Examining and Improving the Effectiveness of Relevance Feedback for Retrieval of Scanned Text Documents

Adenike M. Lam-Adesina Gareth J. F. Jones [*]

School of Computing & Centre for Digital Video Processing, Dublin City University, Glasnevin, Dublin 9, Ireland

Abstract

Important legacy paper documents are digitized and collected in online accessible archives. This enables the preservation, sharing, and significantly the searching of these documents. The text contents of these document images can be transcribed automatically using OCR systems and then stored in an information retrieval system. However, OCR systems make errors in character recognition which have previously been shown to impact on document retrieval behaviour. In particular relevance feedback query-expansion methods, which are often effective for improving electronic text retrieval, are observed to be less reliable for retrieval of scanned document images. Our experimental examination of the effects of character recognition errors on an ad hoc OCR retrieval task demonstrates that, while baseline information retrieval can remain relatively unaffected by transcription errors, relevance feedback via query expansion becomes highly unstable. This paper examines the reason for this behaviour, and introduces novel modifications to standard relevance feedback methods. These methods are shown experimentally to improve the effectiveness of relevance feedback for errorful OCR transcriptions. The new methods combine similar recognised character strings based on term collection frequency and a string edit-distance measure. The techniques are domain independent and make no use of external resources such as dictionaries or training data.

Key words: document image retrieval, scanned text, image retrieval, relevance feedback, query expansion

[*] Corresponding author.
Email addresses: adenike@computing.dcu.ie (Adenike M. Lam-Adesina), gjones@computing.dcu.ie (Gareth J. F. Jones).

Preprint submitted to Information Processing and Management 17 June 2005

1 Introduction

The amount of information available from new and existing documents continues to expand rapidly. While many new documents are in the form of electronic text which is easily searchable, a substantial proportion of this existing material is in paper form (Lynam et al., 2003). In addition to new documents, librarians and other archive maintainers around the world are creating digitized copies of existing paper document collections. These archives have often been gathered over hundreds of years and frequently contain unique material which is becoming increasingly fragile with the passing of time. These materials have not previously been available for easy consultation. Digitized archives of newly generated and legacy paper documents can form an important information source for many users. For example, the Bodleian library at the University of Oxford is currently engaged in an ambitious project to digitize thousands of manuscripts and printed works from its collection of over 8 million items (Bodleian, 2004; ODL, 2004), while on a smaller scale the *Irish Script on Screen (ISOS)* project in Dublin is preserving and making widely available online the cultural heritage of the Irish language (ISOS, 2004). Readers wishing to consult these collections had previously been required to travel to the library holding the volume they wished to consult, and even then, such is the value of the original documents, access was often restricted to professional scholars. Digital copies can be consulted remotely by any interested party without any possibility of damage to the original collection.

Digitizing these documents enables them not only to be preserved and accessed remotely, but also importantly potentially makes them searchable by users of document retrieval systems. In order to maximize the exploitation of these resources, the documents can be indexed for full-text searching to find potentially interesting material. Searching scanned documents in this way requires that their contents first be recognized by an Optical Character Recognition (OCR) process prior to entry into a retrieval system. Unfortunately OCR systems make errors in recognition, which then impact on the behaviour of retrieval systems. While retrieval from error-free text collections has been, and continues to be, the subject of much research work (Spärck Jones & Willett, 1997); detailed performance analysis and development of techniques specifically to address retrieval issues for documents indexed using OCR have received comparatively little attention. The increasing potential for exploiting these digitized collections for scholarship and cultural understanding, as the size and diversity of these collections grows, means that it is important to properly understand the retrieval issues associated with OCR indexed data and develop methods to overcome identified weaknesses.

The effectiveness of electronic text retrieval is improved by techniques such as term weighting, and also by methods such as *relevance feedback (RF)*. Term weighting aims to increase the significance of terms with high selectivity which are able to distinguish relevant from non-relevant documents with respect to a search request. In RF an initial search is performed for an information search request, in response to this a number of potentially useful or relevant documents are returned. These documents are then used to modify the initial search request to make it a better expression of the user's need for information. The modified request is then applied to the information retrieval (IR) system for a further, hopefully improved, retrieval run. In previous work we have demonstrated that term weighting methods are effective for baseline document image retrieval (DIR) (Jones & Lam-Adesina, 2002). However, while RF has been shown to be very effective for retrieval of the errorful automated transcripts of spoken documents (Johnson et al., 1999), its behaviour for DIR has been much less successful, either failing to improve retrieval performance significantly (Taghva, Borsack & Condit, 1996a) or actually reducing it (Jones & Lam-Adesina, 2002; Taghva et al., 2004). This paper provides a careful analysis of RF for DIR using a comparison of native electronic text retrieval and DIR for an ad hoc retrieval task. The study first establishes the causes of the unreliability of RF for DIR, and then secondly uses this analysis to propose extensions to the term selection process. Two domain independent modified term selection methods are introduced and evaluated. The first simple method ignores potential expansion terms with low collection frequency. The second more sophisticated method seeks to combine similar character strings from the top ranked documents which are within an edit distance criterion. The motivation for this technique is described later in the paper. Both of these methods are shown to make RF effective for a DIR task.

The remainder of the paper is organised as follows. Section 2 gives a more detailed introduction to RF and reviews existing related work, Section 3 describes our experimental task, Section 4 summarises the IR techniques used in our work, Section 5 gives our experimental results including a careful analysis of observed behaviour, and introduces and evaluates our extended methods for term selection in DIR, and finally Section 6 concludes with directions for further work.

2 Relevance Feedback and Document Image Retrieval

In RF information collated from relevant (and sometimes non-relevant) documents retrieved using an initial search is used to improve IR. RF is typically implemented via two processes: modification of the existing search query to add or remove terms from the query (query modification) and/or reweighting of the search query terms (term reweighting). Two approaches can be taken to gathering data for RF: true RF where users are asked to indicate the relevance of individual documents retrieved in the initial search, or pseudo RF (PRF) where high ranked documents are assumed to be relevant. The PRF approach generally gives a smaller average improvement in performance than true RF, since information from non-relevant documents is often included in the modifications for the feedback stage, but it is completely automatic requiring no input from the user.

