# Relevance Feedback for Cross Language Image Retrieval

Paul Clough and Mark Sanderson

Department of Information Studies, University of Sheffield, Sheffield, UK.
{p.d.clough,m.sanderson}@sheffield.ac.uk

**Abstract.** In this paper we show how relevance feedback can be used to improve retrieval performance for a cross language image retrieval task, an area of CLIR which is different from existing problems, but has thus far received little attention from CLIR researchers. Using the ImageCLEF test collection, we simulate user interaction with a CL image retrieval system, and in particular the situation in which a user selects one or more relevant images from the top $n$. Using textual captions associated with the images, relevant images are used to create a feedback model in the `Lemur` language model for information retrieval, and our results show that feedback is beneficial, even when only one relevant document is selected. This is particularly useful fo cross language retrieval where problems during translation can create a poor initial ranked list with few relevant in the top $n$. We find that the number of relevant document in the feedback model, and the influence of the initial query on the feedback model most affect retrieval performance on the ImageCLEF test collection using the `Lemur` mixture model.

## 1 Introduction

Relevance feedback is a method aimed at improving initial free-text search results by incorporating user feedback into further iterative retrieval cycles, e.g. by expanding the initial query with terms extracted from documents selected by the user (see, e.g. [1]). Typically, the effect is to improve retrieval performance by either retrieving more relevant documents, or pushing existing relevant documents towards the top of a ranked list. However this can vary greatly across different queries, where the effect of additional query terms will improve some, but degrade others even though overall query expansion appears to improve retrieval (see, e.g. [2] and [3]).

Factors such as vocabulary mismatch between a user's search request and collection documents, the inability of the user to successfully formulate their initial search request, and mismatches due to differences between the language of the query and document collection can vary initial retrieval performance. The success of query expansion using relevance feedback may also vary because poor query terms are suggested as additional terms, caused by factors including the relevance of those documents selected for feedback to the initial query, terms

being selected from non-relevant passages, few documents available for feedback, and irrelevant terms being selected from relevant texts (see, e.g. [4]).

In this paper, we address the problem of matching images to user queries expressed in natural language, where the images are indexed by associated textual captions. The task is cross language, because the captions are expressed in one language and the queries in another, thereby requiring some kind of translation to match queries to images (e.g. query translation into the language of the document collection). Flank [5] has shown cross language image retrieval is viable on large image collections, and a program of research has been undertaken to investigate this further in the Cross Language Evaluation Forum (CLEF) [6]. Like other CLIR tasks, we have shown that retrieval performance for cross language image retrieval is affected by the quality of the translation resource, causing initial retrieval performance to vary across both language and topic [7].

The failure of query translation can vary across topics from mistranslating just one or two query terms, to not translating a term at all. Depending on the length of the query, and whether un-translated terms are important for retrieval, this can lead to unrecoverable results where the only option left to the user is to reformulate their query, or use alternative translation resources. However, there are situations when enough words are translated to recover a small subset of relevant documents, and in these cases relevance feedback can be used to expand the initial translated query with further terms and thereby improve retrieval performance by finding more relevant documents. Although previous work has shown that multilingual image retrieval is feasible on large collections [5], less investigation of relevance feedback in this scenario has occurred. In this paper, we pay particular attention to the situation in which initial retrieval is poor, caused by translation errors.

Most previous work in relevance feedback in CLIR has focused on pre and post-translation using blind (or pseudo) relevance feedback methods where the top $n$ documents are assumed to be relevant (see, e.g. [8] and [9]). Although this approach involves no user interaction, it becomes ineffective when few relevant documents appear in the top $n$ as non-relevant are also used for feedback. Past research has shown that substantial benefits can be obtained using an additional collection similar to that of the document collection in the query language during pre-translation query expansion (see, e.g. [9]), however this assumes that a similar document collection can be found in the source language. Such a resource is unlikely for many image collections, such as historic photographs or commercial photographic collections, therefore we focus on what benefits post-translation expansion can achieve.

In typical CLIR scenarios, unless the user can read and understand documents in the language of the document collection, they cannot provide relevance feedback. However, image retrieval provides a unique opportunity for exploiting relevance feedback in CLIR because for many search requests, users are able to judge relevance based on the image itself without the image caption which makes relevance judgments *language independent*. In addition to this, users are able to judge the relevance of images much faster than for document retrieval, thereby

allowing relevant documents in low rank positions to be used in the feedback process, which otherwise might not be examined if documents to retrieve were cross language texts. As Oard [10] comments: "an image search engine based on cross-language free text retrieval would be an excellent early application for cross-language retrieval because no translation software would be needed on the user's machine."

In these experiments, we show that relevance feedback can help improve initial retrieval performance across a variety of languages, especially when translation quality is poor. The paper is structured as follows: in section 2 we describe the evaluation framework including the language model IR system, our method of evaluating relevance feedback, the measure of retrieval effectiveness used and the translation resource employed. In section 3, we present results for initial retrieval, the variation of retrieval performance with changes in parameter values, and the effect of relevance feedback. Finally, in section 4 we conclude this paper and present ideas for future work.

