DCU at WikipediaMM 2009: Document Expansion from Wikipedia Abstracts

Jinming Min, Peter Wilkins, Johannes Leveling, 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 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. Since the metadata is short for retrieval by query words, we decided to expand the metadata using a typical query expansion method. In our experiments, we use the Rocchio algorithm for document expansion. Our best run is in the 26th rank of all 57 runs which is under our expectation, and we think that the main reason is that our document expansion method uses all the words from the metadata documents which contain words which are unrelated to the content of the images. Compared with our text retrieval baseline, our best document expansion run improves MAP by 11.17%. As one of our conclusions, we think that the document expansion can play an effective factor in the image metadata retrieval task. Our content-based image retrieval uses the same approach as in our participation in ImageCLEF 2008.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval

General Terms

Measurement, Performance, Experimentation

Keywords

Query formulation, Relevance feedback, Document Expansion

1 Introduction

This is DCU's first participation in the WikipediaMM task of CLEF. This task aims to find relevant images based on the query text and image. In the image collection, every image is associated with a metadata file. The file usually contains the description, copyright, author, camera parameters, date and location of the image; for the topics, the data consists of the image and the query text. So this task can be performed by text retrieval or content-based image retrieval or by a combination of these two methods. Our main research efforts are on the text retrieval. Since the useful information in the metadata is usually very short, this text retrieval task is different from the ad-hoc retrieval for news or articles, and we call it short-length documents retrieval.

Since the metadata files contain only very few words to describe the content of the images, we decided to expand the metadata document from an external resource, the Wikipedia abstracts collection DBpedia[1]. Document expansion is quite similar to query expansion in information retrieval research, and we use the Rocchio algorithm as our document expansion method [1]. We have also tested the combination of query expansion from external resource with document expansion from the external resource, but the result is not as effective as the document expansion in this task. With proper expansion from the Wikipedia abstracts corpus, our text retrieval experiment improves MAP by 11.17% percent compared with the baseline system.

2 Text Retrieval System Description

In our approach, we use the Lemur toolkit[2] as the retrieval system. We have tested all different retrieval models in the Lemur toolkit and found that the tf-idf model performs well in this task. Our formal runs employ Lemur's tf-idf method as the retrieval model.

We parse the metadata to be used for indexing and document expansion. For the text part of the topics, we directly extract all words to form a query. The system overview is shown in Figure 1.

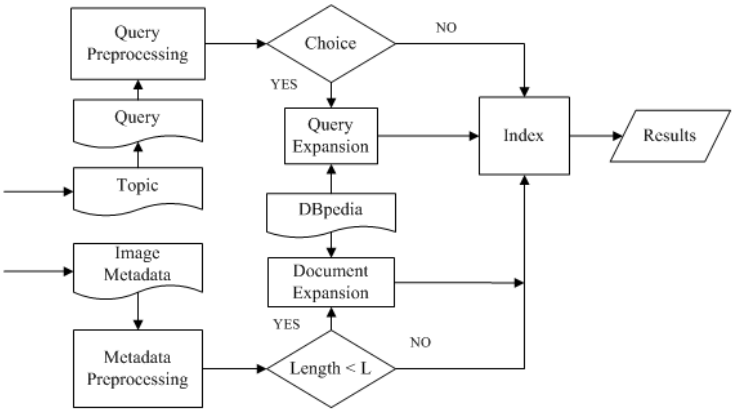

Figure 1: System Overview.

For the queries, we choose to expand queries based on external resources or not which could lead different results; for the metadata document, we will expand it when it's length is less than a threshold $L < 200$. We found long-length documents in the metadata documents contain more noise to get useful relevant feedback terms.

In the following parts, we will describe the preprocessing, the retrieval model used in the task, and the document expansion algorithm.

2.1 Preprocessing

For WikipediaMM 2009, 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. An example is shown in Figure 2:

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

```
<article>
<name id="455">812_image_17.jpg</name>
<image xlink:type="simple" xlink:actuate="onLoad"
xlink:show="embed" xlink:href="../pictures/812_image_17.jpg"
id="455" part="images-110000">812_image_17.jpg</image>
<text>
Paris hugs Butters.
<p/>
<wikitemplate parameters="1">
<wikiparameter number="0" last="1">
<value>tv-screenshot</value>
</wikiparameter>
</wikitemplate>
</text>
</article>
```

Figure 2: Metadata Example.

From the image metadata, all the tags are removed and only the text in the field "text" will be used, thus the document for expansion will be "Paris hugs Butters. tv-screenshot". To transform the metadata into the query we process it by:

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

For every metadata document, after parsing we set the remaining text as the query.

For the DBpedia, we use the text as index and remove the stop words computed from itself. An example document from DBpedia is:

```
<DOC>
<DOCNO>1969 Paris Open</DOCNO>
<TEXT>The 1969 Paris Open was a professional tennis tournament played on
indoor carpet courts. It was the 2nd edition of the Paris Open (later
known as the Paris Masters).
</TEXT>
</DOC>
```

Figure 3: DBpedia Example.

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

2.2 Text Retrieval Model

After testing different information retrieval models on the text based image retrieval task, we found that the *tf-idf* model outperforms state-of-art models such as Okapi *BM25* or language modeling in Lemur toolkit. So we choose the *tf-idf* model as our baseline retrieval model in this task. 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 and b are parameters set to 1.2 and 0.75 respectively. The *idf* of a term is given by $log(N/n_t)$. 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:

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

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

2.3 Document Expansion

Our document expansion method is similar to a typical query expansion process. We use the pseudo-relevance feedback as our document expansion method with Rocchio's algorithm [1]. 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 ($\alpha = 1$, $\beta = 1$). We choose the DBpedia as the external resource for document expansion. The reasons are:

1. the DBpedia dataset contains only the Wikipedia terms definition sentences which contains less noise than full articles;

2. the DBpedia documents are also derived from Wikipedia documents which share some characteristics with our image metadata documents form Wikipedia.