The principal difference between retrieval of typed electronic text and retrieval of OCR indexed scanned documents in DIR is the recognition errors and consequent indexing errors introduced by the OCR process. OCR techniques tend to misrecognise individual characters within words. This introduces novel letter string units into the indexing vocabulary. This can lead to matching problems between indexing units in the search queries and the documents, and also to inappropriate estimation of parameters in the IR system. A number of existing studies have explored the impact of these indexing errors on retrieval behaviour for a range of retrieval models. These have generally concluded that while effective retrieval can be achieved, baseline retrieval is somewhat degraded relative to accurate transcriptions, but also that the performance of RF is less reliable for DIR. The remainder of this section reviews key findings of this existing work in DIR and in particular RF applied to DIR.

The only existing comparative study of DIR for alternative retrieval systems was conducted within the TREC-5 Confusion Track (Kantor & Voorhees, 2000), other significant work has been carried out at ETH Zurich (Mittendorf & Schäuble, 2000), and of most interest with respect to our work, a number of studies have been reported by the University of Nevada, Las Vegas (Taghva, Borsack & Condit, 1996a,b; Taghva et al., 2004). The findings of these and other studies are summarised below.

2.1 TREC-5 Confusion Track

The TREC-5 Confusion track consisted of a "known-item" search, a task in which the system attempts to find a single, partially-remembered, target document from within a document collection. The document collection contained approximately 55,000 documents from the *Federal Register*. Participants in the track were provided with three text versions of the data: the original electronic typed text which was regarded as a baseline, a second version with an estimated error rate of 5% obtained by scanning the hardcopy, and a third version obtained by downsampling the original page images and having an error rate of around 20%.

Groups participating in this task adopted a variety of indexing and best-match IR strategies. The indexing methods used can generally be divided into the following: n-gram character string matching, fuzzy matching of word and character strings, and automated “correction” of words using a dictionary. In addition, several participants explored the use of automated feedback strategies to perform a second retrieval pass.

The CLARITECH submission explored both correction of OCR output and feedback (Tong et al., 1997). Their query expansion method involved expanding the original query by using variants of each query term (search item) taken from the document corpus. These expansion terms were judged similar to the query terms by computing an edit distance based on the minimum number of character changes needed to get to the original term. A maximum of 10 variants within an edit distance of 3 were added for each query term. All original query terms were upweighted relative to the selected expansion terms. Expansion improved the average rank of retrieved relevant known-items, but the key task measure of mean reciprocal rank (MRR) was reduced.

The George Mason University team explored the use of a $tf \times idf$ based PRF method for this task with subword n-gram based indexing (Grossman et al., 1997). This produced a 40% reduction in MRR for the 5% degraded text relative to a no feedback run, but actually gave a 50% reduction for baseline accurate text for this task, although it worked well for a different standard ad hoc IR task.

Overall results of the participants’ submissions were generally inconclusive; the track co-ordinators were unsure as to whether retrieval effectiveness was affected more by the indexing method or the retrieval strategy adopted (Kantor & Voorhees, 2000). However, overall as would be expected, increasing error rates in indexing reduced retrieval effectiveness.

In this known-item search only a single document is identified as relevant to the search request. The scope for investigation of RF in this task is limited since it is not possible to explore issues of improved recall with RF and only to observe variation in precision to a very limited degree. If this study were extended to a standard ad hoc retrieval task with full assessment of potentially relevant documents, then results of these experiments would give a better understanding of behaviour with respect to precision and recall.

2.2 ETH Zurich

Mittendorf and Schäuble working at ETH Zurich carried out a careful theoretical and experimental analysis of the impact of recognition errors on DIR behaviour (Mittendorf & Schäuble, 2000). Working with the TREC-5 Confusion Track collection, they concluded that attempts to post-process the output of an OCR system to correct errors using a dictionary-based method will be unstable with respect to IR effectiveness if the dictionary does not have complete coverage of the document collection vocabulary. For free-text indexing complete dictionary coverage of important search words, such as previously unseen proper nouns, will often not be possible. While attempts to correct the transcript may be successful for words in the dictionary, accurately recognized novel words may often actually be corrupted in the "correction" process, reducing retrieval effectiveness for queries containing these words. They also suggest that adopting word-based indexing is best for lightly corrupted data, such as that used in the investigation described in this paper.

While not exploring RF directly, this study emphasizes the importance of selecting expansion terms that have been recognized reliably by the OCR system. Our investigation follows these recommendations by using word-based indexing and using the raw output of the OCR system without any attempt at dictionary-based correction of recognition errors. Our experimental investigation also confirms the importance of basing expansion term selection on accurately recognized terms.

2.3 University of Nevada, Las Vegas

Experiments at the University of Nevada, Las Vegas were conducted on a locally developed ad hoc IR task using the standard Ide dec-hi and the Rocchio RF methods (Taghva, Borsack & Condit, 1996b). After an initial run and manual relevance assessment of the top ranked documents for feedback, retrieval effectiveness on the residual, or remaining document collection below those assessed, improved for both the correct and OCR generated versions of the collection when the query was expanded and a feedback run carried out. However, while performance on the correct text continued to improve as more expansion terms were added to the query, improvement in RF effectiveness for the OCR collection levelled off at a lower degree of improvement. It was the authors' observation that in general good expansions terms were selected, but that OCR errors in the documents prevented them being retrieved at improved ranks.

A study based on n-gram indexing of degraded versions of the standard CACM, NPL, TIME and WSJ collections is reported in (Harding, Croft & Weir, 1997). These experiments use a query expansion technique similar to that adopted by CLARITECH for the TREC-5 Confusion Track. The queries were expanded to include 1, 2 or 3 of the nearest matching terms in the document vocabulary. This improved average precision retrieval effectiveness in all but one case. However, a further study reported in (Marukawa et al., 1997) again showed the ineffectiveness of query expansion for retrieval from corrupted text. In this research 1083 Japanese news articles were searched using 50 test queries. Query terms were expanded by replacing characters by similar ones to form new queries. This method again resulted in lower average precision for the expanded queries compared to the initial query set.

3 Data Collection

Our experimental investigation was carried out using a DIR research collection adapted from the TREC-8 spoken document retrieval (SDR) task (Johnson et al., 1999). Adapting this existing collection avoided the cost of developing a completely new ad hoc retrieval test collection, and enabled comparison between SDR and DIR behaviour (Jones & Lam-Adesina, 2002). The original SDR test collection consisted of the documents, search requests and relevant documents for each request. For our investigation we developed a parallel document image collection consisting of scanned images generated from manual transcriptions of the audio data. The TREC-8 SDR collection is based on the English broadcast news portion of the TDT-2 News Corpus. The existing SDR collection of text and spoken document sets was augmented by forming a corresponding scanned document collection. The scanned document collection was based on the 21,759 "NEWS" stories in TDT-2 Version 3 (December 1999).

