play an effective factor in the image metadata retrieval task.

1 Introduction

In this paper, we describe our document expansion (DE) method developed for the WikipediaMM task at CLEF 2009 [1]. This information retrieval task is challenging since the image metadata usually contains less terms which can leads the vocabulary mismatch between the user query and image metadata. We propose a document expansion method to enrich the vocabulary of image metadata documents. With proper expansion of metadata from external resource DBpedia[1], our text retrieval experiment improves MAP by 17.89% compared to the baseline system.

2 Retrieval Model and Document Expansion

After testing different IR models on the text-based image retrieval task, we choose the *tf-idf* model in Lemur toolkit[2] as our baseline model in this task [2]. The document term frequency (tf) weight we use in *tf-idf* model is:

$$tf\,(q_i, D) = \frac{k_1 \cdot f(q_i, D)}{f(q_i, D) + k_1 \cdot (1 - b + b\frac{l_d}{l_c})} \tag{1}$$

$f(q_i, D)$ is the frequency of query term q_i in Document D, l_d is the length of document D, l_c is the average document length of the collection, and k_1

[1] http://dbpedia.org/
[2] http://www.lemurproject.org/

and b are parameters set to 1.2 and 0.75 respectively. The *idf* of a term is given by $log(N/n_t)$, where N is number of documents in the collection and n_t is the number of documents containing term t. The query tf function (qtf) is defined similarly with a parameter representing average query length. The score of document D against query Q is given by:

$$s(D,Q) = \sum_{i=1}^{n} tf(q_i, D) \cdot qtf(q_i, Q) \cdot idf(q_i)^2 \quad (2)$$

qtf is the tf for a term in queries and it is computed using the same method with the tf in documents.

For WikipediaMM 2009 task, we use the following data: the topics, the metadata collection and DBpedia. All these collections are preprocessed to be used in our task. For the topics, we select the title part as the query; for the metadata collection, the text is selected as the query to perform the document expansion and all the tags are removed and only the text in the field "text" will be used. To transform the metadata into the query we process it by:

1. removing useless punctuation in metadata;
2. removing special HTML encoded characters;

The English DBpedia includes 2,787,499 documents corresponding to a brief form of a Wikipedia article. We select 500 stop words by ranking the term frequencies from DBPedia and remove all the stop words before indexing it.

Our document expansion method is similar to a typical query expansion process. In the official runs, we use the pseudo-relevance feedback as our document expansion method with Rocchio's algorithm [3]. The Rocchio algorithm reformulates the query from three parts: the original query, the feedback words from the assumed top relevant documents and the negative feedback terms from the assumed non-relevant documents. For the described experiments, we do not use negative feedback. In our implementation of Rocchio's algorithm, the factors for original query terms and feedback terms are all set to be 1 ($\alpha = 1$, $\beta = 1$). For every metadata document, after preprocessing we use the remaining text as the query. We retrieve the top 100 documents as the assumed relevant documents. With all the words from the returned top 100 documents we first remove all the stop words. We select the top five words as the document expansion words. Then the expanded terms will be added into the metadata document and the index is rebuilt.

In our official runs, we did not index the image name which leads to a loss of related information. We rebuild our index with the image name information and use Equation 3 to select the expansion terms from DBpedia. Here the $r(t_i)$ means the number of documents which contain term t_i in the top 100 assumed relevant documents. *idf* uses the same method as Equation 2.

$$S(t_i) = r(t_i) * idf(t_i) \quad (3)$$

For the number of feedback words, we select the top l_d words ranked using Equation 3, where l_d is the length of the original query document. This strategy is

taken from the method successfully adopted in [4]. So after the official runs, we re-indexed the metadata files and got our new highest result (Run: Document Expansion) in Table 1 and the best results are also from the combination of document expansion and query expansion. A simple explanation to the techniques used in our document expansion research is:

- DEE: document expansion from external resource
- QEE: query expansion from external resource
- QE: query expansion from original metadata documents

For query expansion part in our research, we are using the standard Okapi feedback method for query expansion and we are selecting 10 feedback terms in top 30 assumed relevant documents in the prior retrieval. For our content-based image retrieval run, it can be refereed in [2].

3 Results and Analysis

Table 1. Results of the WikipediaMM 2009.

Run	Modality	Methods	MAP	P@10
dcutfidf-baseline	TXT	BASELINE	0.1576	**0.2600**
dcutfidf-dbpedia-qe	TXT	DEE	0.1685	**0.2600**
dcutfidf-dbpediametadata-dbpediaqe	TXT	QEE+DEE+QE	0.1641	0.2378
dcutfidf-dbpediametadata-qe	TXT	DEE+QE	**0.1752**	0.2578
dcuimg	IMG	BASELINE	0.0079	0.0244
Document Expansion	TXT	DEE+QE	**0.1858**	**0.2844**

In Table 1, our results show that document expansion from external resource can be a very effective approach in text-based image retrieval task. DE can improve 6.92% comparting to baseline run and improve 11.17% when combing with QE. After adding the image name information, the MAP was improved 17.89% comparing to baseline. We can conclude the image name is a very important information to describe the content of the image. Furthermore, we perform the significance test for our results. We are comparing the baseline and our new DE result. In Figure 1 we give the scatter plot for the 45 topics' MAP from these two runs. The paired t-test was used to compute statistical significance ($p = 0.0191$), and this difference is considered to be statistically significant by conventional criteria.

