# Document Expansion for Image Retrieval

Jinming Min[1] Johannes Leveling[1] Dong Zhou[2] Gareth J. F. Jones[1]

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

[2] Centre for Next Generation Localisation
University of Dublin
Trinity College
Dublin 2, Ireland
dong.zhou@cs.tcd.ie

## ABSTRACT

Successful information retrieval requires effective matching between the user's search request and the contents of relevant documents. Often the request entered by a user may not use the same topic relevant terms as the authors' of these documents. One potential approach to address problems of query-document term mismatch is document expansion to include additional topically relevant indexing terms in a document which may encourage its retrieval when relevant to queries which do not match its original contents well. We propose and evaluate a new document expansion method using external resources. While results of previous research have been inconclusive in determining the impact of document expansion on retrieval effectiveness, our method is shown to work effectively for text-based image retrieval of short image annotation documents. Our approach uses the Okapi query expansion algorithm as a method for document expansion. We further show improved performance can be achieved by using a "document reduction" approach to include only the significant terms in a document in the expansion process. Our experiments on the WikipediaMM task at ImageCLEF 2008 show an increase of 16.5% in mean average precision (MAP) compared to a variation of Okapi BM25 retrieval model. To compare document expansion with query expansion, we also test query expansion from an external resource which leads an improvement by 9.84% in MAP over our baseline. Our conclusion is that the document expansion with document reduction and in combination with query expansion produces the overall best retrieval results for short-length document retrieval. For this image retrieval task, we also conclude that query expansion from external resources does not outperform the document expansion method.

## Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search and Retrieval -search process, query formulation



## General Terms

Algorithms, Design, Experimentation, Performance

## Keywords

Information Retrieval, Document Expansion, Query Expansion, Pseudo-relevance feedback, Wikipedia

## 1. INTRODUCTION

One of the key issues for successful information retrieval (IR) is the matching of terms in a user search request against the contents of relevant documents. Specifically searchers may not use the same topic relevant terms as the authors of the documents. An obvious way to address this problem is to modify either or both of the query and document to encourage suitable matching in an attempt to retrieve relevant content at improved ranks. Query expansion (QE) has been the subject of many studies in relevance feedback for improving IR. QE can be completely automatic or by a process of recommendation involving the user as offered by search engines such as *Google* and *Bing*. Often one of the problems of QE is the need to perform significant amounts of computation when the query is entered. Document expansion (DE) seeks to address the mismatch problem from the opposite perspective of seeking to include additional topically relevant indexing terms in the document representation in the search index. Compared with QE, DE has the potential advantage that it does not introduce additional computation at query time. While it has potential benefits for IR, DE has received comparatively little research attention compared to QE, and existing results have been mixed with some of the best known work suggesting that it is not effective for IR [1].

In this paper we revisit DE in the context of retrieval of images annotated with brief textual labels. This task is challenging for IR since such annotations are generally short, often with no redundancy of description, and typically do not follow any particular standard in terms of vocabulary selection or level of detail, leading to a high likelihood of mismatch with user queries. Thus, if we can build an improved connection between image annotations and user queries, there is potential to greatly benefit retrieval effectiveness. In this context DE becomes at attractive option, if it can be shown to work reliably. In this paper we use a novel DE technique to demonstrate improvement in IR for image search. Since QE has been proven to be an effective way to solve the term mismatch problem in IR research, it is

reasonable to compare our DE method against state-of-art QE. To be fair, we test QE using the same external resource that we use for DE. The same Okapi feedback algorithm is used for both DE and QE.

This paper is structured as follows: Section 2 overviews the background and related work to our investigation, Section 3 describes our DE method, Section 4 describes our method for QE from external resource, Section 5 introduces our experimental setup and results, Section 6 analyzes our results, and finally Section 7 gives conclusions and directions for further work.

## 2. BACKGROUND AND RELATED WORK

Image retrieval can be performed in several different ways. Approach one is to search the metadata associated with the image source and treat it as a text retrieval task [3], another approach is to analyze the image contents and treat it as a content-based image retrieval task [11]. Alternatively a combination of these two methods can be used to achieve better retrieval results. In this research, our focus is on improving text-based image retrieval.