3.1 TDT-2 Document Set

The TREC-8 SDR portion of the TDT-2 News Corpus covers a period of 5 months from February to June 1998. The news data is taken from 4 sources as follows: CNN "Headline News" (about 80 stories in 4 programmes per day), ABC "World News Tonight" (about 15 stories in 1 programme per day), PRI "The World" (about 20 stories in 1 programme per day) and VOA English news programmes (about 40 stories from 2 programmes per day). The sampling frequencies are approximate and all sources were prone to some failures in the data collection process.

Each broadcast is manually segmented into a number of news stories which form the basic document unit of the corpus. Each news story is uniquely identified by a "DOCNO" indicating the source, date and time of the broadcast, and the location of the story within the broadcast. An individual news story was defined as containing two or more declarative statements about a single event. Miscellaneous data including commercial breaks, music interludes, and trailers were excluded from the data set. The collection contains a total of 21,754 stories with an average length of 180 words totalling about 385 hours of audio data.

There is no high-quality human reference transcription available for TDT-2 - only "closed-caption" quality transcriptions for the television sources and rough transcripts quickly made for the radio sources by commercial transcription services. Additionally, transcriptions are available generated using automatic speech recognition. These have higher error rates than the "rough" manual transcriptions, and are not used in the investigation reported in this paper. The manual transcriptions are used here to assess baseline retrieval performance and as the document source for out scanned document collection.

3.2 Scanned Document Collection

The objective in the design of the printed version of the collection was to create story hardcopy similar in style to newspaper clippings. In order to simulate the differences in formatting of stories from different newspaper sources, each story was printed in one of four fonts: *Times*, *Pandora*, *Computer Modern* and *San serif*. The stories were divided roughly equally between these font types. Material from CNN, ABC, PRI and VOA was assigned to each font on a sequential basis evenly over the time span of the collection. The stories were printed in 1 of 6 different widths and in 1 of 3 different font sizes. The text column width and font size were assigned sequentially from the beginning of each broadcast. These were printed in single column format running onto a second page if necessary. The stories were printed using a Epson EPL-N4000 laser printer. Further details of the collection design are contained in Jones & Han (2001).

Since the stories had been newly printed onto white paper using a high quality laser printer, the best available current OCR technologies would have been able to provide almost perfect transcriptions. However, the operational target of scanned document retrieval will often be legacy documents printed using mechanical methods, for which the print quality is usually inferior to current printings. In addition, the contrast between print and paper may be reduced by discolouration of the paper. In order to explore retrieval behaviour with a more errorful transcription, an OCR transcription was performed with suboptimal system settings. After some ad hoc exploration of OCR accuracy versus the system parameters, the transcription was created as follows. All documents were scanned using an HPScanJet ADF at 200 dpi in Black & White at a threshold of 100. OCR was carried out using Page Keeper Standard Version 3.0 (OCR Engine Version 271) (SR3). Errors include typical OCR mistakes such as the recognition of `journal` as `joumal`. This scanning and recognition strategy obviously assumes that the type of errors created will be similar to those observed for legacy documents with high quality OCR, however this has not been experimentally verified. The OCR process introduced some errors into story name labels, these were all manually verified and errors corrected in order to ensure accuracy of story names. Story names are regarded as metadata which must be accurately recorded in order to manage the archive reliably.

Further versions of the scanned document set could be formed by one or more generations of photocopying the printed stories with varying settings on the photocopier. These different document sets could then be interleaved to generate collections of varying quality which would then experience differing levels of OCR performance across the collection. However, it was decided to have all documents with as uniform a level of degradation quality as possible, so that the retrieval behaviour of the data could be explored without the additional complexity of uneven image quality.

3.3 TREC-8 SDR Test Collection

The TREC-8 SDR retrieval test collection contains a set of 50 search topics and corresponding relevance assessments. The goal in creating the topics was to devise topics with a few (but not too many) relevant documents in the collection to appropriately challenge test retrieval systems. Retrieval runs submitted by the TREC-8 SDR participants were used to form document pools for manual relevance assessment. The average topic length was 13.7 words and the mean number of relevant documents for each topic was 36.4. Note: only 49 of the topics were ultimately adjudged to have relevant documents within the TREC-8 SDR corpus (Johnson et al., 1999).

The subsequent TREC-9 SDR evaluation used the same document collection with a new set of topics and relevance assessments. In this latter case two sets of topics were provided: a standard set with average length 11.7 words, and a set of “terse” query statements of average length 3.3 words which seek to express the same information need. These terse query statements were developed to explore the impact of recognition errors on very short query statements. Evaluations reported by Johnson et al. (2000) show there to be almost no difference in retrieval behaviour between standard and terse query statements for SDR. We were not able to undertake evaluation with the TREC-9 SDR test set due to a relabelling of the document set by the collection providers, but believe, based on our baseline results for TREC-8 SDR, that DIR would be similarly unaffected for terse query statements.

4 Information Retrieval Techniques

The basis of the experimental system was the City University research distribution version of the Okapi system (Robertson et al., 1995). The Okapi retrieval model has been shown to be very effective in a number of comparative evaluation exercises in recent years for Text Retrieval and SDR tasks (Johnson et al., 1999) and has been adopted in many IR research systems. The retrieval strategy adopted in this investigation follows standard practice for best-match ranked retrieval. The documents and search topics were first processed to remove common stop words from a list of around 260 words, suffix stripped using the Okapi implementation of Porter stemming (Porter, 1980) to encourage matching of different word forms, and terms were further indexed using a small set of synonyms.

4.1 Term Weighting

Following preprocessing document terms are weighted using the Okapi BM25 weight (Robertson et al., 1995). The BM25 weight for a term is calculated as follows,

$$cw(i,j) = cfw(i) \times \frac{tf(i,j) \times (k_1 + 1)}{k_1 \times ((1-b) + (b \times ndl(j))) + tf(i,j)}$$

where $cw(i,j)$ represents the weight of term i in document j, $cfw(i) = log(N/n(i))$ the standard collection frequency (inverse document frequency) weight, $n(i)$ is the total number of documents containing term i, and N is the total number of documents in the collection, $tf(i,j)$ is the within document term frequency, and $ndl(j) = dl(j)/\text{Av}.dl$ is the normalized document length where $dl(j)$ is the length of j. k_1 and b are empirically selected tuning constants for a particular collection. The matching score for each document is computed by summing the weights of terms appearing in the query and the document, which are then returned in order of decreasing matching score.