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. The stop words list is produced from the DBpedia document collection, and we compute the term frequency from the DBpedia collection and set the top 500 words as the stop words. For the top 100 relevant documents in DBpedia, we compute a word frequency list and remove the stop words and the original words from the query. We select the top five words as the document expansion words.

The number of relevant documents for document expansion is higher than normal because the Wikipedia abstract corpus usually has very short documents. If we only used 10 or 20 as the assumed relevant documents, it would be very hard for us to get useful feedback terms from the relevant documents. Furthermore, the original metadata documents are short so we only select the top 5 terms as the feedback terms. Then the expanded terms will be added into the metadata document and the index is rebuilt.

3 Content-Based Image Retrieval

For content-based image retrieval we make use of the following six global visual features defined in the MPEG-7 specification:

- **Scalable Colour (SC):** derived from a colour histogram defined in the HSV colour space. It uses a Haar transform coefficient encoding, allowing scalable representation.

- **Colour Structure (CS):** based on colour histograms, the feature represents an image by both the color distribution (similar to a color histogram) and the local spatial structure of the colour.

- **Colour Layout (CL):** compact descriptor which captures the spatial layout of the representative colours on a grid superimposed on an image.
- **Colour Moments (CM):** similar to Colour Layout, this descriptor divides an image into 4x4 subimages and for each subimage the mean and the variance on each LUV color space component is computed.
- **Edge Histogram (EH):** represents the spatial distribution of edges in an image. Edges are categorized into five types: vertical, horizontal, 45 degrees diagonal, 135 degrees diagonal and non directional.
- **Homogeneous Texture (HT):** provides a quantitative representation using 62 numbers, consisting of the mean energy and the energy deviation from a set of frequency channels.

Our work for visual querying was the same approach as undertaken in ImageCLEF 2008. For a visual query, we take the topic images and extract from each their representation of the image by each of the six features above. For each feature we query its associated retrieval expert (i.e. visual index and ranking function) to produce a ranked list. The ranking metric for each feature is as specified by MPEG-7 and is typically a variation on Euclidian distance. For our experiments we kept the top 1000 results. Each ranked list was then weighted and the results from all ranked lists are normalized using MinMax [2], then linearly combined using CombSUM [2].

The weighting we employed was linear, using an approach where the weights are determined at *query-time*. This approach is the same as used in our previous ImageCLEF experiments, with an explanation found in [3].

4 Results

We have submitted 5 runs to the WikipediaMM task including 4 runs by the text retrieval system and 1 run by image retrieval system. The main technique used in the text retrieval is: query expansion (QE), query expansion from external resource (QEE), document expansion from external resource (DEE). Our 4 runs are combinations of these techniques and the baseline run uses only the tf-idf IR model without additions. The results for English monolingual WikipediaMM 2009 task are shown in Table 1.

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

Table 1: Official Results for the WikipediaMM 2009.

Our best result ranks in the middle of all the official runs in the WikipediaMM 2009 task. Compared with our baseline result, document expansion plays an important role in our best result. Document expansion can improve the MAP from 0.1576 to 0.1685, but query expansion from external resource in combination with our methods does not show much improvement. Our image run gets bad results due to some computation error which will fixed in the future research.

5 Conclusion

We presented our system for the WikipediaMM task of CLEF 2009 focusing on document expansion. Document expansion has not been thoroughly researched for information retrieval. From the past research, whether the document expansion can improve the retrieval effectiveness or how to

improve it is not obvious [4, 5]. Our results show that the document expansion could play a role in the image metadata retrieval task. Also the documents usually contain too much information unrelated to the content of picture such as the copyright and author information. This information used in the document expansion will greatly harm the expansion results. In future experiments we will try to remove some noise from the documents and use the words related with the content of the image as the query to perform document expansion.

From this task, our main finding is that the document expansion can improve the retrieval effectiveness much when the document length is short. On the other hand, query expansion from the external resource does not improve performance since the query text usually is very accurate and does not need to be expanded with more words.

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) project.

References

[1] J.J. Rocchio. Relevance feedback in information retrieval. In *In Gerard Salton, editor, The SMART Retrieval System-Experiments in Automatic Document Processing*, pages 313–323, Englewood Cliffs, NJ, USA, 1971.

[2] Edward A. Fox and Joseph A. Shaw. Combination of Multiple Searches. In *Proceedings of the Third Text REtreival Conference (TREC-1994)*, pages 243–252, Gaithersburg, MD, USA, 1994.

[3] Anni Jarvelin, Peter Wilkins, Tomasz Adamek, Eija Airio, Gareth Jones, Alan F. Smeaton, and Eero Sormunen. Dcu and uta at imageclefphoto 2007. In *ImageCLEF 2007 - The CLEF Cross Language Image Retrieval Track Workshop*, Budapest, Hungary, 2007.

[4] Amit Singhal and Fernando Pereira. Document Expansion for Speech Retrieval. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 34–41, Berkeley, California, USA, 1999.

[5] Bodo Billerbeck and Justin Zobel. Document expansion versus query expansion for ad-hoc retrieval. In *The Tenth Australasian Document Computing Symposium*, pages 34–41, Sydney, Australia, December 2005.