## 2 Experimental Setup

### 2.1 The ImageCLEF Test Collection

The ImageCLEF test collection consists of a document collection, a set of user needs expressed in both natural language and with an exemplar image, and for each user need a set of relevance judgments [6]. The document collection consists of 28,133 images from the St Andrews Library photographic collection and all images have an accompanying textual description consisting of 8 fields in a semi-structured format. These fields can be used individually or collectively to facilitate image retrieval. The 28,133 captions consist of 44,085 terms and 1,348,474 word occurrences; the maximum caption length is 316 words, but on average 48 words in length. All captions are written in British English, although the language also contains colloquial expressions. Approximately 81% of captions contain text in all fields, the rest generally without the description field. In most cases the image description is a grammatical sentence of around 15 words. The majority of images (82%) are in black and white, although colour images are also present in the collection.

The test collection consists of 50 queries (topics) which are designed to provide a range of user requests to a cross language image retrieval system. Each topic contains a few keywords describing the user need (e.g. metal railway bridges) which have been translated into French, German, Italian, Dutch, Spanish and Chinese. The test collection also consists of a set of relevance judgments for each topic based primarily on the image, but also assisted by the image caption. Topics and relevance judgments are provided for an ad hoc retrieval task which is this: given a multilingual statement describing a user need, find as many relevant images as possible from the document collection. This retrieval task simulates when a user is able to express their need in natural language, but requires a visual document to fulfill their search request.

### 2.2 The Lemur Retrieval System

In the Lemur implementation of language modeling for IR, documents and queries are viewed as observations from generative[1] unigram language models (see, e.g. [11] for more information). Queries and documents are represented as estimated language models with word probabilities derived from the documents, queries and the collection as a whole. The estimated query and document language models ($\hat{\theta}_Q$ and $\hat{\theta}_D$ respectively) are compared and ranked using the KL-divergence measure, an approach which can be likened to the vector-space model of retrieval where queries and documents are represented by vectors rather than language models.

The estimated document language model $\hat{\theta}_D$ is computed from the smoothed probability of a query word seen in the document, the model smoothed by using the collection language model to estimate word probabilities when a query word is not seen in the document. Lemur supports three methods of smoothing: Jelinek-Mercer, Bayesian smoothing using Dirichlet Priors, and absolute discounting (see [11] for a comparison of retrieval performance using these smoothing methods).

Without feedback, the probability of picking a word from the estimated query model $\hat{\theta}_Q$ is computed using the maximum likelihood estimator based entirely on the query text. However, the shorter the query, the more unstable and inaccurate this estimate will be; therefore a common method of improving this estimate is to expand the query model using relevance feedback. By exploiting documents the user has judged as relevant (or using the top $n$ documents from an initial search), retrieval performance is often improved because it helps to supplement the initial user query with collection-specific words obtained from a feedback model. In Lemur, the process of feedback is to update the query language model with extra evidence contained in the feedback documents, rather than simply adding additional terms to the initial query. Given an estimated feedback model $\hat{\theta}_F$ based on a set of feedback documents $F = (d_1, d_2, .., d_n)$ and the original query model $\hat{\theta}_Q$, the updated query model $\hat{\theta}_{Q'}$ is computed from interpolating the two models:

$$\hat{\theta}_{Q'} = (1 - \alpha)\hat{\theta}_Q + \alpha\hat{\theta}_F \quad (1)$$

The *feedback coefficient* $\alpha$ controls the influence of the feedback model. If $\alpha = 0$ estimates are based entirely on the initial query; whereas if $\alpha = 1$ estimates are based entirely on the set of feedback documents. Allowing the impact of the feedback model to vary can be used to investigate the effects of including the initial query in the updated query model on retrieval performance. This is particularly relevant to cross language evaluation where the quality of the initial query is, amongst other factors, dependent on the quality of the translation resource. In Lemur, two methods can be used to estimate the feedback model $\hat{\theta}_F$ : (1) a two component mixture model, and (2) divergence/risk minimisation and both approaches have been shown to be effective for relevance feedback

[1] Generative in the sense that a query or document is generated by picking words from the query or document language model.

[12]. In both methods, a second parameter $\lambda$ is used to control whether word estimates for the feedback model are derived entirely from the set of feedback documents ($\lambda = 0$), the collection language model ($\lambda = 1$), or somewhere in between ($0 < \lambda < 1$). This can be used to promote words which are common in the feedback documents, but not common according to the collection language model.

In these experiments, we use the KL-divergence language model with the absolute discounting method of smoothing with $\Delta = 0.7$. We focus primarily on the two-component mixture model to estimate word probabilities in the feedback model because initial experiments showed this gave higher retrieval performance than the divergence/risk minimisation approach. In all experiments, we stem both query and documents using the Porter stemmer, and remove stopwords using a list of 249 common English words. Images are indexed on all caption fields for both the retrieval and feedback indices.