4.2 Relevance Feedback

The main issues for implementing RF are the selection of appropriate expansion terms and calculation of revised term weights. In the standard Okapi approach potential expansion terms are ranked using the Robertson selection value (rsv) (Robertson, 1990), defined as,

$$rsv(i) = r(i) \times rw(i) \qquad (1)$$

where $r(i)$ is again the number of relevant documents containing term i, and $rw(i)$ is the standard Robertson/Sparck Jones relevance weight (Robertson et al., 1995) defined as,

$$rw(i) = \log \frac{(r(i) + 0.5)(N - n(i) - R + r(i) + 0.5)}{(n(i) - r(i) + 0.5)(R - r(i) + 0.5)}$$

where $n(i)$ and N have the same definitions as before and R is the total number of relevant documents for this query. The top ranking terms are then added to the query. Term reweighting is carried out by replacing $cfw(i)$ with $rw(i)$ in the BM25 weight. In this study we explore only query expansion since we generally observe this to be the dominant factor in RF.

Standard RF and PRF methods treat the whole document as relevant, the implication of this being that using terms from non-relevant sections of these documents for expansion may cause query drift. Further problems can arise in PRF when terms are taken from assumed relevant documents that are actually non-relevant. To reduce the number of expansion terms taken from non-relevant material, we adopt a term selection method based on document summaries (Lam-Adesina & Jones, 2001). Our results using this method have been very encouraging for other tasks. including our previous investigation of Text Retrieval and SDR for the test collection used in this paper (Jones & Lam-Adesina, 2002), and we use it again in this investigation. By focusing on the key elements of the document our method seeks to exclude possible expansion terms not closely associated with the main focus of the document and query. The summary is formed by taking a fixed number of sentences from each document selected using a combination of statistical and heuristic techniques (Lam-Adesina & Jones, 2001). Potential expansion terms are selected from the top R_1 documents assumed relevant, but the $rsv(i)$ is calculated using a separate larger R value; we find that this technique gives more effective $rsv(i)$ values.

5 Experimental Investigation

This section describes our investigation of PRF for scanned document images. The experiments begin by establishing baseline retrieval performance, first without and then with the application of PRF, for text and scanned document images. PRF behaviour for DIR is then explored through a range of experiments using variations of term weighting and expansion term sets. The characteristics of the OCR collection are then examined, and the results of this analysis used to understand term selection for PRF query expansion.

Table 1
Baseline and summary-based feedback results for both collections Media Text OCR.

Media	Text				OCR			
	P10	P30	AvP	RelRet	P10	P30	AvP	RelRet
Baseline	0.551	0.354	0.468	1608	0.557	0.352	0.454	1581
Fbk 5	0.580	0.392	0.506	1639	0.574	0.380	0.498	1578
Chg bl.	+5.3%	+10.7%	+8.1%	+31	+3.1%	+7.9%	+9.7%	-3
Fbk 20	0.598	0.396	0.514	1631	0.539	0.352	0.440	1385
Chg bl.	+8.5%	+11.9%	+9.8%	+23	-4.1%	-0%	-3.1%	-196

Finally, these results are used to develop two extended methods for expansion term selection in PRF with DIR. Evaluation of these methods shows them to be effective for our DIR task.

Results are shown for retrieval precision at 10 and 30 document cutoff, standard TREC average precision and the total number of relevant documents retrieved. The number of relevant documents retrieved in each case can be compared for Recall to the total number of relevant documents across all topic statements in the TREC-8 SDR test set of 1818.

The BM25 values were set empirically as $K1 = 1.4$ and $b = 0.6$ using the baseline retrieval system without PRF. The parameters for the summary-based PRF were set as follows using the the electronic Text collection. Summaries were based on the most significant 6 sentences, the top 5 ranked documents are the source of potential feedback terms (R_1), and the top 20 documents assumed relevant for computation of $rsv(i)$ for term selection (R). The weight of the original query terms was in each case multiplied by 1.5 relative to the expansion terms, since the original terms have been chosen by the searcher themself. These values were used successfully for Text Retrieval and SDR in our earlier investigations of PRF for these collections (Jones & Lam-Adesina, 2002)

5.1 Baseline and Standard PRF Results

Table 1 shows baseline retrieval results without feedback (Baseline) and PRF results adding 5 (Fbk 5) and 20 (Fbk 20) expansion terms for both the uncorrupted Text and OCR document collections, together with their changes from the baseline. From Table 1 it can be seen that there is a small reduction in both baseline average precision and the number of relevant documents retrieved for OCR Text compared to the uncorrupted Text. The BM25 weighting scheme is thus shown to be effective for baseline retrieval with this OCR indexed collection without additional processing. Further investigation of these results for randomly selected queries shows that some terms that occur in the original query have been distorted in the OCR collection during the recognition process e.g. *Government* → *Govrcment*, *Financial* → *Linanci*, and that these term recognition errors were a major factor in the failure to retrieve 27 relevant documents from the OCR collection compared to the Text collection.

Looking now at the PRF results, it can be seen that performance in terms of both precision and number of relevant documents retrieved improves for both collections when 5 expansion terms are added. However, while performance continues to improve for the Text collection when 20 terms are added, the reverse is the case for the OCR collection. In this case the average precision decreases by -3.1% relative to the baseline and the number of relevant document retrieved by -196. This poor result is somewhat surprising and represents an average loss of 4 relevant documents per request, as well as reduced precision. Overall PRF results show it operating reliably for Text retrieval, and that while it can be effective for DIR, it is clearly less robust as the number of expansion terms chosen is increased. The next section describes a series of experiments and analysis of the data to better understand the causes of this behaviour.

5.2 Analysis of OCR text PRF Performance

One question that arises in analysing these results is; does DIR perform better with less expansion terms because: poor expansion terms are selected as the number increases, due to matching problems between the query terms and the errorful OCR text in the document, because of issues of term weighting associated with expansion terms, or a combination of these factors. In order to explore these effects a series of experiments were performed. PRF works well for uncorrupted Text for both 5 and 20 term expansion, hence we can make use of this collection in our analysis of PRF for the OCR test collection.

In order to explore the impact of the selected expansion terms on retrieval performance, the 20 expansion term sets were swapped. Thus uncorrupted Text expansion terms were added to the baseline queries for the OCR collection, with the reverse experiment performed for the Text document data with OCR expansion terms. Results for this first experiment are shown in Table 2 (SwQy1). The SwQy1 results show that there is little change in the Text collection retrieval results when using the expansion terms selected from the OCR collection, while there is a further small reduction in the precision and relevant documents retrieved for the OCR collection when the Text expansion terms are used. This would suggest that in general the selection of expansion terms for the OCR collection is robust to recognition errors, however analy-

Table 2
Retrieval results with summary based feedback swapping 20 expansion terms and term weights between the collections.

