Document Expansion for Text-based Image Retrieval at WikipediaMM 2010

Jinming Min, Johannes Leveling, and Gareth J. F. Jones

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

Abstract. We describe and analyze our participation in the WikipediaMM task at ImageCLEF 2010. Our approach is based on text-based image retrieval using information retrieval techniques on the metadata documents of the images. We submitted two English monolingual runs and one multilingual run. The monolingual runs used the query to retrieve the metadata document with the query and document in the same language; the multilingual run used queries in one language to search the metadata provided in three languages. The main focus of our work was using the English query to retrieve images based on the English metadata. For these experiments the English metadata data was expanded using an external resource - DBpedia. This study expanded on our application of document expansion in our previous participation in ImageCLEF 2009. In 2010 we combined document expansion with a document reduction technique which aimed to include only topically important words to the metadata. Our experiments used the Okapi feedback algorithm for document expansion and Okapi BM25 model for retrieval. Experimental results show that combining document expansion with the document reduction method give the best overall retrieval results.

Keywords: text-based image search, metadata-based search, relevance feedback, document expansion

1 Introduction

This paper describes our participation in the WikipediaMM task at CLEF 2010 [1]. Our approach to this task was based only on text retrieval using the metadata provided for each image. This is a challenging information retrieval (IR) task since the image metadata usually contains less terms than would be found in text documents more typically used in IR. This can lead to problems of vocabulary mismatch between the user query and image metadata. For our particpation in CLEF 2010, we continued to explore the document expansion research for this task which we utilized in WikipediaMM 2009 [2]. This year our document expansion method was combined with a document reduction technique.

This paper is structured as follows: Section 2 introduces the retrieval model used in this work, Section 3 describes our document expansion and document reduction methods, Section 4 records and analyzes our experimental results, and finally Section 5 gives conclusions and directions for further work.

2 Retrieval Model

After testing different IR models on the text-based image retrieval task, we chose the *tf-idf* model in the Lemur toolkit[1] as our baseline model for this task [2]. The document term frequency (tf) weight used in the *tf-idf* model is:

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

where $f(q_i, D)$ 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 and b are fixed parameters set to 1.2 and 0.75 respectively for this task (default values in Lemur toolkit). 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:

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

qtf is the tf for a term in queries, computed using the same method as the tf in the documents.

For the WikipediaMM 2010 task, we use the following data: the topics, the metadata collection and English DBpedia collection (version 3.5). All these collections were preprocessed for use in our work. For the topics, we selected the English title as the query; for the metadata collection, the text including the image name, description, comment and caption was selected as the query to perform the document expansion and all the tags were removed. To transform the metadata into the query was processed as follows:

1. removing punctuation in metadata text;
2. removing URLs from the metadata text;
3. removing special HTML encoded characters;

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

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

3 Document Expansion

Our document expansion method is similar to a typical query expansion process. In the official runs, we used pseudo-relevance feedback (PRF) as our document expansion method with the Okapi feedback algorithm [3]. The Okapi feedback algorithm reformulates the query from two parts: the original query, feedback words from the assumed top relevant documents. In our implementation of the query expansion process, the factors for original query terms and feedback terms are all set to be 1 ($\alpha = 1$, $\beta = 1$) which has been applied successfully in previous document expansion work [4]. 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. We first remove all the stop words from the returned top 100 documents. We select the top five words as the document expansion terms. The selected expansion terms are then added to the metadata document and the index is rebuilt.

In our official runs, we 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 4.

$$S(t_i) = r(t_i) * idf(t_i) \tag{3}$$

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]. Our best results are from the combination of document reduction, document expansion and query expansion. Use of our document expansion techniques is designated as follows:

- DE: document expansion from external resource
- DR: document reduction for the metadata documents
- QE: query expansion from original metadata documents

3.1 Document Reduction

In previous research on DE, usually all the words in the document are associated with the same weight as the "query" terms to find relevant documents prior to expansion. Given an example document "blue flower shot by user", an obvious problem is easily identified. In this document the phrase "blue flower" is an accurate description of the image. If we leave the noise words "shot by user" in the query, it will not help us find good relevant documents. So our method first computes the importance for each term in a document. To do this we compute the weight of each term as its significance using the Okapi BM25 function.