As described in the introduction, textual image annotations which are often very short introduce significant problems of query mismatch. Document expansion (DE) is a technique for enriching source documents by adding topically related terms, and as such is a potentially useful method to apply for text-based image retrieval. DE was initially introduced in the field of speech retrieval where automatic transcriptions are noisy and lead to mismatch problems [10]. During the expansion process, documents were used as queries to search an external collection of documents. A number of terms from this external collection were then selected from the top-ranked documents returned by this search, and then added to the original (query) document. Singhal et al. [10] introduced DE as a method to recover those words that might have been in the original speech data, but had been misrecognized during speech retrieval process. They found that enriching the documents via a process of DE yielded retrieval effectiveness that improved not only over the original erroneous transcription, but also over a perfect manual transcription, since not only misrecognized words were added but also topically related words that had not been spoken. In the area of cross-lingual information retrieval, Levow published a series of experiments [7, 6, 5] exploring pre- and post-translation DE for both spoken and text documents in Mandarin-English cross-lingual retrieval and showed some improvements. In language modeling IR, Tao et al. [13] constructed a method to expand every document with a probabilistic neighborhood. The cosine similarity was used to compute the neighborhood relations of documents. In this work, Tao found that DE helps more for short-length documents. By contrast, an attempt to employ DE in image retrieval during the CLEF campaign degraded the performance by 28.24% in MAP when using the web as the reference corpus [2]. Documents were expanded from the top ranked snippets from a web search engine, but in this case only the document title was used as the query to search for relevant documents. A study reported by Billerbeck and Zobel [1] showed DE to only have limited effects and concluded that the technique is unpromising. Min et al. [8] use the whole document as the query to find relevant documents in DBpedia[1] and expand the top 5 feedback terms into the original document from the top 100 relevant documents. Their results show improvement of 11.17% for MAP. Leveling et al. [4] investigate document expansion using an entry vocabulary module for an ad-hoc retrieval task, but DE combining with Okapi BM25 retrieval model did not show significant improvement.

In summary, previous investigations of DE have met with mixed results. However, compared to QE it has been relatively neglected as an area of research, and the positive findings of some investigations, particularly for short or noisy documents, indicate that is has promise to improve retrieval performance for at least some IR tasks with suitable parameters.

## 3. DOCUMENT EXPANSION METHOD

In this section we propose a DE method for text-based image retrieval. Our DE method is similar to a typical QE process. We use pseudo-relevance feedback (PRF) as our DE method with the Okapi feedback algorithm [9]. The Okapi feedback algorithm reformulates the query from two parts: the original query and the feedback words from the assumed top relevant documents. In our implementation of Okapi feedback, the weight of original query terms and feedback terms are all set to be 1.

Figure 1 presents a system overview. DE terms are extracted from the top ranked relevant documents retrieved from the external resource. The expanded documents are indexed in the image retrieval system. Compared to the previous DE method, the key stage here is to select key terms from the documents prior to expansion in a process we refer to as *document reduction*. The objective of document reduction is to focus the DE "query" on the most significant elements of the document. The remaining terms are used to formulate a query for DE on an external resource. In the following section we introduce our document reduction and DE methods.

### 3.1 Document Reduction

In previous research on DE, usually all the words in the document are associated with the same weight as "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 every 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 Figure 2, 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

[1]http://dbpedia.org

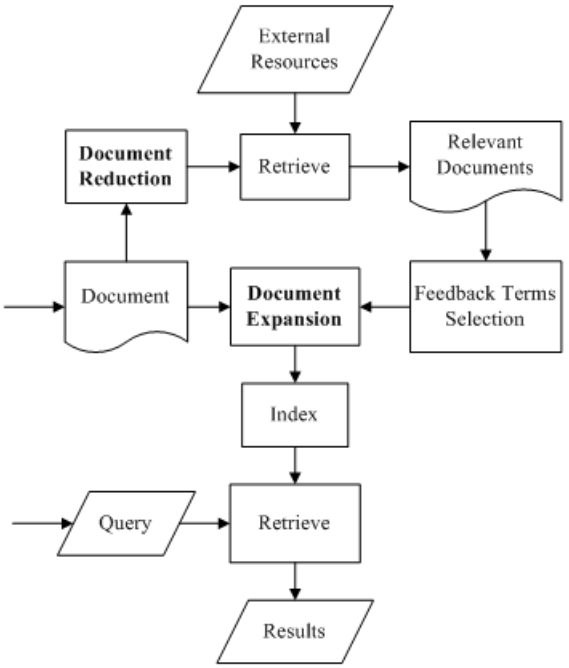


**Figure 1: System Overview.**

collection vocabulary as defined in Equation 1.