Media	Text				OCR			
	P10	P30	AvP	RelRet	P10	P30	AvP	RelRet
SwQy1	0.608	0.396	0.518	1630	0.516	0.350	0.420	1364
chg bl.	+10.3%	+11.9%	+10.7%	+22	-7.4%	-0.6%	-7.5%	-217
SwWgt	0.557	0.355	0.465	1598	0.553	0.346	0.450	1593
chg bl.	+1.1%	+0.3%	-0.6%	-10	-0.7%	-1.7%	-0.9%	+12
SwQy2	0.600	0.399	0.498	1504	0.584	0.378	0.493	1597
chg bl.	+8.9%	+12.7%	+6.4%	-104	+4.8%	+7.1%	+8.6%	+16
SwQyWgt	0.592	0.390	0.503	1501	0.606	0.395	0.515	1640
chg bl.	+7.4%	+10.2%	+7.9%	-107	+8.8%	+12.2%	+13.4%	+59

sis of the OCR expansion terms shows the presence of some corrupted words which will not match with any terms in the Text collection. The failure of the Text collection expansion terms to improve OCR collection retrieval effectiveness is perhaps more surprising, and may be attributable to some of these terms not being useful terms for retrieval from the OCR collection, perhaps due to term matching or term weighting issues.

In order to analyse the impact of term weighting on retrieval effectiveness two further swapping experiments were performed. The SwWgt row in Table 2 show results for exchanging term weights for baseline runs without feedback, while the SwQy2 results show the result of repeating the 20 term PRF experiments from Table 1 with the swapped weights. The results for SwWgt show very little change from the original baseline. This suggests, at least for the terms that have matched in the documents, that the weights for the terms in the original queries are well estimated in the OCR collection relative to the Text collection. For the OCR collection, SwQy2 shows performance improvement with respect to both the average precision and the number of relevant documents retrieved relative to the results in Table 1. It can be seen that applying the appropriate term weights estimated from clean text is effective in removing the negative effects of PRF. This suggests that an important component in the problem of PRF for the OCR collection is inappropriate term weight estimation due to recognition errors. In fact a number of the OCR expansion terms actually have zero weights assigned to them when the weights are swopped since misrecognised terms never appear in the Text collection. Inclusion of these terms in the expanded query has an adverse affect on PRF in the OCR collection which is removed when the weights are estimated for the Text collection. On the other hand, for the Text collection the application

Table 3
The term occurrence statistics for Text and OCR collections.

	Text	OCR
No. of Unique Terms	78611	124810
Terms Occurring Only Once	46626	73612
Terms Occurring More Than Once	31985	51199
Correct Terms	78611	66254
Incorrect Terms	–	58556

of estimated term weights from the OCR collection with the original expanded queries from the Text collection results in a small relative reduction in precision and a large drop in the number of relevant document retrieved. Thus some of the terms which appear in the Text collection, must either not appear or be assigned poor term weights in the OCR collection.

Finally, the SwQyWgt results in Table 2 show performance for swapping both the expansion terms and term weights between the collections. For the OCR collection these figures are even better than those observed for SwQy2. In this case all expansion terms will have non-zero weights as estimated on the Text collection. The comparable results for the Text collection show similar reduction in the average precision and number of relevant documents retrieved for PRF to that for SwQy2; comparing these results to the Text SwQy1 result further illustrates the impact of poor OCR term weights for some expansion terms.

5.3 Analysis of Collections

In practice we will not be able to swap weights between an uncorrupted Text collection and a parallel OCR collection to resolve the problems associated with PRF for DIR. In order to better understand what happens to the estimation of term weights for OCR collection, and to find ways of addressing the problems that this presents for PRF, we investigated the collection statistics. This examination explores how the collection statistics are affected by recognition errors and their implications for term weights.

Table 3 shows the term occurrence statistics for the two collections. The types of difference between the OCR and error-free Text collections are typical of OCR collections (Taghva, Borsack & Condit, 1996a). In particular we can see that there is a 47% increase in the number of unique terms for the OCR collection. Some of these errors arise from words that are broken off at the end of a line in free text e.g *benefi- cial*, while others arise from errors of one or more characters within a word. In Table 3, out of the 73612 terms that occur only

Table 4
Examples of corrupted words with $n(i)$ counts and corresponding $cfw(i)$ term weight estimates.

word	$n(i)$	$cfw(i)$	word	$n(i)$	$cfw(i)$
GOVERNMENT	2757	0.991	LEWINSKY	544	2.754
GOVERNMENT	2390	1.153	LEWINSKY	523	2.764
GOVERNMT	1	8.643	LEWINSKI'S	1	8.643
GOVGRMENT	1	8.643	LEWTNSKY	1	8.643
GOV-CRMMCNT	1	8.643	LEWIUSKY	1	8.643
GOVEMMEUT	2	8.133	LEWINSLY	1	8.643
GOVEMMEUT	2	8.133	LEWINSTCI	1	8.643
GOVERNTN	3	7.796	LEWINSXI	2	8.133
GOVEN MENT	3	7.796	LEWISH	2	8.133

once less than 60% are correct terms which means that some 26,986 incorrectly recognised terms may be classified as highly important terms by our IR system due to their rarity. Further some 19,214 incorrect terms occur more than once. We could of course try to correct these errors in post-processimg with a dictionary, but as observed earlier, Mittendorf & Schäuble (2000) indicate that this can introduce more problems for IR than it solves.

So what exactly happens to cause the disparity between the occurrence statistics of the Text and OCR collections? Table 4 shows examples of how words can be misrecognised with their collection frequency counts $n(i)$ and $cfw(i)$ values calculated based on the collection size $N = 21,759$.

In Table 4 the first line shows the original word from the Text collection and the corresponding correct $n(i)$ and $cfw(i)$ values. The second line shows the same word in the OCR collection and subsequent lines show some examples of corrupted versions of these words found in the OCR collection. It can be seen that each corrupted word is very rare with a correspondingly high $cfw(i)$ value. In some cases the distortion of the word is very minor while on other occasions it is so bad that it is almost impossible to recognise the corrupted word as the original word e.g. *Lewinski* → *LewiIrtci*. As can be seen from Table 4, OCR errors are not consistent, different characters can be changed at different times. For example, in one case the word *Government* is changed to *Govgrment* and in another case to *Govemment*. These inconsistencies in character errors make correcting or managing these errors even more complex.