For example, considering the following document from the WikipediaMM collection in Fig 1, the document will be "billcratty2 summary old publicity portrait of dancer choreographer bill cratty. photo by jack mitchell. licensing promotional" after preprocessing. If we manually select the important words from the document, we could form a new document: "old publicity portrait of

dancer choreographer bill cratty". Using the reduced document as the query document is obviously better than the original one in terms of locating potentially useful DE terms. For automatic reduction of the document, we first compute all the term *idf* scores of the collection vocabulary as defined in Equation 4.

$$idf(t_i) = log\frac{N - n(t_i) + 0.5}{n(t_i) + 0.5} \tag{4}$$

here t_i is the ith term, and N is the total number of documents in this collection; $n(t_i)$ is the number of the documents which contain the term t_i. So for every word t_i in document D, we can compute its BM25 weight using Equation 5:

$$weight(t_i, D) = idf(t_i)\frac{f(t_i, D)(k_1 + 1)}{f(t_i, D) + k_1(1 - b + b\frac{|D|}{avgdl})} \tag{5}$$

here $f(t_i, D)$ is the frequency of word t_i in document D; k_1 and b are parameters ($k_1 = 2.0$, $b = 0.75$, starting parameters suggested by [3]); $|D|$ is the length of the document D; and $avgdl$ is the average length of documents in the collection. For the above example, the BM25 score of each term is shown in Table 1 after removing the stopwords.

Table 1. Document *BM25* Score Example

Term	Score
billcratty2	13.316
cratty	12.725
choreographer	12.046
dancer	10.186
mitchell	8.850
bill	7.273
jack	7.174
publicity	6.238
portrait	5.515
promotional	4.389
photo	2.696
summary	2.297
licensing	2.106

We propose to reduce documents by ranking their terms using their BM25 score in decreasing order and removing all terms below a given cut-off value (given as a percentage here). If we choose 50% as the number to reduce the document length, we get the new document "billcratty2 cratty choreographer dancer mitchell bill" for the above example. We call the cut-off value the document reduction rate, which can be defined as: If the reduction rate is $r\%$, we will keep $r\%$ of the original length for the document, and the length of a document means the number of all terms in a document. Using the new reduced document

as the query to obtain documents for expansion produces some differences in the top ranked documents compared to the DE method without DR process. Thus it will select different feedback words from the relevant documents.

```
</article>
<?xml version="1.0"?>
<article>
<name id="23918">BillCratty2.jpg</name>
<text>
 <h2>Summary</h2> Old publicity portrait of dancer
  choreographer Bill Cratty. Photo by Jack Mitchell.
 <h2>Licensing</h2>
 <value>Promotional</value>
</text>
</article>
```

Fig. 1. Document Example.

4 Results and Analysis

For our participation in this task we submitted three official runs as shown in Table 2. Our best result comes from the combination of document reduction, document expansion and query expansion. In our document reduction experiment, the document reduction rate is set 50%. For the run dcuRunOkapi, the English metadata was expanded from the English DBpedia; for the run dcuRunOkapiAll, the index is built from the combination of the expanded English metadata and the original French and German metadata. These two runs produce the same retrieval results since the French and German metadata do not affect the English query very much. For the run dcuRunOkapiFR, French topics were used to search the French metadata. The retrieval effectiveness was found to be relatively low compared to the English runs. The reason for this is due to the generally very significant lack of French metadata in the WikipediaMM image collection. To compare with our DE result, we also provide another English baseline run without document expansion - baselineEnRun. Comparing DE run with baseline run, DE run improves 12.96% by MAP criteria. Applied paired t-test, the two runs are significantly different ($p \leq 0.005$). In Figure 2, TL means topic language and AL means annotation language.

5 Conclusion

This paper presented and analyzed our system for the WikipediaMM task at CLEF 2010 focusing on document reduction and document expansion. In our

Table 2. Results of the WikipediaMM 2009.

Run	Modality	Methods	TL	AL	MAP	P@10
dcuRunOkapi	TXT	DR+DE+QE	EN	EN	0.2039 (+12.96%)	**0.4271**
dcuRunOkapiAll	TXT	DR+DE+QE	EN	EN+FR+DE	0.2039 (+12.96%)	**0.4271**
baselineEnRun	TXT	QE	EN	EN	0.1805	0.4200
dcuRunOkapiFR	TXT	QE	FR	FR	0.1192	0.3243

past research, document expansion from external resources has been shown to be effective in the text based image retrieval task. This year, document expansion combined with document reduction produces effective results in this task.

Our main findings in this research are as follows. Document expansion can improve the retrieval performance for our text-based image retrieval task. For this year, using the improved document expansion method with document reduction still gives us a good retrieval result in this task. From the overall results from this task, the combination of content-based image retrieval and text based image retrieval methods performs better than the single method and this will form one of our future research directions.

6 Acknowledgments

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

References

1. Popescu, A., Tsikrika, T., Kludas, J.: Overview of the Wikipedia Retrieval task at ImageCLEF 2010. In: Working Notes of CLEF 2010, Padova, Italy (2010)
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. Robertson, S., Spärck Jones, K.: Simple, proven approaches to text retrieval. TR UCAM-CL-TR-356, University of Cambridge, Computer Laboratory (1994)
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