$$idf(t_i) = log\frac{N - n(t_i) + 0.5}{n(t_i) + 0.5} \quad (1)$$

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 2:

$$weight(t_i, D) = idf(t_i)\frac{f(t_i, D)(k_1 + 1)}{f(t_i, D) + k_1(1 - b + b\frac{|D|}{avgdl})} \quad (2)$$

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 [9]); $|D|$ 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.

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.

**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 |

```
</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>
```

**Figure 2: Document Example.**

## 3.2 Document Expansion

The documents for expansion are retrieved from an external resource (DBpedia in our experiments) with the top 100 ranked documents as the assumed relevant documents. From all the words in the top 100 documents we first remove all the stop words. The stop word list was produced from the DBpedia document collection, for which we computed the term frequency in the DBpedia collection and select the top 500 words as the stop words. For the top 100 relevant documents, we compute a word frequency list and remove the stop words and ignore the original words contained in the "query". Equation 3 is used to rank the terms. 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 1.

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

The number of assumed relevant documents for DE is higher than would normally be considered for QE because the documents in DBpedia are usually very short length. If we only used 10 or 20 as the assumed relevant documents, it was found to be difficult to get useful feedback terms from the relevant documents. For the number of feedback words, we select the top L words ranked using Equation 3, where L is the length of the original query document. This strategy is taken from the method successfully adopted in [10] and we don't get higher result when trying fixed number as standard query expansion method.

## 3.3 Retrieval Model

After testing different IR models on the text-based image retrieval task, we selected the *tf-idf* model in the Lemur toolkit [2] as our baseline retrieval model. Details of the *tf-idf* model can be found in [16]; this is essentially a variation of the Okapi BM25 model. The document term frequency ($tf$) weight used in *tf-idf* model is shown in Equation 4.

$$tf\,(t_i, D) = \frac{k_1 \cdot f(t_i, D)}{f(t_i, D) + k_1 \cdot (1 - b + b\frac{l_d}{l_c})} \quad (4)$$

$f(t_i, D)$ is the frequency of query term $t_i$ 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 1.0 and 0.3 respectively since our target documents are of short-length [9]. The *idf* of a term is given by $log(N/n(t_i))$, where $N$ and $n(t_i)$ have the same definitions as before.

The query $tf$ function ($qtf$) is also defined using Equation 4 where $k_1$ and $b$ are set to 1000 and 0, so $qtf$ will usually be approximately equal to 1. The score of document $D$ against query $Q$ is calculated as shown in Equation 5.

$$s(D, Q) = \sum_{n}^{i=1} tf(t_i, D) \cdot qtf(t_i, Q) \cdot idf(t_i)^2 \quad (5)$$

In the retrieval process, we also test the effectiveness of query expansion (QE). Using PRF for QE, we set the number of feedback documents to 5, and the number of feedback terms as 20. These feedback terms are added to the query with a factor 1. All these parameters are adjusted manually to get the best result.

# 4. QUERY EXPANSION FROM EXTERNAL RESOURCE

QE is a proven way to address the vocabulary mismatch problem in IR. In this work, we also explore QE from external resources [15] to compare with our DE method. In [15], the authors report that QE from snippets of web search engine results can get better results for TREC collections. We found that for our image retrieval task almost all the queries are noun phrases and usually top-ranked documents returned from a search engine include the Wikipedia link. For this reason, we chose DBpedia as our external resource for QE experiment. Our QE method uses the standard Okapi feedback methods [9]. We set the top $R$ documents as the assumed relevant documents, and the number of feedback terms is $k$ ($R = 30$, $k = 10$ in our experiment). For the expansion process, we also adjusted the factor for the original query terms and feedback terms. In our implementation, we adjust the factor for the original query terms to 2 and the feedback terms to 1 where we get the best result.