Comparing the Document Expansion Run with our baseline run, we have 26 of 45 topics in Document Expansion run improves. To analyze the effect of DE in more detail, we selected the topic 92 "bikes" as the good example for DE which improves MAP from 0.0159 to 0.2681. And we will compare the top 5 results from baseline run and DE run. The top 5 results for topic 92 in baseline run is 104197, 72035, **29066**[3], 171738, and 201403. In these results, only the

[3] Bold font means it is relevant with the topic

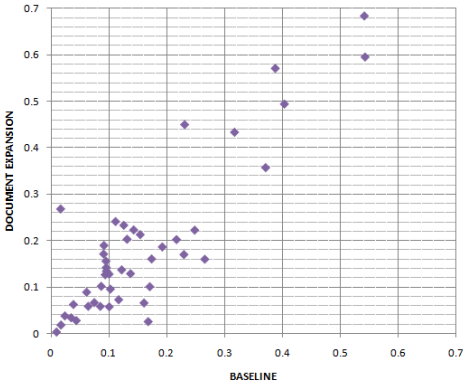

Fig. 1. Average Precision Difference.

third result is judged as relevant ($P@5 = 0.2$). After we have done DE, we get different top 5 results using the same retrieval model which are **126160**, **244171**, 171738, **256625**, and 10283 ($P@5 = 0.6$). We choose the document 126160 as an example to observe what happened after DE.

```
<DOC>
<DOCNO>126160</DOCNO>
<TEXT>
 <ORIGINAL>mountain bike image of a mountain bicycle frog perspective copy
  pierpaolo corona vajont italy</ORIGINAL>
 <EXPANSION>bike bicycle racing cycling cross racer frog bikes stationary
  exercise bicycles race</EXPANSION>
</TEXT>
</DOC>
```

After expansion, many words related to "bikes" are added to the document which make this document more focus on the "bikes"and it is also the main meaning of this image metadata. Document 104792 was ranked first before DE. After DE the document is expanded with words related to "chile" and it is not included in the results for experiment using DE. The document is relevant to the topic.

```
<DOC>
<DOCNO>104197</DOCNO>
<TEXT>
 <ORIGINAL>my bike from santiago of chile</ORIGINAL>
 <EXPANSION>chile santiago metropolitan chilean spanish airline</EXPANSION>
</TEXT>
</DOC>
```

Through our observation, we find that DE strengthens the main meaning of the document. For the image metadata, usually the main meaning can be de-

scribed by a few key words. So we are finding which words are the most important and expand the document using related terms extracted from DBpedia.

4 Conclusion

We presented and analyzed our system for the WikipediaMM task at CLEF 2009 focusing on document expansion. From past research, whether the document expansion can improve the IR effectiveness or how to improve it is not obvious [5]. Our main findings in this research are as follows. DE can improve the retrieval performance for our text-based image retrieval task. The reason is that image metadata can be viewed as short-length documents which usually contain few words to describe the content of the image. When expanding the metadata from the related external resources, it will help to solve the query-document mismatch problem in this task. Since our external resources are also short-length documents, we choose a higher number as the assumed relevant documents in the pseudo relevant feedback process. Finally, we find DE's main impact will take effect in the final QE process. Combining document reduction, DE and QE produces the best results in text-based image retrieval. Furthermore we will continue the research by exploring the use of document expansion in ad-hoc IR tasks.

5 Acknowledgments

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (CNGL) project.

References

1. Tsikrika, T., Kludas, J.: Overview of the WikipediaMM Task at ImageCLEF 2009. In: Working Notes for the CLEF 2009 Workshop, Corfu, Greece (2009)
2. Min, J., Wilkins, P., Leveling, J., Jones, G.: DCU at WikipediaMM 2009: Document Expansion from Wikipedia Abstracts. In: Working Notes for the CLEF 2009 Workshop, Corfu, Greece (2009)
3. Rocchio, J.: Relevance feedback in information retrieval. In: In Gerard Salton, editor, The SMART Retrieval System-Experiments in Automatic Document Processing, Englewood Cliffs, NJ, USA (1971) 313–323
4. Singhal, A., Pereira, F.: Document Expansion for Speech Retrieval. In: Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval, Berkeley, California, USA (1999) 34–41
5. Billerbeck, B., Zobel, J.: Document expansion versus query expansion for ad-hoc retrieval. In: The Tenth Australasian Document Computing Symposium, Sydney, Australia (2005) 34–41