### 2.3 The translation Resource: Systran

In these experiments, we translated cross language topics into English using Systran, a popular machine translation (MT) resource which has been used in past CLIR research (see, e.g. [5]). Using Systran as a translation resource, like any form of translation method, will result in erroneous queries because of difficulties encountered during translation including: short queries resulting in little if none syntactic structure to exploit, errors in the original cross language text (e.g. spelling mistakes or incorrect use of diacritics), lack of coverage by the translation lexicon, incorrect translation of phrases, mis-translation of proper names, and incorrect translation of ambiguous words (e.g. selecting the wrong sense of a noun or verb). The effect of translation errors on retrieval performance for ImageCLEF topics is discussed in [7]. For more information on Systran, see e.g. [13].

### 2.4 Evaluating Relevance Feedback

To evaluate the effects of relevance feedback on cross language image retrieval, we used the following experimental procedure. First, the cross languages topic titles from the ImageCLEF test collection were translated into English using Systran. Using the translated topic titles, we performed an initial retrieval, the baseline, using the set of *strict intersection* relevance judgments that accompany the test collection. To simulate relevance feedback, we extracted relevant documents from the initial ranked lists to create sets of relevance judgments which contain only those relevant found within rank position $n$. A proportion of these relevant were used to build a feedback model from which additional query terms were selected. A second retrieval was performed with the updated query model to simulate the situation in which a user selects relevant images from the top $n$. For both initial retrieval and feedback, the answer set was limited to 1000 images.

At the end of the feedback cycle, the effect of relevance feedback is evaluated by comparing it to initial performance. In some cases initial retrieval is poor

as no relevant images appear in the top $n$. In these situations, `Lemur` does not generate any query terms and retrieval after feedback results in an empty answer set and no evaluation scores. This affects retrieval after feedback and makes results appear worse than they are. Therefore, rather than evaluate an empty answer set, the initial retrieval results were used instead. This means results with no relevant in the top $n$ will have no effect on retrieval performance.

Various approaches to evaluating the effects of relevance feedback have been suggested; most derived from the early work on the SMART system (see, e.g. [14]) and include rank freezing, using a residual collection, and using test and control groups. Methods proposed in previous work attempt to eliminate improvements in retrieval due to re-ranking, rather than finding new relevant documents. Measures such as average precision are based on the number of relevant, and rank position and unless documents used for feedback are eliminated from the results, a large increase in performance can be observed due to re-ranking. This is particularly noticeable in feedback over several iterations. In these experiments documents from the feedback set are not disregarded after feedback, but this does not affect the validity of our results because we focus on an evaluation measure based on recall at $n$ not rank position where increases in retrieval performance can only come from new relevant documents found in the top $n$ after feedback. We compute an absolute measure of effectiveness before and after feedback and compare the results from this in our evaluation. Although from the user-perspective documents selected for relevance feedback would appear again in the answer set, it is not uncommon to find image retrieval systems which offer this kind of relevance feedback interface (e.g. COMPASS[2] and WebSeek[3]).

### 2.5 Evaluation Measures

In these experiments, the goal of relevance feedback is to find further relevant documents, particularly when few relevant documents appear in the top $n$. We use $n = 100$ as a cut-off point for evaluation because from our experiences in building the ImageCLEF test collection, judging the relevance for a large number of images is feasible. The goal is then given that users identify relevant images from the top 100, to automatically find further relevant images (if any more exist) and bring these into the top 100. In this evaluation, a recall measure is used to evaluate retrieval performance in which the movement of feedback documents is ignored and we assume that finding new relevant images is more important to the user than their rank position. Precision at $n$ measures the proportion of relevant documents found in the top $n$ rank positions, regardless of position, and precision at 10/20 is often quoted in relevance feedback. Because we assume users to search the first 100 images, we use precision at 100 ($P_{100}$). However, this measure is affected by the total number of relevance assessments in the test collection. This does not affect precision at 10 or 20, but in the ImageCLEF test

[2] COMPASS: http://compass.itc.it/
[3] WebSeek: http://www.ctr.columbia.edu/WebSEEk/

collection 90% of topics have fewer than 100 relevant which causes the $P_{100}$ score to be less than 1, even if all relevant images are found in the top 100.