Since many queries to DBpedia can directly return the definition of the query, we call the document containing the definition of the query the "definition document". We emphasize the terms from the definition document since it is directly related to the original query. We introduce a method to identify whether a document is the definition document for a query. Given a query $Q = q_1, q_2, ..., q_n$ and a document with title $T = t_1, t_2, ..., t_m$, if $Q$ and $D$ satisfy the following conditions $D$ is classified as the definition document of $Q$:

1. $m \geq n (m, n \geq 0)$;
2. for every $q_i$, we can find a term $t_j$ in $T$ which satisfies $t_j = q_i (1 \leq i \leq n, 1 \leq j \leq m)$.

We search for the definition document in the top R returned documents for a query. If we find it, we build a definition vocabulary set $S = s_1, s_2, ..., s_m$. In the expansion process, if we find a feedback term f stratifying $f \in S$, we give higher weight to it ($w = 2$, in our experiment). If the definition document is not found, we apply the standard Okapi feedback process.

# 5. EXPERIMENTS AND RESULTS

In this section, we describe our experimental setup and results. Experiments were conducted using the collection from the ImageCLEF WikipediaMM task. The corpus is taken from the (INEX MM) Wikipedia image collection and includes 151,520 images [14]. Every image is associated with a metadata file. Another important resource we use is DBpedia which is used as the external resource for DE and QE. DBpedia is a Wikipedia abstract collection and includes 2,452,726 documents. We chose the English DBpedia as the external resource for document expansion since: 1) the DBpedia dataset contains only the definition sentences of Wikipedia terms and so contains less noise than full articles; 2) the DBpedia corpus covers all kinds of topics which promises that we can find relevant documents in it.

In our experiments, there are two kinds of query expansion modules that should be clarified: standard query expansion from the target corpus (QE), and query expansion from the external resource (QEE).

**Table 2: Impact of Document Expansion to Query Expansion (DR Rate as 50%).**

| Runs | QE | MAP | P@10 | R-Prec |
|---|---|---|---|---|
| Baseline | - | 0.2612 | 0.3680 | 0.3094 |
| Baseline + QE | +4.44% | 0.2728 | 0.3680 | 0.3095 |
| DE | - | 0.2620 | 0.3533 | 0.3106 |
| DE + QE | +7.33% | 0.2812 | 0.3520 | 0.3208 |
| DR + DE | - | 0.2866 | 0.3707 | 0.3176 |
| DR + DE + QE | +5.44% | **0.3022** | **0.3907** | **0.3342** |

## 5.1 Comparing with our Baseline

We carried out two baseline experiments. One using the *tf-idf* model with QE and another without QE. For our main investigation, one set of experiments used DE only and another document reduction (DR) combined with DE. These configurations were then further combined with QE. Thus overall we have 6 runs as shown in Table 2. From the results we can see that DE combined with DR and QE yields the best result in terms of MAP, P@10 and R-Precision scores. Here we are using a DR rate of 50% as an empirical starting point. Also in Table 2, we show percentage impact of including QE for each of the DE strategies.

## 5.2 Query Expansion from External Resource

In this paper, we test QE from an external resource (QEE) as a comparison with our DE method. Table 3 compares the QEE method and DE method for our results. As Table 3 shows QEE can improve our results compared with the baseline system, but this result does not outperform our DE

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

result. Figure 3 gives information about QEE and Baseline runs.

We compare the run QEE+QE and the Baseline in Figure 3. This passes a paired t-test (p value is 0.0133) for significance. We chose the paired t-test [12] as our method to do the significance test.

**Table 3: Query Expansion from External Resources (DR Rate as 50%).**

| Runs | MAP | P@10 | R-Prec |
|---|---|---|---|
| Baseline | 0.2612 | 0.3680 | 0.3094 |
| Baseline + QE | 0.2728 | 0.3680 | 0.3095 |
| QEE | 0.2695 | 0.3760 | 0.3092 |
| QEE + QE | 0.2869 | **0.3773** | 0.3222 |
| QEE + DR + DE | 0.2864 | 0.3347 | 0.3253 |
| QEE + DR + DE + QE | **0.2893** | 0.3413 | **0.3319** |

