Examining the Effectiveness of IR Techniques for Document Image Retrieval

Gareth J. F. Jones   Adenike M. Lam-Adesina
Department of Computer Science,
University of Exeter, Exeter EX4 4QF, U.K.

email: G.J.F.Jones@exeter.ac.uk, A.M.Lam-Adesina@exeter.ac.uk

Abstract

Effective Document Image Retrieval (DIR) requires the use of appropriate Information Retrieval (IR) methods. Our research indicates that some IR techniques found to be effective for electronic text retrieval do not transfer to DIR in a simple predictable manner, and that modifications must be made to these methods to enable them to work effectively for DIR.

1 Introduction

Information derived from the indexing of scanned document images is an essential component of applications such as the retrieval of printed legacy materials, and can provide additional sources of information for retrieval and management of digital multimedia media content, such as video sources. In order to make use of this information a natural starting point is to exploit techniques from the large body of research in information retrieval and other related natural language technology areas, such as information filtering and information extraction, derived from experience with electronic text. An underlying assumption of doing this is that these techniques will be as effective for image retrieval tasks as for the media for which they were originally developed. To date there has been only limited experimental investigation of the behaviour and effectiveness of IR techniques for document images and associated index information.

This very limited amount of experimental work examining the behaviour of standard IR methods on document image retrieval (DIR) tasks suggests that before, or certainly as well as, considering more advanced tasks such as semantic interpretation, we should spend some time examining fundamental retrieval behaviour. It may be tempting to compare image retrieval with other multimedia tasks such as spoken document retrieval (SDR). Examining the results of the SDR tasks in TREC-6 to TREC-9, it can be seen that techniques such as term weighting, query expansion, document expansion, are all highly effective for SDR [1], sometimes more so than for electronic text retrieval. However, any assumption that this trend will carry over into image retrieval needs to be handled with care.

In this paper we give a brief review of major relevant existing studies and describe our current research which shows unexpected results for the application of standard IR techniques in DIR, and suggests that more work is needed to better understand this behaviour.

2 Existing Work

To date there has only been a very limited amount of research in DIR reported in the IR literature. Much of this work is fairly inconclusive, and investigation incorporating careful analysis of results is even more limited.

The only generally available IR task for DIR is the TREC-5 Confusion task. This consists of a collection of around 55,000 documents in parallel electronic text and two indexed document image collections with 5% and 20% error rates. This task is a *known-item* search requiring only a single relevant document to be correctly retrieved for each search topic. As such this task does not examine the recall effects of the retrieval techniques, or indeed the general precision behaviour of documents in the collection which might be deemed relevant apart from the single relevant known-item. Participants in the TREC-5 track applied a variety of indexing, term weighting and feedback methods, but the overall outcome of the task was very inconclusive [2]. The main and fairly unsurprising result being that retrieval performance is affected adversely by increasing indexing error rates.

A number of interesting results were found in the extensive work on DIR carried out at the University of Las Vegas at Nevada [3]. Of particular note was their conclusion about the importance of using an appropriate within document frequency model for errorful document image index data, and the limited success they achieved using relevance feedback for document images compared to a parallel text retrieval task.

3 Information Retrieval for Mixed-Media Collections

For the last two years we have been working on a project called *Information Retrieval for Mixed-Media Collections (IRMMC)*. This project is concerned with retrieval from collections containing a mixture of documents from different sources. In our case we have been using a collection containing electronic text, spoken documents and scanned document images. An important component of the experimental work for this project was to examine the retrieval characteristics of documents in the three different sources. In order to do this before working with a mixed-media collection, we carried out an analysis with separate parallel document collections in each media to investigate differing retrieval behaviour.

Our experimental work is based on the TREC-8 SDR collection. The standard task contains an electronic text baseline document collection with near accurate manual transcription of the spoken documents and the output from an automatic speech recognition system. The document collection is taken from the English language broadcast news sections of the TDT-2 data set. The version used in our experiments is Version 3 (November 1999). The English broadcast news data is taken from 4 sources: CNN “Headline News”, ABC “World News Tonight”, PRI “The World” and VOA English news programmes. The broadcasts are taken from the period February to June 1998. The TREC-8 SDR document set comprises 21,754 individual news stories taken from this document set with an average length of 180. A set of 50 search topics of average length 13.7 words were formed by the organizers at NIST. Relevance assessment used a pooling method with an average of 36.4 relevant documents being identified for each topic. In order to compare retrieval for scanned document images we generated a further collection by printing each document in the collection in the format of a newspaper clipping, scanning the output and then performing OCR to generate an index file. Full details of the design of this scanned image collection are given in [4].