Therefore a normalised precision at 100 measure is used that normalises $P_{100}$ with respect to the number of relevant documents for each query. This measures the proportion of relevant documents retrieved in the top 100 (i.e. a recall measure), rather than the proportion of the top 100 which are relevant. Given a $P_{100}$ score and a set of relevance judgments for a query $\Phi$ (the size given by $\mid \Phi \mid$), the normalised precision at 100 score $P_{norm100}$ is given by:

$$P_{norm100} = \frac{P_{100} \times 100}{min(100, \mid \Phi \mid)} \tag{2}$$

The normalised precision score ranges from 0 indicating no relevant in the top 100, to 1 which indicates either all relevant are in the top 100 (if $\mid \Phi \mid \leq 100$) or that all top 100 documents are relevant (if $\mid \Phi \mid > 100$). We also compute the number of good and bad topics within the top $n$. A *good* topic is one in which all relevant documents are found within the top $n$ and for which relevance feedback would not be necessary because no further relevant documents exist. For $n = 100$, these are indicated by $P_{norm100} = 1$. A *bad* topic is one in which no relevant documents are found in the top $n$ and for which relevance feedback will therefore be unsuccessful (unless the user is willing to go beyond the top $n$). For $n = 100$, these are indicated by $P_{norm100} = 0$. In the results we still quote Mean Average Precision (MAP) scores to show the effects of re-ranking, recall (precision at 1000), $P_{100}$, and $P_{norm100}$.

## 3 Results and Discussion

### 3.1 Initial Retrieval Performance

Table 1 summarises retrieval performance for initial retrieval without feedback at $n = 100$. Also shown are the number of *failed* topics, those which either return no images at all, or no relevant in the top 1000. These topics cannot be helped and require the user to reformulate their initial query, or turn to browsing the collection to find relevant images. The number of topics which can be improved are those which are not classed as good or bad.

According to MAP, French would appear to perform the best out of the cross language results at an average of 75.5% of monolingual (85% of $P_{norm100}$). German and Italian perform similarly, but there are differences. Although the German MAP score is higher than Italian, the topic characteristics tell a different story. German has the highest number of good topics cross language, but also shares the highest number of bad topics (with Dutch), and the highest number of failed topics. Since the user tends not to know the number of relevant images for a topic, then the benefits of systems with a high number of good topics are probably less significant than a system which returns no relevant in the top 100. The effects of rank position on MAP can be observed, the $P_{norm100}$ scores higher reflecting recall rather than the position of the image. Table 1 also shows

**Table 1.** A summary of initial retrieval performance averaged across all topics

| | MAP | %mono MAP | Recall | Avg $P_{100}$ | Avg $P_{norm100}$ | %mono $P_{norm100}$ | Topics good | Topics bad | Topics to imp. | Topics failed |
|---|---|---|---|---|---|---|---|---|---|---|
| Mono | 0.5514 | - | 0.8129 | 0.1800 | 0.8132 | - | 22 | 1 | 27 | 1 |
| German | 0.4042 | 73.3% | 0.6173 | 0.1272 | 0.6450 | 79.3% | 19 | 9 | 22 | 6 |
| French | 0.4161 | 75.5% | 0.8489 | 0.1624 | 0.6912 | 85.0% | 18 | 4 | 28 | 2 |
| Italian | 0.4018 | 72.9% | 0.7898 | 0.1442 | 0.6626 | 81.5% | 14 | 7 | 29 | 3 |
| Dutch | 0.3806 | 69.0% | 0.7018 | 0.1134 | 0.5832 | 72.0% | 15 | 9 | 26 | 4 |
| Spanish | 0.3940 | 71.5% | 0.7638 | 0.1464 | 0.6504 | 80.0% | 16 | 4 | 30 | 1 |
| Chinese | 0.2794 | 50.7% | 0.7255 | 0.1156 | 0.5416 | 67.0% | 13 | 8 | 29 | 5 |

the proportion of relevant images found in the top 100, for example on average around 81% of all relevant images are found in the top 100 for monolingual and 54% for Chinese retrieval. The results illustrate that the $P_{100}$ and $P_{norm100}$ scores also vary because $P_{100}$ measure is sensitive to the number of relevant judgments for a topic. For example, $P_{100}$ for Dutch is lower than the $P_{100}$ for Chinese, but vice-versa for the $P_{norm100}$ scores.

**Table 2.** Average rank position (geometric mean) of first 5 relevant images and average number of relevant in top 100 (incl. standard deviation)

| | 1st rel | 2nd rel | 3rd rel | 4th rel | 5th rel | Avg rel in top 100 |
|---|---|---|---|---|---|---|
| Mono | 1.5 | 3.4 | 5.9 | 9.1 | 10.8 | 18 (19.29) |
| German | 3.8 | 6.2 | 9.8 | 14.2 | 16.4 | 16 (16.11) |
| French | 3.4 | 6.5 | 11.4 | 16.3 | 18.0 | 18 (21.00) |
| Italian | 2.4 | 5.4 | 9.9 | 13.7 | 15.5 | 17 (15.99) |
| Dutch | 4.1 | 6.6. | 10.4 | 15.0 | 17.2 | 14 (12.18) |
| Spanish | 3.2 | 7.4 | 11.3 | 14.9 | 17.8 | 16 (16.04) |
| Chinese | 5.3 | 12.3 | 17.8 | 22.2 | 28.9 | 14 (17.46) |

Table 2 shows the average position of the first 5 relevant images in results from each language and the mean number of relevant in the top 100, ignoring those topics for which $P_{norm100} = 0$. In general, relevant images for cross language retrieval generally appear in lower rank positions than for monolingual and in particular users are going to have to search the furthest in the Chinese and Dutch results. On average, users will find fewer relevant images in cross language retrieval than monolingual, resulting in fewer available images for relevance feedback in these cases. Again, Chinese and Dutch results have, on average, fewest relevant images in the top 100.