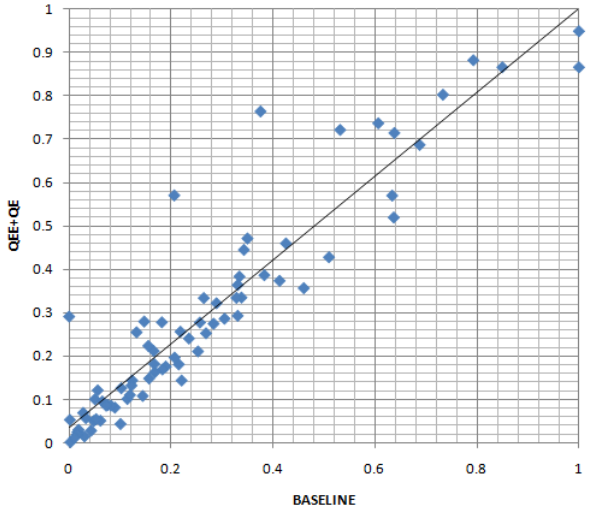


**Figure 3: Average Precision Difference for QEE.**

## 5.3 Document Reduction Rate

We also investigated the choice of the document reduction rate. Results for a range of DR rates are shown in Table 4, and all these results are combined with DE method. In this table, all the runs use the same DE setting as used in Table 2 and include QE. The results show that a DR rate of 70% gives the best retrieval performance in terms of MAP.

Using DE and QE in combination gave an improvement in MAP of 7.66% compared to our baseline. Furthermore, we found that using the whole document as a query was less effective at locating good terms for DE, than using an approach incorporating a document reduction stage. By incorporating document reduction, we get a 16.54% improvement in MAP when combining document reduction with A rate OF 70% with DE and QE.

Also we evaluated QE effectiveness when combining with DR and DE as shown in Table 2. In our baseline experiments, QE produced a 4.44% increase in MAP. When QE is combined with DE, the increase of MAP becomes 7.33%. When QE is combined with DR and DE, we get a 5.44% increase in MAP compared with using DR and DE.

| DR Rate | MAP | P@10 | R-Prec |
|---|---|---|---|
| 10% | 0.2438 | 0.3387 | 0.2704 |
| 20% | 0.2736 | 0.3587 | 0.3149 |
| 30% | 0.2817 | 0.3773 | 0.3190 |
| 40% | 0.3023 | 0.3893 | 0.3326 |
| 50% | 0.3022 | **0.3907** | 0.3342 |
| 60% | 0.3032 | 0.3867 | 0.3312 |
| 70% | **0.3044** | 0.3827 | 0.3368 |
| 80% | 0.2997 | 0.3720 | 0.3386 |
| 90% | 0.2975 | 0.3693 | **0.3393** |
| 100% | 0.2812 | 0.3520 | 0.3208 |

**Table 4: Document Reduction Rate.**

Performing significance tests for our results, there are 75 topics for the WikipediaMM 2008 task. For every topic we give the average precision difference in the Figure 4. We compare the results from the baseline experiment without QE (Baseline) with the combination of document reduction, DE and QE (DR + DE + QE). For t-test the two-tailed P value is 0.0003. So by conventional criteria, this difference is considered to be extremely statistically significant. The increase in MAP of the results from DE+QE to DR+DE+QE is also significant (p=0.0326).

## 5.4 Efficiency Issue

Since DE will make the index size much bigger than the original one, we test our index time and query time. For image retrieval, our metadata are relatively small compared to documents such as news data, so even expanded image metadata documents are relatively small search items. We use Table 5 to describe our index size.

Also we test the query time with the several different runs. We do not find significant change in the query time.

**Table 6: Average Query Time.**

| Runs | Query Time (s) |
|---|---|
| Baseline | 1.714 |
| Baseline + QE | 2.596 |
| DE | 1.852 |
| DE + QE | 2.734 |

# 6. ANALYSIS

Why does DE improve the text-based image retrieval effectiveness? From our observations, the image metadata text has very similar characteristics to a typical query text. It consists of few words and focuses on a single topic. In standard ad-hoc retrieval tasks (such as those at TREC and elsewhere) for text retrieval, documents are typically news articles which are longer and may cover more than one topic. Expanding a long-length document covering more than one topic using a QE algorithm could be an improper choice since it is hard to find documents which are relevant to the documents. In our experiments the metadata document is usually of very short length, which is an intrinsic advantage for the metadata document to make use of the DE algorithm. Using the metadata document as the query, it has a better chance of locating relevant documents within the related external resources. Selecting the top feedback terms and adding them into the metadata document enriches the