Our investigation has shown that some of these erroneous terms find their way into the expanded query for the OCR collection, and that these terms can have a negative effect on PRF retrieval performance. When Text collection weights

Table 5
PRF for OCR collection with expansion terms added only if they occur in the Text collection.

	P10	P30	AvP	RelRet
Fbk5	0.578	0.373	0.491	1604
chg bl.	+3.8%	+6.0%	+8.2%	+23
Fbk20	0.600	0.386	0.501	1567
chg bl.	+7.7%	+9.7%	+10.4%	-14

are applied, these corrupted terms will usually have a term weight of zero since they are never observed in the Text collection. Results in Table 2 show that assigning zero weights to errorful terms in this way can help overcome the problems of PRF for OCR collections. Table 5 shows results for a further query expansion experiment for the OCR collection. In this experiment expansion terms are added only if they occur in the Text collection; the Text collection here merely acts as a filter on expansion terms and all term weights are taken from the OCR collection itself.

The results in Table 5 show that when erroneous terms, as defined by their presence or absence in the Text collection, are eliminated from query expansion for the OCR collection improvements of +8.2% and +10.4% are observed over the no feedback baseline for the addition of 5 and 20 expansion terms respectively. Of course, as emphasized earlier, in practice an error-free collection for estimating expansion terms usefulness will not generally exist, but this result illustrates that if we can select appropriate terms, PRF using only the OCR collection has the potential to be effective. The next section develops and evaluates two extended PRF methods which aim to reduce the likelihood of including incorrectly recognized terms in the expanded queries and improve the effectiveness of expansion term selection.

5.4 Effective PRF for OCR Collections

From the analysis in the previous section, the question arises: how might we exclude harmful terms from the list of potential expansion terms for an OCR indexed collection? In this section we introduce and evaluate two adaptations to the standard query expansion term selection procedure. The first is based on a simple removal of expansion terms with low $n(i)$ values, the second more complex strategy aims to merge incorrectly recognised terms with correctly recognised ones in the context of the top ranked documents retrieved in response to a specific search request.

Table 6
Results of excluding terms with $n(i)$ values of between 1 and 3, from the 20 expansion terms used in Table 1.

Media	Text				OCR			
	P10	P30	AvP	RelRet	P10	P30	AvP	RelRet
$n(i) < 2$	0.610	0.391	0.522	1637	0.588	0.385	0.496	1593
chg bl.	+10.7%	+10.5%	+11.5%	+29	+5.6%	+9.4%	+9.3%	+12
$n(i) < 3$	0.610	0.391	0.520	1637	0.588	0.385	0.496	1593
chg bl.	+10.7%	+10.5%	+11.1%	+29	+5.6%	+9.4%	+9.3%	+12
$n(i) < 4$	0.606	0.391	0.517	1637	0.584	0.385	0.491	1593
chg bl.	+10.0%	+10.5%	+10.5%	+29	+4.8%	+9.4%	+8.1%	+12

5.4.1 Removal of Expansion Terms with low $n(i)$ counts

From Table 4 we can see that most corrupted words usually occur with an $n(i)$ of between 1 and 3 with a large percentage occurring with an $n(i)$ value of 1. One simple way to eliminate these terms is to delete potential expansion terms with $n(i)$ values below a given threshold. Thus, we explored this strategy as a simple adaptation of the standard term selection method. Table 6 shows results of an experiment in which potential expansion terms with $n(i)$ values < 2, < 3 and < 4 are ignored. When terms with $n(i) < 3$ are ignored, average precision improves by about +9% relative to the baseline, and reverses the loss in relevant documents retrieved observed in Table 1 with 12 additional relevant documents retrieved. Ignoring terms with $n(i) < 4$ looks to be too aggressive with a slightly smaller observed improvement in average precision. Interestingly this technique also gives a small improvement in retrieval performance over standard unflitered PRF for the Text collection; we intend to explore this behaviour further in future work in the context of larger text retrieval tasks.

The results in Table 5 suggests that expanding by 20 non-zero weighted terms may be more effective. Table 7 shows results of a further experiment excluding terms with low $n(i)$ values as in Table 6, but selecting expansion terms until 20 terms were added to the initial query. Interestingly in all cases these results are a little worse for precision than those in Table 6, although a very small improvement in relevant retrieved is observed for the Text documents.

5.4.2 Merging Expansion Terms using Edit Distance Comparison

This first simple strategy for addressing the problems of PRF for DIR works reasonably well for this collection. However, the optimal value of $n(i)$ may be sensitive to the statistics of individual collections. Additionally, there are two

Table 7
Results of excluding terms with $n(i)$ values of between 1 and 3, and ensuring 20 expansion terms are added.

Media	Text				OCR			
	P10	P30	AvP	RelRet	P10	P30	AvP	RelRet
$n(i) < 2$	0.602	0.385	0.509	1649	0.574	0.376	0.488	1592
chg bl.	+9.3%	+8.8%	+8.8%	+41	+3.1%	+6.8%	+7.5%	+11
$n(i) < 3$	0.602	0.383	0.511	1649	0.574	376	0.488	1592
chg bl.	+9.3%	+8.2%	+9.2%	+41	+3.1%	+6.8%	+7.5%	+11
$n(i) < 4$	0.606	0.388	0.512	1648	0.573	0.376	0.487	1592
chg bl.	+10.0%	+9.6%	+9.4%	+41	+3.1%	+6.8%	+7.5%	+11

notable problems with this approach. First, correctly recognized rare words that would actually be good expansion terms will be deleted along with the incorrectly recognized words, and thus not be available as potential expansion terms. Second, many incorrectly recognized, apparently rare, words can be recognized manually as corrupted versions of correct terms appearing in assumed relevant documents. For example, while the correct term *GOVERNMENT* from Table 4 may appear in some top ranked documents, the incorrectly recognized terms *GOVGRMENT* ($n(i) = 1$) and *GOVEMMEUT* ($n(i) = 2$) may also appear in these or other top ranked documents. These variant forms would be obvious to a human reader based on string similarity and the linguistic content in which they were found. Further, for the OCR collection, the $rsv(i)$ values of the correctly recognized terms will be wrong when a term i has been incorrectly recognized in other top ranked documents. This arises because the $r(i)$ value from Equation 1 can often be significantly underestimated due to these recognition errors. For example, while for *GOVERNMENT* $n(i) = 2390$, its $r(i)$ value in the top 5 ranked documents may only be $r(i) = 2$. If the variant terms had been correctly recognized a value of $r(i) = 4$ or 5 (depending on whether the variants and correct version appear in the same or different documents) may have been recorded. This is potentially a significant problem leading to distortion in the ranking of the $rsv(i)$ ordered list, reduction in the likelihood of choosing correctly recognized useful expansion terms (since their $r(i)$ is underestimated), and potentially consequential reduction in the effectiveness of expansion term selection. Deleting incorrectly recognized words with low $n(i)$ values from top ranked documents will not address this problem of underestimation of the $r(i)$ value. Based on these observations we next describe another modified term selection method.