### 3.2 Effect of Parameters on Relevance Feedback

At least four parameters can be varied using `Lemur` for relevance feedback, including: $\alpha$, $\lambda$, the number of relevant documents and the number of selected feedback terms. In this section we show that retrieval performance across language is most affected by $\alpha$ and the number of relevant selected for feedback.

Varying the influence of the collection language model has little effect on retrieval performance using this retrieval system, feedback model and test collection, and these results coincide with the findings of Zhai and Lafferty [12]. Using default settings of 30 terms selected and $\alpha = 0.5$, we find that with 1 feedback document, as $\lambda$ increases, retrieval performance decreases slightly; with 5 feedback documents the reverse is true, although the variation in retrieval performance is under 2%. In all cases, the $P_{norm100}$ score after relevance feedback is always higher than for initial retrieval, and a similar pattern emerges in the cross language results.

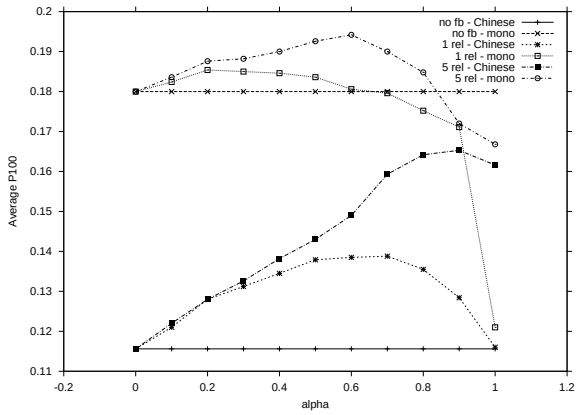


**Fig. 1.** A plot of average $P_{100}$ for monolingual and Chinese retrieval using 1 or 5 relevant documents for feedback and varying $\alpha$

Using default settings of $\alpha = \lambda = 0.5$, and selecting between 5 and 100 terms for 1 and 5 feedback documents, we find on average that in both cases selecting fewer than 30 terms gives lowest performance. With 5 feedback documents, retrieval performance reaches its highest with 30 terms, and highest at 100 terms with 1 feedback document, although little variation exists after 30 terms are selected. The results are also similar across language, and again, regardless of the number of terms selected retrieval performance never goes below initial performance.

The parameters which most affect retrieval performance are $\alpha$ and the number of relevant documents used in the feedback model. Fig. 1 shows the average $P_{100}$ score as $\alpha$ is varied for both monolingual and cross language retrieval, and

using 1 and 5 relevant documents in the feedback model. Although this coincides with the findings of Zhai and Lafferty, what is interesting are differences across language which result from the quality of the initial query and its influence on term estimates. In cross language retrieval, the initial query tends to be worse than monolingual because of translation errors. In the feedback model this means that when the initial query is good (monolingual), then it should be used to influence term weights. In Fig. 1 with 1 feedback document, retrieval performance is highest with $\alpha = 0.2$, whereas with 5 documents, $\alpha = 0.6$ is highest as the feedback model gives better term estimates. Retrieval performance goes below the baseline particularly when $\alpha = 1$ and the initial query is ignored (see Table 3), but the case is very different for other languages, especially Chinese.

**Table 3.** Results after relevance feedback when ignoring the initial query from the feedback model ($\alpha = 1$) and selecting 30 terms from up to 10 relevant documents

| | Avg $P_{norm100}$ | %increase | Topics good | Topics bad |
|---|---|---|---|---|
| Mono | 0.7961 | -2.1% | 18 | 1 |
| German | 0.6777 | 5.0% | 17 | 9 |
| French | 0.7416 | 7.3% | 17 | 4 |
| Italian | 0.6968 | 5.2% | 14 | 7 |
| Dutch | 0.6587 | 12.9% | 15 | 9 |
| Spanish | 0.7403 | 13.8% | 18 | 4 |
| Chinese | 0.6804 | 25.6% | 14 | 8 |

In Chinese retrieval, the initial query tends to be worse and therefore larger values of $\alpha$ give higher results than for monolingual. Given 1 feedback document, $\alpha = 0.7$ is highest, whereas for 5 documents $\alpha = 0.9$ is highest. What is most surprising is that even if the initial query is ignored and excluded from the query after feedback ($\alpha = 1$), retrieval performance is still above the baseline, especially when more relevant are used to estimate term probabilities. Upon inspection, we find a possible cause for this is that terms which exist in the initial monolingual query are often generated from the feedback model in the cross language case when the initial query is not effective. In these cases discarding the initial query makes little difference and in fact results in high improvements in $P_{norm100}$ scores over initial retrieval as shown in Table 3. Fig. 2 shows for Chinese retrieval that $\alpha > 0.7$ will give highest results. This trend is similar across language, especially for those languages in which initial retrieval is poor because of translation quality, e.g. Dutch and Italian.