Table 5: Index Statistics.

| Runs | Index Time (s) | Index Size (Mb) | Vocabulary | Average Document Length |
|---|---|---|---|---|
| Baseline | 17.005 | 51.6m | 193417 | 24 |
| Document Expansion | 20.780 | **69.5m** | **203613** | **35** |

metadata document vocabulary, but does not weaken its meaning. Thus the expanded metadata document will have more opportunities to be searched effectively by users with an improved chance of query document match. Overall the effects are similar to that of QE. Another aspect in the experiments is the related external resource. The retrieval task is conducted on Wikipedia data so we selected the Wikipedia abstract collection as the document expansion resource. The related external resource is thus an appropriate resource for the DE process.

We believe that the most important difference between DE and QE is that the former can be improved by the process of document reduction since using the whole document as the query to find relevant documents is not the best way for DE. Document reduction can help to remove the noise from the query document and get better relevant documents rank list. Another difference is that DE usually selects expanded terms up to the length of the original documents. In QE, usually the number of the feedback terms is set empirically as a fixed number. In our method, we expanded the documents by doubling their length which has been successfully applied in speech retrieval [10].

And we also get improvement for retrieval effectiveness from QEE. Comparing DE with QEE, our DE method still outperforms the QEE method. Since in this task, the lack of information for image metadata is still the main issue. QEE can only expand using the limited knowledge from external resource into the original queries; but for DE, we can integrate lots of useful information from Wikipedia into metadata documents. In previous research, researchers claim that DE will often make the original documents drift to another topic. But in our task, we think the expansion can still be beneficial for most documents since our expansion method already extracts key terms from the original documents which helps to ensure we expand the key meaning of the original document.

From Table 2 we can see that without DE, QE improves the MAP from 0.26 to 0.27, but with DE, the MAP is improved from 0.26 to 0.28. We think the reason for this is that DE introduces more related words into the documents, so that the QE process can also benefit from it. So in the process of the QE, the feedback words will be more useful for obtaining relevant results in the second retrieval process.

## 6.1 Per-topic Analysis

We have 75 topics for this collection. Comparing the Baseline and DE method, for 47 topics the MAP improves and for 27 topics it decreases while for 1 topic the MAP is unchanged. We select an example document to observe the details of the document expansion process.

For topic 23, the query terms are "british trains". Before DE, the document IDs fro the top 10 results are: **1980516**[3], 222020, 316360, **228342**, **1032854**, **1475020**, **1192327**, **1487499**, **1125229**, **2227472**. Before DE, the P@10 is 0.8. And after DE, we got the P@10 as 1.0. All the top

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

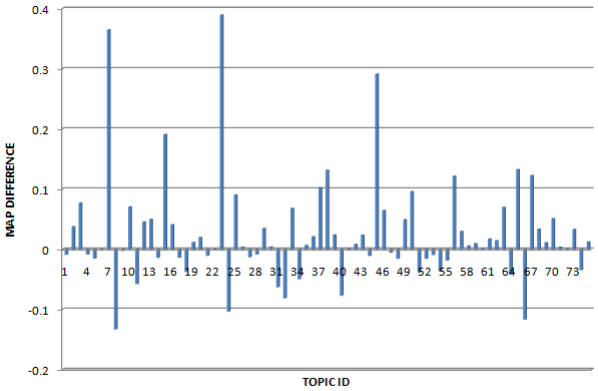


**Figure 4: Average Precision Difference for DE.**

ten documents are relevant document: **1487499**, **1125229**, **1423946**, **1032854**, **1475020**, **1192327**, **1185704**, **1109791**, **2329048**, **1239902**. We select document *1423946* as an example shown in Figure 5 to observe the effectiveness of DE, since its rank for topic 23 improves from 116 in the Baseline run and 48 in the Baseline+QE run to 3 in DR+DE+QE run. In this example, we can find the term "train", it does not appear in the original document but after expansion it does appear.

```
<DOC>
<DOCNO>1423946</DOCNO>
<TEXT>
<ORIGINAL>norwich british rail
class 960 class on 31st january 2004 at the
time this unit was painted in railtrack blue
green livery it has since been reclassified
as british rail and repainted in network rail
yellow livery image by phil scott</ORIGINAL>
<EXPANSION>rail units multiple unit diesel blue
electric locomotives green train livery services
type locomotive introduced freight car passenger
vehicles theotokos steam</EXPANSION>
</TEXT>
</DOC>
```