Our first experiments with this new collection examined the effectiveness of the Okapi term weighting function for DIR. The term frequency function used in Okapi was shown to be very effective for DIR [5], overcoming the shortcomings previously identified in [3].

We have subsequently carried out a comparative investigation into retrieval of electronic text, spoken documents and document images. In this investigation we explored the behaviour of term weighting and pseudo relevance feedback (PRF) for each collection. These experiments used our summary-based PRF method described in [7] based on a modification of the Robertson expansion term selection methods [6]. Experiments were carried out to optimise the number of assumed relevant documents and the number of expansion terms to be added. Results from this investigation showed that as we might expect, PRF gave improvement for both text retrieval and spoken document retrieval. Adding either the 5 or 20 top ranked expansion terms improved average precision in both cases relative to a no feedback baseline. However, for DIR while adding 5 terms gave an improvement in average precision, adding 20 terms actually reduces performance compared to the baseline with respect to both precision and the number of relevant documents retrieved [8].

Past research work in DIR concluded that retrieval performance is not affected adversely by errors in OCR generated documents [3], [9] although the effect of PRF on OCR text is still unclear. The work reported here is aimed at investigating the instability and unpredictability of retrieval performance and relevance feedback in DIR. In order to understand the behaviour of PRF for DIR better we have performed a further series of experiments comparing retrieval for the IRMMC scanned image and electronic text collections.

4 Experimental Investigation

This section presents results from our current investigation. The experiments used the research distribution of the City University Okapi system [6] used in our previous experiments [5] and our own summary-based PRF methods [6]. The Okapi parameters were set as follows K1 = 1.4 and b = 0.6. For the PRF runs we assume that the top 5 ranked documents are relevant as the source of potential expansion terms and that the top 20 are relevant for the ranking of potential expansion terms. Feedback summaries consisted of the best scoring 6 sentences as calculated using the summarization methods described in [6].

The results tables show precision at cut off of 5, 10 and 30 documents retrieved, standard TREC average precision and the total number of relevant documents retrieved in this run. The total number of available relevant documents in the TREC-8 SDR collection being 1808. The Figures show query-by-query breakdown comparisons of average precision in each case.

4.1 Baseline Results

Baseline	Text	DIR
P. 5	0.633	0.649
P. 10	0.551	0.557
P. 30	0.354	0.352
Av. Prec.	0.468	0.454
Rel. Ret	1608	1581

Table 1: baseline retrieval results for Te and OCR collection prior to feedback.

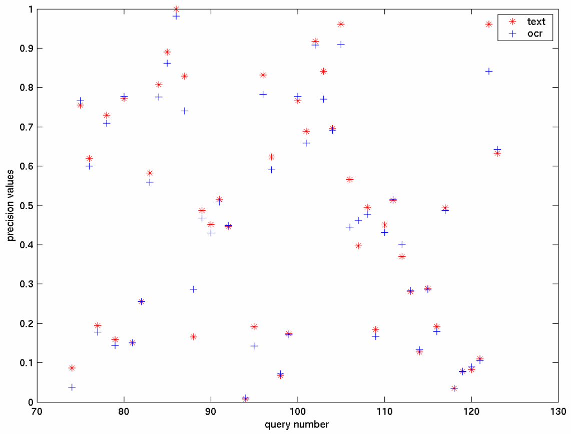

Figure 1: query-by-query comparison of average precision for baseline retrieval results.

Table 1 shows baseline results for text and document image retrieval prior to the application of PRF. The table above shows that initial retrieval performance for the collections is comparable with respect to all the measures used. Retrieval results seem not to be affected by errors in the OCR text. This result supports earlier conclusions reported in [3] [9].

