Dublin City University at CLEF 2005: Experiments with the ImageCLEF St Andrew's Collection

Gareth J. F. Jones and Kieran McDonald

Centre for Digital Video Processing & School of Computing
Dublin City University, Dublin 9, Ireland
email: {gjones,kmcdon}@computing.dcu.ie

Abstract. The aim of the Dublin City University's participation in the CLEF 2005 ImageCLEF St Andrew's Collection task was to explore an alternative approach to exploiting text annotation and content-based retrieval in a novel combined way for pseudo relevance feedback (PRF). This method combines evidence from retrieved lists generated using text-based and content-based retrieval to determine which documents will be assumed relevant for the PRF process. Unfortunately the experimental results show that while standard text-based PRF improves upon a no feedback text-only baseline, at present our new approach to combining evidence from text-based and content-based retrieval does not give further improvement.

1 Introduction

Dublin City University's participation in the CLEF 2005 ImageCLEF St Andrew's collection task [1] explored a novel approach to pseudo relevance feedback (PRF) combining evidence from separate text-based and content-based retrieval runs. The underlying text retrieval system is based on a standard Okapi model for document ranking and PRF [2]. Three sets of experiments are reported for the following topic languages: Chinese (simplified), Dutch, French, German, Greek, Italian, Japanese, Portuguese, Russian and Spanish (european), along with corresponding monolingual English results as a baseline for comparison. Topics were translated into English using the online Babelfish machine translation engine. The first set of experiments establish baseline retrieval performance without PRF, the second set of experiments incorporate a standard PRF stage, and finally the third set investigates our new combined method for PRF.

This paper is organised as follows: Section 2 briefly outlines the details of our standard retrieval system and describes our novel PRF method, Section 3 gives results for our experiments, and finally Section 4 concludes the paper.

2 Retrieval System

2.1 Standard Retrieval Approach

Our basic experimental retrieval system is a local implementation of the standard Okapi retrieval model [2]. Documents and search topics are processed to remove stopwords from the standard SMART list, and suffix stripped using the Snowball implementation of Porter stemming [3] [4]. The resulting terms are weighted using the standard BM25 weighting scheme with parameters ($k1$ and b) selected using the CLEF 2004 ImageCLEF test collection data as a training set.

Standard PRF was carried out using query expansion. The top ranked documents from a baseline retrieval run were assumed relevant. Terms from these documents were ranked using the Robertson selection value (RSV) [2], and the top ranked terms added to the original topic statement. The parameters of the PRF stage were again selected using the CLEF 2004 ImageCLEF test set.

2.2 Combining Text and Content-based Retrieval for PRF

The preceding text-based retrieval methods have been shown to work reasonably effectively for the St Andrew's ImageCLEF task in earlier workshops [5]. However, this approach makes no use of the document or topic images. In our participation in the CLEF 2004 ImageCLEF task we attempted to improve text-only based retrieval by performing a standard data fusion summation combination of retrieved ranked lists from text-only retrieval and the provided context-based retrieval lists generated using the GIFT/Viper system. The results of these combined lists showed little difference from the text-only runs [5].

Analysis of the GIFT/Viper only runs for the CLEF 2004 task showed them to have very poor recall, but reasonable precision at high cutoff levels. However, further investigation of this showed that this good high cutoff precision is largely attributable to a good match on the topic image which is part of the document collection. This topic image is relevant for the topic and typically found at rank position one. Our analysis suggests that there is little to be gained from data fusion in this way, certainly when content-based retrieval is based on low-level features. Indeed it is perhaps surprising that this method does not degrade performance relative to the text-only retrieval runs.

Nevertheless, we were interested to see if the evidence from content-based retrieval runs might be usefully combined with the text-only retrieval runs in a different way. For our CLEF 2005 experiments we hypothesized that documents retrieved by both the text-based and content-based methods are more likely to be relevant than documents retrieved by only one system. We adapted the standard PRF method to incorporate this hypothesis as follows. Starting from the top of lists retrieved independently using text-based retrieval with the standard PRF method and content-based retrieval, we look for documents retrieved by both systems. Documents retrieved by both systems are assumed to be relevant and are used to augment the assumed relevant document set for a further run of

Table 1. Text-only baseline retrieval runs using Babelfish topic translation

		English	Chinese (s)	Dutch	French	German	Greek
Prec.	5 docs	0.557	0.264	0.471	0.393	0.486	0.379
	10 docs	0.500	0.254	0.436	0.375	0.418	0.404
	15 docs	0.460	0.250	0.402	0.355	0.374	0.386
	20 docs	0.427	0.230	0.377	0.323	0.343	0.370
Av Precision		0.355	0.189	0.283	0.244	0.284	0.249
% chg.		—	-46.8%	-20.3%	-31.3%	-20.0%	-29.9%
Rel. Ret.		1550	1168	1213	1405	1337	1107
chg. Rel. Ret.		—	-382	-337	-145	-213	-443

		English	Italian	Japanese	Portuguese	Russian	Spanish (e)
Prec.	5 docs	0.557	0.300	0.393	0.407	0.379	0.336
	10 docs	0.500	0.296	0.368	0.368	0.354	0.325
	15 docs	0.460	0.269	0.336	0.343	0.329	0.307
	20 docs	0.427	0.266	0.311	0.323	0.314	0.280
Av Precision		0.355	0.216	0.259	0.243	0.247	0.207
% chg.		—	-39.2%	-27.0%	-31.5%	-30.4%	-41.7%
Rel. Ret.		1550	1181	1304	1263	1184	1227
chg. Rel. Ret.		—	-369	-246	-287	-366	-323

the text-only based retrieval system with the standard query expansion PRF method.