**Figure 5: Document Expansion Example.**

# 7. CONCLUSIONS AND FUTURE WORK

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

helps 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. We find that using the whole document as the query to do DE can introduce too much noise, and we reduce the document by selecting important words, then use the reduced document as the query to get the relevant documents. This process can help to achieve higher retrieval performance. 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. For QEE, we get significant improvement based on our baseline system. Comparing QEE with the DE method, QEE can only expand limited knowledge into the retrieval process which means that QEE cannot outperform the DE method in the image retrieval task.

Text-based image retrieval is a special case of IR for which our DE method improves the retrieval performance. For this task, one key characteristic is that image metadata can usually be viewed as a short-length document. Using related external resources and extracting words from relevant documents can help solve the query document mismatch in this case. Our future research will focus on whether we can use the same technology for IR on longer documents. In previous DE research, usually the whole document is used as the query to find the relevant documents. Our document reduction method may also be promising for the long-length text retrieval task. Furthermore, we plan to investigate different algorithms to compute the term importance score in the document. This leads to a new research question: is the Okapi BM25 weighting scheme the best method for term selection in document reduction? Also another way to do this would be using a text summarization method. We will continue the research by exploring the use of document expansion in ad-hoc IR tasks.

## 8. 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.

## 9. REFERENCES

[1] B. Billerbeck and J. Zobel. Document Expansion versus Query Expansion for Ad-hoc Retrieval. In *the Tenth Australasian Document Computing Symposium*, pages 34–41, Sydney, Australia, 2005.

[2] Y.-C. Chang and H.-H. Chen. Using an image-text parallel corpus and the web for query expansion in cross-language image retrieval. pages 504–511, 2008.

[3] A. Goodrum. Image information retrieval: An overview of current research. *Informing Science*, 3:2000, 2000.

[4] J. Leveling, D. Zhou, G. J. F. Jones, and V. Wade. TCD-DCU at TEL@CLEF 2009: Document expansion, query translation and laguage modeling. In *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece, 2009.

[5] G.-A. Levow. Issues in Pre- and Post-translation Document Expansion: Untranslatable Cognates and Missegmented Words. In *Proceedings of 4th International Workshop on Information Retrieval in Asian Languages*, pages 77–83, Sapporo, Japan, 2003.

[6] G.-A. Levow. Multi-scale Document Expansion for Mandarin Chinese. In *Proceedings of Workshop on Multilingual Spoken Document Retrieval, at International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 73–78, Hong Kong, 2003.

[7] G.-A. Levow and D. W. Oard. Translingual topic tracking with PRISE. In *Working Notes of the Third Topic Detection and Tracking Workshop*, Tysons Corner, VA, USA, 2000.

[8] J. Min, P. Wilkins, J. Leveling, and G. J. F. Jones. DCU at WikipediaMM 2009: Document expansion from wikipedia abstracts. In *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece, 2009.

[9] S. E. Robertson and K. Spärck Jones. Simple, proven approaches to text retrieval. Technical Report UCAM-CL-TR-356, University of Cambridge, Computer Laboratory, Dec. 1994.

[10] A. Singhal and F. 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.

[11] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:1349–1380, 2000.

[12] M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 623–632, New York, NY, USA, 2007. ACM.

[13] T. Tao, X. Wang, Q. Mei, and C. Zhai. Language model information retrieval with document expansion. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 407–414, New York, USA, 2006.

[14] T. Westerveld and R. van Zwol. The inex 2006 multimedia track. In *N. Fuhr, M. Lalmas, and A. Trotman, editors, Advances in XML Information Retrieval:Fifth International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2006, Lecture Notes in Computer Science/Lecture Notes in Artificial Intelligence (LNCS/LNAI)*, Springer-Verlag, 2007. ACM.

[15] Z. Yin, M. Shokouhi, and N. Craswell. Query expansion using external evidence. In *ECIR '09: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, pages 362–374, Berlin, Heidelberg, 2009. Springer-Verlag.

[16] C. Zhai. Notes on the Lemur TFIDF model, Oct. 2001.