Figure 1 shows a query-by-query breakdown comparison of average precision results for baseline retrieval. It can be seen that in general results for the two documents sets are again very similar, in most cases the results are almost indistinguishable or the text collection result is slightly higher; an overall result consistent with the average precision results shown in Table 1.

4.2 Effect of Expansion Term Selection

The next experiment establishes "baseline" results for application of summary-based PRF with differing numbers of query expansion terms. The original terms were upweighted by a factor of 1.5 relative to the expansion terms. In general it is observed that the addition of a selected number of terms improves retrieval until, number of terms selected reaches a peak value is reached after which there is a gradual loss in performance.

Feedback	5 terms		20 terms	
	Text	DIR	Text	DIR
P. 5	0.669	0.661	0.670	0.612
P. 10	0.580	0.574	0.598	0.539
P. 30	0.392	0.352	0.396	0.352
Av. Prec.	0.506	0.498	0.514	0.440
% chg. from baseline	+8.1%	+9.7%	+9.8%	-3.1%
Rel. Ret.	1639	1578	1631	1385
chg. from baseline	+31	-3	+23	-196

Table 2: retrieval with PRF for Text and OCR collections.

Table 2 retrieval shows retrieval performance for each collection after the application of PRF with 5 and 20 expansion terms. For text retrieval there is an improvement in both average precision and the number of relevant documents retrieved when either 5 or 20 terms are added. For DIR although retrieval performance is improved with the addition of 5 terms, performance is actually worse than the baseline when 20 terms are added. There is a loss of 3% in average precision and almost 200 in the overall number of relevant documents retrieved.

Figures 2 and 3 show the query-by-query breakdown of average precision results. Figure 2 shows a much greater variability in results between the two media than observed previously in Figure 1. In Figure 2 the results are quite different for most queries, in some cases OCR text retrieval is much better than for the original text, overall the results for electronic text are superior on more occasions producing the overall averages shown in Table 2. For this small number of expansion terms presumably on some occasions some better expansion terms are selected from the OCR baseline run than the electronic text run, leading to better results for some queries in the PRF run.

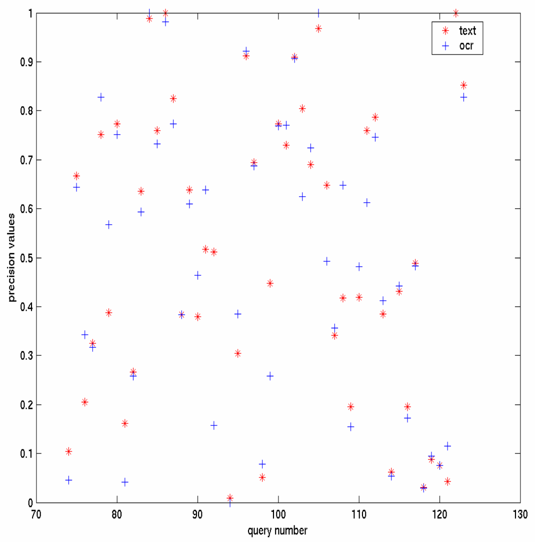

Figure 2: query-by-query comparison of average precision for retrieval with PRF with 5 expansion terms.

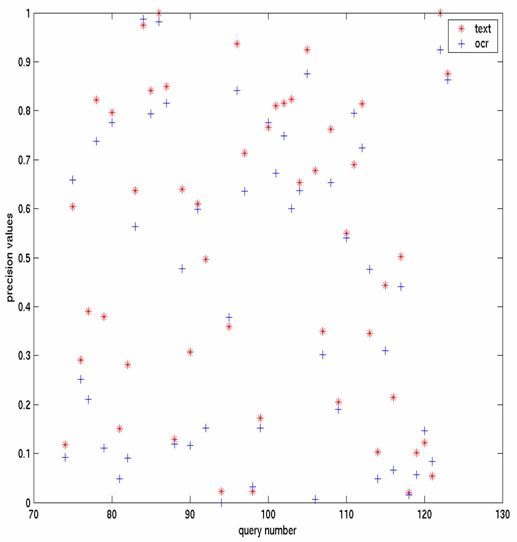

Figure 3: query-by-query comparison of average precision for retrieval with PRF with 20 expansion terms.