Fig. 2 shows the effects of varying the number of feedback documents between 1 and up to 10 (whether the maximum number of relevant are found will be language and topic dependent) at various values of $\alpha$, with $\lambda = 0.5$ and 30 feedback terms selected for query expansion. As before, the results reflect the

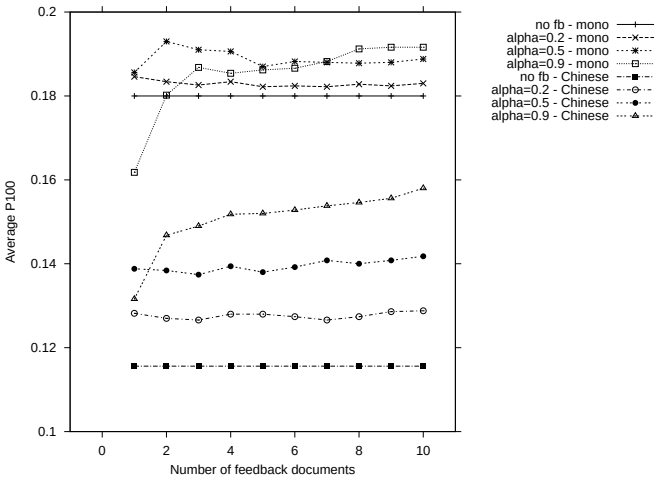


**Fig. 2.** A plot of average $P_{100}$ for monolingual and Chinese retrieval varying $\alpha$ and using up to 1 to 10 relevant images in the feedback model

effects of $\alpha$ on retrieval performance in both monolingual and cross language for Chinese retrieval. However, in the monolingual case retrieval performance tends to level out and even decreases with $\alpha = 0.5$ because the initial retrieval is actually already high and cannot be improved even with more relevant documents found by the user. For $\alpha = 0.9$, retrieval performance levels out after 8 feedback documents.

In the cross language case, again the results are different. For $\alpha = 0.9$, because emphasis is placed on the relevant documents to estimate the query model, the more images judged relevant by the user, the better. For $\alpha = 0.5$, results are relatively stable and it would seem that the number of feedback documents has little effect on retrieval performance. Generally more feedback documents improve word estimates in the feedback model because terms which better generalise the topic concept are promoted over more specific terms emphasised with fewer relevant documents. For most topics, this tends to be reflected in the query terms where more numbers, dates and proper names are generated. Using more feedback documents also tends to improve the discrimination between which feedback words are related to the topic, and those acting as "noise".

### 3.3 Relevance Feedback across Language

Using the `Lemur` mixture model on the ImageCLEF collection, retrieval performance depends on parameter selection as previously shown. This section shows the degree of increase one might expect in retrieval performance for each language in the situations where the user selects just 1 relevant image, and up to 10 relevant from the top 100. The results are shown in Tables 4 and 5 where

$\alpha = \lambda = 0.5$ and 30 feedback terms are selected from the feedback model. Because we know that results will change depending on $\alpha$, we also show the maximum increase of $P_{norm100}$ obtained by testing all values of $\alpha$ between 0.1 and 0.9. These settings are, of course, dependent on the ImageCLEF test collection, topics and our retrieval system, but they do provide an idea of the kind of performance increase one could obtain if automatic parameter selection was used.

**Table 4.** A summary of retrieval performance with relevance feedback from selecting 30 terms from up to 10 relevant in the top 100 ($\alpha = \lambda = 0.5$)

| | MAP | Recall | Avg $P_{100}$ | Avg $P_{norm100}$ | %increase $P_{norm100}$ | Topics good | Topics bad | Maximum increase ($\alpha$) |
|---|---|---|---|---|---|---|---|---|
| Mono | 0.6649 | 0.8356 | 0.1888 | 0.8440 | 3.8% | 25 | 1 | 4.8% (0.6) |
| German | 0.5831 | 0.6222 | 0.1436 | 0.6848 | 6.2% | 23 | 9 | 9.6% (0.8) |
| French | 0.5976 | 0.8587 | 0.1816 | 0.7668 | 10.9% | 24 | 4 | 13.5% (0.8) |
| Italian | 0.5765 | 0.7915 | 0.1624 | 0.7096 | 7.1% | 17 | 7 | 11.4% (0.8) |
| Dutch | 0.4954 | 0.6995 | 0.1304 | 0.6348 | 8.9% | 17 | 9 | 18.1% (0.9) |
| Spanish | 0.5820 | 0.8172 | 0.1650 | 0.7380 | 13.5% | 21 | 4 | 19.2% (0.9) |
| Chinese | 0.4806 | 0.7620 | 0.1386 | 0.6170 | 13.9% | 17 | 8 | 26.9% (0.9) |