For this investigation content-based retrieval used our own image retrieval system based on standard low-level colour, edge and texture features. The colour comparison was based on 5×5 regional colour with HSV histogram dimensions $16 \times 4 \times 4$. Edge comparison used Canny edge with 5×5 regions quantized into 8 directions. Texture matching was based on the first 5 DCT co-efficients, each quantized into 3 values for 3×3 regions. The scores of the three components were then combined in a weighted sum and the overall summed scores used to rank the content-based retrieved list.

3 Experimental Results

The settings for the Okapi model were optimized using the CLEF 2004 ImageCLEF English language topics as follows: $k1 = 1.0$ and $b = 0.5$. These parameters were used for all test runs reported in this paper.

3.1 Baseline Retrieval

Table 1 shows baseline retrieval results for the Okapi model without application of feedback. Monolingual results for English topics are shown in the left side column for each row. Results for each translated topic language relative to English are then shown in the other columns. From these results we can see that cross-language performance is degraded relative to monolingual by between

Table 2. Text-only PRF retrieval runs using Babelfish topic translation

		English	Chinese (s)	Dutch	French	German	Greek
Prec.	5 docs	0.529	0.257	0.450	0.407	0.443	0.439
	10 docs	0.500	0.275	0.425	0.407	0.407	0.432
	15 docs	0.467	0.274	0.407	0.393	0.393	0.410
	20 docs	0.432	0.261	0.382	0.373	0.375	0.396
Av Precision		0.364	0.213	0.308	0.283	0.308	0.302
% chg.		—	-41.5%	-15.4%	-22.3%	-15.4%	-17.0%
Rel. Ret.		1648	1320	1405	1580	1427	1219
chg. Rel. Ret.		—	-328	-243	-68	-221	-429

		English	Italian	Japanese	Portuguese	Russian	Spanish (e)
Prec.	5 docs	0.529	0.264	0.350	0.379	0.371	0.336
	10 docs	0.500	0.279	0.346	0.346	0.357	0.321
	15 docs	0.467	0.255	0.326	0.324	0.350	0.295
	20 docs	0.432	0.245	0.329	0.316	0.338	0.286
Av Precision		0.354	0.215	0.268	0.247	0.280	0.224
% chg.		—	-40.9%	-26.4%	-32.1%	-23.1%	-38.5%
Rel. Ret.		1648	1223	1331	1364	1335	1360
chg. Rel. Ret.		—	-425	-317	-284	-313	-288

around 20% and 45% for the different topic languages with respect to MAP, and by between 150 and 450 for the total number of relevant documents retrieved. These results are in line with those that would be expected for short documents with cross-language topics translated using a standard commercial machine translation system.

3.2 Standard Pseudo Relevance Feedback

Results using the CLEF 2004 ImageCLEF data with the English language topics were shown to be optimized on average by assuming the top 15 documents retrieved to be relevant and by adding the resulting top 10 ranked terms to the original topic, with the original terms upweighted by a factor of 3.5 relative to the expansion terms.

Table 2 shows results for applying PRF with these settings. The form of the results table is the same as that in Table 1. From this table we can see that PRF is effective for this task for all topic languages. Further the reduction relative to monolingual retrieval in each case is also generally reduced. Again this trend is commonly observed for cross-language information retrieval tasks.

Performance for individual topic languages can be improved by selecting the parameters separately, but we believed that optimizing for individual topic languages would lead to overfitting to the training topic set. To explore this issue, we performed an extensive set of post evaluation experiments varying $k1$ and b using the CLEF 2005 test collection. Results of these experiments showed that in all cases average precision and the total number of relevant documents

Table 3. PRF retrieval runs incorporating text and image retrieval evidence using Babelfish topic translation

		English	Chinese (s)	Dutch	French	German	Greek
Prec.	5 docs	0.529	0.264	0.443	0.407	0.443	0.414
	10 docs	0.504	0.268	0.432	0.411	0.414	0.429
	15 docs	0.460	0.271	0.402	0.393	0.391	0.405
	20 docs	0.432	0.259	0.375	0.373	0.371	0.393
Av Precision		0.365	0.210	0.306	0.282	0.308	0.298
% chg.		—	-42.7%	-16.2%	-22.7%	-15.6%	-18.4%
Rel. Ret.		1652	1318	1405	1578	1428	1218
chg. Rel. Ret.		—	-334	-247	-74	-224	-434