Figure 3 shows similar large differences in average precision retrieval performance between the two media, except in this case the DIR result is lower than the electronic text result by a large degree in most cases, again leading to the overall average precision results shown for 20 expansion terms in Table 2.

To better understand this behaviour we sought to investigate the reasons for the observed results. We hypothesized that the poor results for 20 expansion terms with the OCR text collection is attributable to either the selection of poor expansion terms or assignment of poor weights to some of the terms appearing in the expanded query. The following sections describe the results of these of this further experimental investigation.

4.3 Effect of Expansion Term Selection

Results for the baseline runs shown in Table 1 show that there is little different in average retrieval performance prior to the application of feedback. Results in Table 2 show that there is a large difference in behaviour between electronic and OCR text following the application of PRF. In this next experiment we investigated the effect of the expansion term selection in retrieval behaviour. To do this we exchanged the PRF expansion terms selected using the electronic text baseline run with those generated for the DIR baseline run. Results from this experiment should tell us if the difference in PRF retrieval behaviour can be attributed to the selected expansion terms.

Table 3 shows the results of this query swapping experiment. The results indicate that in general the terms selected from the different baseline runs were not a significant factor in the difference in PRF for electronic text and OCR text. Perhaps surprisingly electronic text retrieval performance actually improves very slightly when using the OCR text derived expanded queries. It is interesting to note from Table 1 that retrieval at rank cutoff of 5 DIR outperforms electronic text retrieval suggesting that it will provide a pool of potential expansion terms at least as good as

Swapped Queries	Text	DIR
P. 5	0.706	0.580
P. 10	0.608	0.516
P. 30	0.396	0.350
Av. Prec.	0.518	0.420
% chg. media	+0.8%	-4.5%
Rel. Ret.	1630	1364
chg. media	-1	-21

Table 3: retrieval performance for PRF with 20 expansion terms for Text and DIR with expanded queries swapped between the collections.

those for electronic text. The result for electronic text in Table 3 shows that good expansion terms are indeed selected from the DIR baseline run. The result for DIR with the electronic text expanded queries is more slightly surprising showing a further 4.5% reduction in average precision over the DIR with its own topic statements. Although the exact reason for this is not clear, the slightly lower high rank baseline retrieval performance for the Text collection may lead to selection of less good expansion terms, or perhaps some aspect of the distribution of correctly recognized terms in the OCR collection may not be well matched with the expansion terms required for most effective PRF for the Text Collection.

Overall these results indicate that we must look elsewhere to explain the behaviour of DIR with the application of PRF.

4.4 Effect of Term Weighting

Based on the results above, it can be concluded that poor term weights assigned to some of the expansion terms are responsible for the degradation in retrieval performance for OCR text. To confirm this hypothesis we performed another experiment. In this next experiment the *collection frequency weights (cfw)* were exchanged between the collections. The experiment in Section 4.2 was then repeated using 20 expansion terms in the PRF run.

Results for this experiment are shown in Table 4 and Figure 4. These results are very interesting, there is a large improvement in average precision and the number of relevant documents retrieved for the OCR collection relative to the baseline result in Table 1 when the correct text weights are used. By contrast while the electronic text retrieval average precision result is better than the baseline, it is some 2.5% worse than that with its own *cfw*'s. While the exchange of weights has produced a small reduction in average precision here, this reduction is much less than that observed for the OCR documents with these *cfw* values. This suggests that the relationship between term weighting and retrieval performance is more complex than merely arising from the poor term weights. In general query-document matching will be more reliable for electronic text and it is probable that this compensates to some extent for the reduced quality of term weight estimates. Also notable is the fact that there is a loss of more than 100 in the overall number of relevant documents retrieved in this latter case arising only from the exchange of the term weights. These are likely to be documents at the lower end of the ranked retrieved lists where the number of query-document term matches are likely to be low. This further suggests that query-document matching is an important issue in the reliability of retrieval where the quality of term weights is reduced.

Swapped *cfw* weights	Text	DIR
P. 5	0.686	0.702
P. 10	0.592	0.606
P. 30	0.390	0.395
Av. Prec.	0.503	0.515
% chg. from baseline	+7.5%	+13.4%
Rel. Ret.	1501	1640
chg. from baseline	-107	+59

Table 4: Results of swapping the *collection frequency weights (cfw)* for electronic and OCR document collections.