The benefits of relevance feedback on retrieval performance are clearly seen in Tables 4 and 5. As expected, when initial retrieval is higher the effects are less dramatic and most benefit is gained for Spanish, Dutch and Chinese retrieval. Unsurprisingly, the number of bad topics stays the same as initial retrieval because when no relevant are found, results from feedback are ignored. Relevance feedback, however, does increase both the number of good topics and average $P_{norm100}$ scores in both sets of results. The average increase in performance for $\alpha = \lambda = 0.5$ is similar, but because $\alpha$ is not at its optimum in Table 4, the results are lower than could be obtained if some kind of parameter selection were used. In general more relevant documents are beneficial across all languages (especially non-English), and in the cross language situation, less influence from the initial query in term reweighting gives better results (i.e. $\alpha \geq 0.8$). The results in Table 5 show that for most languages, retrieval performance can be improved by users selecting just one relevant in the top 100. It would seem that encouraging the user to select either just one or up to 10 relevant in the top 100, which in a traditional CLIR task would be unreasonable but possible in a CL image retrieval task, would offer substantial benefits for the user and enable more relevant images to be shown to them in the top 100.

Finally, Fig. 3 shows the effects of relevance on individual topics for monolingual retrieval, and Chinese using 1 and up to 10 relevant by plotting the difference between the $P_{norm100}$ results before and after feedback. Many monolingual topics are not improved and there are topics where both the monolingual

**Table 5.** A summary of retrieval performance with relevance feedback from selecting 30 terms from 1 relevant in the top 100 ($\alpha = \lambda = 0.5$)

| | MAP | Recall | Avg $P_{100}$ | Avg $P_{norm100}$ | %increase $P_{norm100}$ | Topics good | Topics bad | Maximum increase ($\alpha$) |
|---|---|---|---|---|---|---|---|---|
| Mono | 0.6317 | 0.8554 | 0.1856 | 0.8378 | 3.0% | 24 | 1 | 2.3% (0.2) |
| German | 0.5187 | 0.6703 | 0.1448 | 0.6894 | 6.9% | 22 | 9 | 6.9% (0.6) |
| French | 0.5579 | 0.8450 | 0.1694 | 0.7506 | 8.9% | 24 | 4 | 8.9% (0.5) |
| Italian | 0.5134 | 0.7993 | 0.1564 | 0.7018 | 5.9% | 17 | 7 | 5.9% (0.5) |
| Dutch | 0.4708 | 0.7201 | 0.1284 | 0.6312 | 8.2% | 17 | 9 | 8.7% (0.7) |
| Spanish | 0.5581 | 0.8276 | 0.1636 | 0.7328 | 12.7% | 23 | 4 | 13.7% (0.6) |
| Chinese | 0.4332 | 0.8109 | 0.1338 | 0.6130 | 13.2% | 17 | 8 | 18.3% (0.7) |

and Chinese results are improved similarly (e.g. topic 14, 15 and 16). In these topics, the Chinese translations are similar to the monolingual, but both are initial queries which benefit from relevance feedback. Topics in which the monolingual differences are much lower than the Chinese are those in which the initial query is poor. Generally, the effect of more relevant is to improve retrieval performance (e.g. topic 3), but there are topics when the situation is reversed (e.g. topic 50). Although overall the retrieval performance is similar for Chinese using 1 or 10 relevant (for $\alpha = \lambda = 0.5$), the effect on individual topics is to increase the number of topics made worse using 1 relevant, and increase the number of topics improved using 10 relevant. This trend occurs across all languages.

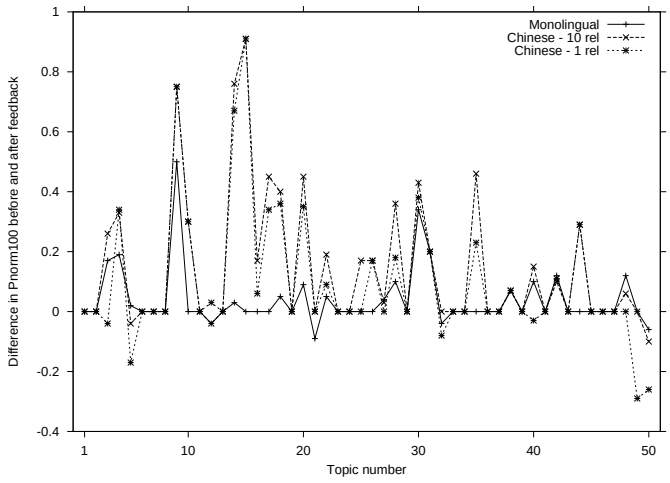


**Fig. 3.** A plot of the difference in $P_{norm100}$ before and after feedback for monolingual and Chinese retrieval (from 1 and up to 10 relevant)

## 4 Conclusions and Future Work

In this paper, we have explored the use of relevance feedback in cross language image retrieval by simulating user involvement based on relevance assessments from the ImageCLEF test collection. We have shown that relevance feedback can improve retrieval performance (using normalised $P_{100}$) across all languages (including monolingual) using the `Lemur` feedback model and ImageCLEF topics for an ad hoc retrieval task. As a cross language task, characteristics include: (1) the number of failed topics varies across language depending on the quality of translation, (2) across languages, relevant images tend to appear in lower rank positions, and (3) fewer relevant images are available for relevance feedback than in the monolingual case. Despite these issues, however, image retrieval is a CL problem in which it is feasible that users are willing and able to search many images, and judge relevance even when they are unable to understand the language of the captions because judgments are also made on the basis of the image itself which is language independent. We have shown in these experiments that on average retrieval performance can be improved by as much as 13.9% (for Chinese) and with parameter selection could increase to 26.9%.