Both of the problems described above can be overcome by identifying misrecognized words within assumed relevant documents and merging them with correct words. String comparison algorithms compute an “edit distance” be-

Table 8
Results for OCR documents using string-comparison term merging.

Media	OCR			
MinEd	P10	P30	AvP	RelRet
1	0.576	0.380	0.489	1593
chg bl.	+3.4%	+8.0%	+7.7%	+12
2	0.582	0.380	0.492	1581
chg.bl.	+4.5%	+8.0%	+8.4%	+0
3	0.596	0.385	0.505	1610
chg. bl.	+7.0%	+9.4%	+11.2%	+29
4	0.588	0.388	0.508	1616
chg. bl.	+5.6%	+10.2%	+11.9%	+35
5	0.592	0.386	0.507	1607
chg. bl.	+6.3%	+9.7%	+11.7%	+26

tween two strings giving the minimum number of changes required to convert one string to the other one (Zobel & Dart, 1996). These algorithms can make mistakes, sometimes merging words that are not related. However, in the constrained context of the small number of documents assumed to be relevant to a search query, it is likely that similar character strings really are the same word, leading to only a very small number of false merges. The small number of false merges that do occur could be expected to have an insignificant impact on PRF effectiveness. In order to explore this merging hypothesis a further experiment was conducted in which all words appearing in the summaries of the top 5 ranked documents were compared using the string comparison algorithm described in Zobel & Dart (1996). Words within a preset edit distance were merged with the word with the larger $n(i)$ value assumed to be the correct spelling. The $r(i)$ values of merged words were added, and the $n(i)$ value taken as that of the larger value. The reduced set of potential expansion terms was then ranked by the $rsv(i)$ computed using the merged $r(i)$ values.

Table 8 shows the result of using this merging approach with 20 expansion terms for edit distance values of 1, 2, 3, 4 and 5. From Table 8 it can be seen that in all cases the merging procedure produces an improvement in average precision over the no feedback baseline in Table 1. The improvement increases as the maximum edit distance allowed to merge two strings increases up to a value of 4 characters. At this point average precision increases by almost +12% relative to the no feedback baseline in Table 1 to 0.508, while there is also an overall increase of +35 in the total number of relevant documents retrieved. Significantly this result is achieved without use of domain or topic specific external linguistic or data resources. There is little variation between results for maximum allowed edit distance values of 3, 4 and 5, suggesting that using a value of 4 will give good average stability across different queries for this collection. While this techniques has been demonstrated to work effectively on this test collection, it remains for it to be tested on alternative DIR tasks when these become available.

The success of this technique can be attributed to the elimination of a number of highly weighted rare misrecognized terms from the feedback terms for two reasons. First, the rank of non-relevant documents in the assumed relevant set which contain these terms is not now promoted by addition of these highly weighted terms to the search query. As shown in Table 4, the relative $cfw(i)$ values of rare terms is very high and thus inclusion of a single one of these terms in an expanded query can be enough to significantly increase the matching score of a document containing this term. The rank of non-relevant documents may still be promoted due to the presence of other expansion terms, but this is a general drawback of query expansion in PRF for all media types.

As shown in Table 4, incorrectly recognised terms can have $n(i)$ values greater than 1 and occur across the collection. Thus in addition to their presence in the assumed relevant document set, individual misrecognized terms can also occur in other documents, effectively with a random distribution. These other documents containing incorrect terms may include some or none of the original query terms, but when the query is expanded to include the highly weighted errorful terms, the matching score of the documents containing them can increase dramatically relative to other documents. While these documents may be relevant to the search request, it is most likely that they will often not be relevant. Their increase in relative matching score can significantly increase their rank in the output of the PRF stage. Thus, a document which previously matched with none of the original query terms, may now match with an errorful expansion terms with say $n(i) = 2$, where the only other occurrence of the terms is in one of the assumed relevant documents. The presence of this errorfully recognized term in these two documents tells us nothing about the potential relevance of either of them, but such errors may have a significant impact on retrieval behaviour. The effect of this behaviour can be seen in Table 1 where expansion leads to a fall of -196 in the total number of relevant documents retrieved. Here non-relevant documents containing expansion terms are being retrieved at the expense of relevant ones. Eliminating these rare terms using the term merging technique described above prevents this problem.

Overall then the merging technique gives better estimation of $rsv(i)$ due to more accurately calculating $r(i)$, and prevents problems of over promotion of documents containing errorful terms. It has the attraction of making no assumptions about word distribution characteristics of the collection, and makes no use of general or domain specific resources such as dictionaries. Although as stated earlier it still needs to be evaluated in alternative collections. While retrieval using this single method of term merging based on a simple edit distance is effective, use of a more sophisticated merging procedure taking into account the observed characteristics of OCR errors could be expected to yield more accurate merging, with a consequential further improvement in PRF effectiveness, albeit probably a small one.

A good starting point for an improved merging technique could be the OCRSpell software described by Taghva & Stofsky (2001). OCRSpell was designed for interactive correction of OCR errors, and does not exploit the topical context of the terms provided in our PRF method. Thus it would need to be adapted for fully automated correction and could be extended to take of the context of individual terms. Alternatively, both out simple edit distance metric and OCRSpell could be used in an interactive mode for PRF. Additionally, term corrections discovered during PRF could be incorporated into the standard document index so that they are available for subsequent search requests.