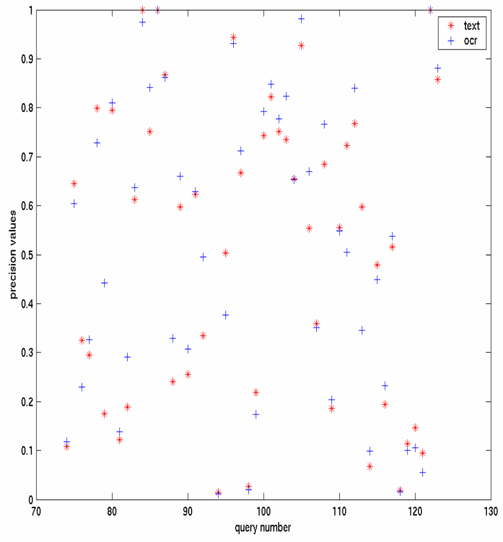

Figure 4: query-by-query comparison of average precision for retrieval with PRF with 20 expansion terms with collection frequency weights swapped between collections.

Improved term weights clearly form part of the solution to better DIR performance. In practice of course there is unlikely to be a parallel text collection available for a DIR task (if there is we might as well use the text collection instead for the retrieval phase since it will be more reliable). However, there will often be access to contemporaneous or related electronic text documents. In the next experiments we seek to obtain the benefits of improved term weight estimates for DIR from an alternative document collection.

4.5 Term Weight Correction

In this section we report results for experiments using related text information to explore the extent to which the benefits of reliable indexing can be derived from non-parallel sources. The text data used here is the also taken from the TDT-2 New Corpus. In addition to the 4 broadcast sources outlined in Section 3, this collection also includes 2 text document sources taken from the same time period as the broadcast news material. These are taken from New York Times Newswire Service (excluding non-NYT sources) and Associated Press Worldstream Service (English content only), and include a total of around 20,000 news stories.

Two sets of experiments were carried out using cfw's calculated using the contemporaneous text data. In the first experiment the *cfw*'s calculated using the contemporaneous text data were used on their own to replace the weights in the test document collections. In the second set of experiments the contemporaneous text collection and the test document collection were combined to calculate the *cfw* values. This latter strategy has been successfully adopted previously by partcipants in the TREC SDR tracks [1], but not to our knowledge explored for DIR.

4.5.1 Contemporaneous text *cfw* values

Contemp. Collection	Text	DIR
P.5	0.633	0.625
P. 10	0.547	0.549
P. 30	0.370	0.361
Av. Prec.	0.494	0.479
% chg. from baseline	+5.6%	+5.5%
Rel. Ret.	1632	1618
Chg. from baseline	+24	+37

Table 5: shows the effect of applying term weights estimated from a comparable collection on the test collections

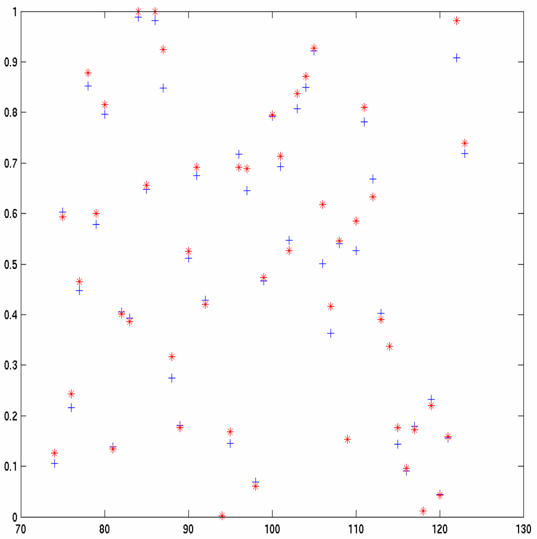

Figure 5: showing the query-by-query effect of using a comparable collection for term weights estimation.

Table 5 and Figure 5 show the results for replacement of *cfw* values in the test collections with those generated from the contemporaneous text collection. These results show estimation of term weights using this source to be beneficial for DIR although the result achieved for Text retrieval is lower than that observed using its own weights.