Using the two-component mixture feedback model in `Lemur`, the model is found to behave similar to previous results in the monolingual case using PRF, but across language using relevance feedback, the results are different. In cross language retrieval, the quality of the initial query reflects the quality of the translation resource, and for languages where translation tends to be worse, e.g. Chinese, we find that the influence of the initial query should be reduced when calculating term weights in the feedback model. We also find that the more documents used to build the feedback model, the better word estimates are which result in more topics improving in retrieval performance, and reducing those which decrease. The number of feedback terms and the influence of the collection model on term estimates has minimal effects on retrieval for this task.

As an initial investigation using relevance feedback to improve the results of retrieval performance in CL image retrieval, the results are encouraging enough to suggest that relevance feedback would play a useful and important role when designing a CL image retrieval system. Given the improvements demonstrated on the ImageCLEF test collection, we would recommend that designers of the CL image retrieval system consider how users can be encouraged to interact with their system and provide feedback. In addition, getting users to identify more relevant images seems beneficial across language which could be achieved through improving initial retrieval, or more feedback iterations which also requires a supporting user interface. A final problem is how to deal with failed queries, particularly in a cross language system where queries might fail because of translation errors. One solution could be to provide a browsing interface which may help to encourage users to continue pursuing their initial search to find at least 1 relevant which would enable relevance feedback to be utilised.

In future work, we aim to pursue a number of directions in relevance feedback for CL image retrieval. In particular, we aim to investigate whether results are improved by using only a limited number of caption fields for retrieval, whether

removing domain stopwords (i.e. "noise") can improve results, what the effects are if the user were to select irrelevant documents for feedback (particularly important in image retrieval where relevance assessment is highly subjective), and whether relevant images used for feedback give similar retrieval performance. This last point addresses topics which are general, e.g. "Postcards of London", where relevant images are likely to exhibit captions which vary in content (e.g. pictures of famous London sights, locations, people etc.). Finally, we aim to investigate automatic parameter selection in the `Lemur` feedback model, the effects of multiple feedback iterations on retrieval performance, and whether content-based retrieval methods can help improve retrieval performance and complement the text-based approaches.

## References


1. Efthimiadis, E., Robertson, S.: Feedback and interaction in information retrieval. In Oppenheim, C., ed.: Perspectives in Information Management, London, Butterworths (1989) 257–272
2. Harman, D.: Relevance feedback revisited. In: Proceedings of SIGIR1992. (1992) 1–10
3. Mitra, M., Singhal, A., Buckley, C.: Improving automatic query expansion. In: Proceedings of SIGIR1998. (1998) 206–214
4. Magennis, M., van Rijsbergen, C.J.: The potential and actual efectiveness of interactive query expansion. In: Proceedings of SIGIR1997. (1997) 324–332
5. Flank, S.: Cross-language multimedia information retrieval. In: Proceedings of Applied Natural Language Processing and the North American Chapter of the Association for Computational Linguistics (ANLP-NAACL2000). (2000)
6. Clough, P., Sanderson, M.: The CLEF cross language image retrieval track. In: Submission, to appear. (2003)
7. Clough, P., Sanderson, M.: Assessing translation quality for cross language image retrieval. In: Submission, to appear. (2003)
8. Ballesteros, L., Croft, B.: Resolving ambiguity for cross-language retrieval. In: Proceedings of SIGIR1998. (1998) 64–71
9. McNamee, P., Mayfield, J.: Comparing cross-language query expansion techniques by degrading translation resources. In: Proceedings SIGIR2002. (2002) 159–166
10. Oard, D.: Serving users in many languages. D-Lib magazine (1997)
11. Zhai, C., Lafferty, J.: A study of smoothing methods for langauge models applied to ad hoc information retrieval. In: Proceedings of SIGIR'2001. (2001) 334–342
12. Zhai, C., Lafferty, J.: Model-based feedback in the kl-divergence retrieval model. In: Tenth International Conference on Information and Knowledge Management (CIKM2001). (2001) 403–410
13. Hutchins, W., Somers, H.: An Introduction to machine Translation. Academic Press, London, England (1986)
14. Salton, G.: The SMART Retrieval System - Experiments in Automatic Document Processing. Prentice-Hall Inc., Englewood Cliffs, N.J. (1971)