		English	Italian	Japanese	Portuguese	Russian	Spanish (e)
Prec.	5 docs	0.529	0.264	0.343	0.371	0.371	0.343
	10 docs	0.504	0.279	0.350	0.350	0.354	0.318
	15 docs	0.460	0.248	0.321	0.319	0.350	0.291
	20 docs	0.432	0.241	0.325	0.309	0.339	0.284
Av Precision		0.365	0.215	0.268	0.247	0.279	0.224
% chg.		—	-41.1%	-26.6%	-32.3%	-23.6%	-38.6%
Rel. Ret.		1652	1227	1336	1366	1331	1361
chg. Rel. Ret.		—	-425	-316	-286	-321	-291

retrieval can be improved slightly. In a few cases relatively large improvements were observed (for example, for PRF with Japanese topics average precision improved from 0.268 to 0.303, and with Italian topics from 0.215 to 0.266). There was a very wide variation in the optimal $k1$ and b for the various topic languages, and often between baseline and PRF runs for the same language. For further comparison we ran a similar set of experiments to optimize $k1$ and b for the CLEF 2004 ImageCLEF collection. We observed similar variations in optimal values between the topic languages, baseline and PRF runs, and also generally between the 2004 and 2005 topic sets for the same language and run condition. This variation between topic sets would appear to justify our original decision to adopt the same $k1$ and b values for all our submitted test runs.

3.3 Text and Image Combined Pseudo Relevance Feedback

The combination of features for content-based image retrieval was also optimized using the CLEF 2004 ImageCLEF task using only the topic and document images. Based on this optimization the matching scores of the features were combined as follows: $0.5 \times colour + 0.3 \times edge + 0.2 \times texture$.

The selection depth of documents in the ranked retrieved text-based and image-based lists from which the additional assumed relevant set could be selected was also determined using the CLEF 2004 ImageCLEF data. We carried out extensive investigation of the optimal search depth for a range of topic languages. There was no apparent reliable trend across the language pairs, and we

could not be confident that values chosen for a particular pair on the training data would be suitable for a new topic set. Based on analysis of overall trends across the set of language pairs, we decided to set the search to a depth of 180 retrieved documents for the text-only list and for the image-only list to a rank of 20 documents. Documents occurring in both lists down to these rank positions were assumed to be relevant and added to the text-only run top 15 documents assumed to be relevant for term selection in text-only PRF.

Results from these experiments are shown in Table 3. Comparing these results to those using the standard PRF method in Table 2 we observe very little change in the results. In general the results for our new method are marginally reduced in comparison to the standard method. Examination of the outputs from the component systems revealed that the main reason for the similarity between results in Tables 2 and 3 is that very few additional assumed relevant documents are found in the comparison of the text-only and image-only retrieval lists. This arises largely due to the failure of the image-only retrieval system to retrieve relevant documents within the upper ranks[1] of the retrieved lists. Thus when comparing the text-only and image-only retrieved lists very few matches were found. The poor performance of the image-only retrieval system is to be expected since we are using standard low-level image matching techniques on the St Andrew's collection which is very heterogeneous, but we had hoped that combining with the text-only evidence would prove useful.

Similar to the text-only runs, it is likely that these results could be improved marginally by adjusting the search depth of the lists for the PRF stage. However post fitting to the test data does not represent a realistic search scenario, is unlikely to give any clear increase in results, and, as shown in the previous section, will generally not be reliable for different topic sets and languages.

4 Conclusions and Further Work

Results from our experiments for the CLEF 2005 St Andrew's ImageCLEF task show expected performance trends for our baseline system and a PRF augmented text-based retrieval system each using the standard Okapi model. Our proposed new PRF approach combining retrieval lists from text-based and image-based retrieval for this task failed to improve on results obtained using a standard PRF method. A clear reason for the failure of this technique is the absence of relevant documents in the ranked lists retrieved by the image-only retrieval system. Despite the current results, it would be interesting to explore this technique further in a task where the image collection is more homogeneous and image-based retrieval is more effective.

[1] In determining the system parameters we explored searching the image retrieval lists to a depth of 200 documents for our new combination method.

References

[1] Clough, P., Müeller, H., Deselaers, T., Grubinger, M., Lehmann, T., Jensen, J., and Hersh, W.: The CLEF 2005 Cross-Language Image Retrieval Task, Proceedings of the CLEF 2005: Workshop on Cross-Language Information Retrieval and Evaluation, Vienna, Austria, 2005.

[2] Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., and Gatford, M.,: Okapi at TREC-3. In D.K. Harman, editor, Proceedings of the Third Text REtrieval Conference (TREC-3), pages 109-126. NIST, 1995.

[3] *Snowball* toolkit `http://snowball.tartarus.org/`

[4] Porter, M. F.: An algorithm for suffix stripping. *Program* 14:10-137, 1980.

[5] Jones, G. J. F., Groves, D., Khasin, A., Lam-Adesina, A. M., Mellebeek, B., and Way, A.: Dublin City University at CLEF 2004: Experiments with the ImageCLEF St Andrew's Collection, Proceedings of the CLEF 2004: Workshop on Cross-Language Information Retrieval and Evaluation, Bath, U.K., pp653-663, 2004.