One of the characteristics of printed text is that it sometimes contains more news (background of the news item and probably a small narration of how the event have been evolving) than its broadcast counterpart. In this case the word distribution between the collections is likely to be different. This might account for the low average precision, shown above for retrieval using cfw weights from the contemporaneous text collection on the Text test collection. While this effect would also apply to the OCR test collection, the overall accuracy of the cfw estimates derived from the contemporaneous text collection more than compensates for the difference in term distributions between the collections. Thus we explored the idea of merging the comparable collection with the original collection to correct this problem.

4.5.2 Combined collections *cfw* values

Table 6 shows the results for retrieval using *cfw* values calculated by combining the contemporaneous text collection with the test collection in each case. These results show that using term weight estimation from the merged collection resulted in improved retrieval performance for both collections. For DIR approximately 18% retrieval performance improvement over the baseline is achieved by using term weights estimated from the merged collection which is about 12% more than that achieved when only the comparable collection was used for term weights estimation in the previous experiment. This result is particularly encouraging for the development of effective DIR tools. The result for Text retrieval is also very

Merged Collection	Text	DIR
P. 5	0.718	0.722
P. 10	0.622	0.614
P. 30	0.414	0.409
Av. Prec.	0.541	0.534
% chg. from baseline	+15.6%	+17.6%
Rel. Ret.	1656	1656
chg. from baseline	+48	+75

Table 6: shows the effect of term weights estimation from the merged collection

good indicating the usefulness of using additional material for collection-based term weighting when the test collection itself is relatively small.

5 Conclusion

Our investigation shows that although baseline retrieval performance is not adversely affected by errors in OCR text, degradation in retrieval performance occurs when PRF is applied. Experiments showed that the methods used for selection of expansion terms for PRF are robust to OCR text recognition errors, and that the problem with PRF for DIR is caused by poor estimation of weights for at least some of the terms added to the initial query. If a comparable collection for such collection exists, this can be used to correctly estimate term weights before it is applied on the OCR text. Further analysis of these results is needed to examine the effects of inidivdual term weight errors and their origins within the OCR collection.

Further investigations are also needed to test the effectiveness of contemporaneous collections with a larger test collection before definite conclusions can be made. We are also keen to determine the effects of varying degree of errors on these methods, the relationship between the contents of the test collection and the comparable document set, and what can be done to correct term weights in the absence of a suitable comparable collection.

Acknowledgement

The IRMMC project was supported by UK EPSRC grant GR/N04034.

References

[1] J.S. Garafolo, C.G.P. Auzanne, and E.M. Voorhees. The TREC Spoken Document Retrieval Track: A Success Story. In *Proceedings of the RIAO 2000 Conference: Content-Based Multimedia Information Access*, pages 1-20, Paris, 2000.

[2] P. B. Kantor and E. M. Voorhees. The TREC-5 Confusion Track: Comparing Retrieval Methods for Scanned Text. *Information Retrieval*, 2:165-176, 2000.

[3] K. Taghva, J. Borsack, and A. Condit. Results of applying probabilistic IR to OCR text. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 202-211, Dublin, 1994. ACM.

[4] G. J. F. Jones and M.Han. Information Retrieval from Mixed-Media Collections: Report on Design and Indexing of a Scanned Document Collection. Technical Report 400, Department of Computer Science, University of Exeter, January 2001.

[5] G. J. F. Jones and M. Han. Retrieving scanned documents from a mixed-media document collection. In *Proceedings of the BCS-IRSG European Colloquium on IR Research*, pages 136-149, Darmstadt, April 2001.

[6] A. M. Lam-Adesina and G. J. F. Jones. Applying Summarisation Techniques for Term Selection in Relevance Feedback. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1-9, New Orleans, 2001. ACM.

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

[8] G.J.F.Jones and A.M.Lam-Adesina. An Investigation of Mixed-Media Information Retrieval. In *Proceedings of the 6th European Conference on Digital Libraries*, pages 463-478, Rome, 2002. Springer Verlag.

[9] K. Taghva, J. Borsack and A. Condit. Evaluation of model-based retrieval effectiveness with OCR text, *ACM Transactions on Information Systems (TOIS)*, 14:1, 1996