5.5 Analysis of True RF

Based on our results and analysis we can compare the behaviour for query expansion for OCR documents collections for PRF and true RF as described in Taghva, Borsack & Condit (1996a). The problems of adding inaccurate recognized terms for PRF have been explored here. For true RF, effectiveness is evaluated in terms of its impact on a residual collection. This collection consists of documents not previously assessed for relevance. Thus rare incorrectly recognized terms are likely to be added to the expanded query, as observed in our experiments, even when true RF mean that all documents used for feedback are actually relevant. However, since these terms are rare with most of their occurrences actually in the documents assessed for relevance, their impact on ranking the residual collection will be limited to terms with $n(i)$ values greater than 1. Also since these terms have been added only due to their presence in relevant documents, their impact on lower ranked documents in the feedback run is likely to be lower than in the PRF case where these terms are taken from both relevant and non-relevant documents which have been assumed to be relevant. This analysis exactly predicts the type of behaviour shown in Table XVII of (Taghva, Borsack & Condit, 1996a) where increasing the number of expansion terms for the OCR collection produces small unpredictable perturbations in the average precision, while increasing the number of expansion terms produces continued improvements in a parallel Text collection. More recent experiments by the same group on automatic query expansion using interactive RF adding additional terms decreased retrieval performance compared to a no feedback baseline Taghva et al. (2004). This was attributed to over expansion of the query to include terms not associated with document relevance causing the focus of the query to drift. However, this result is also consistent with our analysis, unfortunately (Taghva et al., 2004) does not report any analysis of the individual selected expansion terms.

6 Conclusions and Further Work

Scanned images of paper documents form an important part of information repositories and maximizing their potential utility requires effective retrieval mechanisms. This paper has demonstrated and analysed the failure of standard PRF methods to transfer simply to DIR. The reasons for this failure have been examined in detail and two extensions to standard PRF were proposed and evaluated. Best results were achieved using a method based on string comparison and merging of related words which delivered an increase of almost +12% in average precision over a baseline no feedback run, and overcame the significant loss in relevant documents retrieved observed for standard PRF. In comparison this average precision result was 98.8% of the result achieved using a parallel uncorrupted Text collection in conjunction with a standard PRF method, and 97.3% of the best result achieved for this Text collection using one of the extended methods. The corresponding figures for the total number of relevant documents retrieved are 99.1% and 98.7% respectively. The extended PRF techniques make no use of external linguistic resources, and as such should be applicable to other scanned documents without collection specific adjustment beyond possibly adjusting the string edit distance merging decision criterion.

Further work will investigate the comparative stability of our modified PRF methods for individual search queries with respect to each other and to the Text collection, confirmation of our analysis for the behaviour of true RF and further possible applications of string comparison methods making use of term and document context. For example, for OCR text derived from frames of video data. It would be interesting to explore the incorporation of more complex term merging procedures, such as the methods used in OCRSpell (Taghva & Stofsky, 2001). It is also important to investigate the behaviour of our PRF methods on scanned document collections of varied image quality. Of particular interest would be its potential to improve retrieval effectiveness on images with very poor OCR performance.

Ongoing work is also focusing on further investigation of the wider issues of retrieval behaviour of documents within OCR collections, in particular preliminary results suggest that retrieval effectiveness of short documents is disproportionately affected by the use of the output of automatic OCR. We intend to explore this further and propose solutions to this problem.

References

Bodleian Library, University of Oxford (2004). http://www.bodley.ox.ac.uk

Grossman, D. A., Lundquist, C., Reichart, J., Holmes, D., Chowdbury, A., & Frieder, O. (1997). Using Relevance Feedback within the Relational Model for TREC-5. In: Voorhees, E. & Harman, D. (Eds.) *Proceedings of the Fifth Text REtrieval Conference (TREC-5)*, NIST Special Publication 500-238, pp. 477-487.

Harding, S. M., Croft, W. B., & Weir, C. (1997). Probabilistic Retrieval of OCR Degraded Text Using N-grams. In *Proceedings of the First European Conference on Research and Development for Digital Libraries*, Pisa, pp345-359, Springer.

Irish Script on Screen (2004). http://www.isos.dias.ie

Johnson, S. E., Jourlin, P., Spärck Jones, K., & Woodland, P. C. (1999). Spoken Document Retrieval for TREC-8 at Cambridge University. In Voorhees, E. & Harman, D. (Eds.) *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, Gaithersburg, MD, pp. 109-126, NIST.

Johnson, S. E., Jourlin, P., Spärck Jones, K., & Woodland, P. C. (2000). Spoken Document Retrieval for TREC-9 at Cambridge University. In Voorhees, E. & Harman, D. (Eds.) *Proceedings of the Eighth Text REtrieval Conference (TREC-9)*, Gaithersburg, MD, pp. 117-126, NIST.

Jones, G. J. F., & Han, M. (2001). 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.

Jones, G. J. F., & Lam-Adesina, A. M. (2002). An Investigation of Mixed-Media Information Retrieval. In *Proceedings of the 6th European Conference on Research and Development for Digital Libraries*, Rome, pp. 463-478, Springer.

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

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

Lynam, P, Varian, H. R., Swearingen, K., Charles, P., Good, N., Lamar Jordan, L., & Pal, J. (2003). How Much Information? 2003. http://www.sims.berkeley.edu/research/projects/how-much-info-2003/

Mittendorf, E. & Schäuble, P. (2000). Information Retrieval can Cope with

Many Errors. *Information Retrieval*, 3:189-216.

Marukawa, K., Hu, T., Fujisawa, H., & Shima, Y. (1997). Document Retrieval Tolerating Character Recognition Errors - Evaluation and Application. *Pattern Recognition*, 30(8):1361-1371.

The Oxford Digital Library (2004). `http://www.odl.ox.ac.uk`

Porter, M. (1980). An algorithm for suffix stripping. *Program*, 14(3):130-137.

Robertson, S. E. (1990). On Term Selection for Query Expansion. *Journal of Documentation*, 46(4):359-364.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., & Gatford, M. (1995). Okapi at TREC-3. In Harman, D. (Ed.)*Proceedings of the Third Text REtrieval Conference (TREC-3)*, Gaithersburg, MD, pp. 109-126. NIST.

Spärck Jones, K. & Willett, P. (1997). *Readings in Information Retrieval*, Morgan Kaufman.

Taghva, K., Borsack, J., & Condit A. (1996). Evaluation of Model-Based Retrieval Effectiveness with OCR Text. *ACM Transactions on Information Systems*, 14(1):64-93.

Taghva, K., Borsack, J., & Condit, A. (1996). Effects of OCR errors on Ranking and Feedback using the Vector Space Model. *Information Processing and Management*, 32(3):37-332.

Taghva, K. & Stofsky, E. (2001). OCRSpell: an interactive spelling correction system for OCR errors in text. *Internation Journal of Document Analysis and Recognition*, 3(3):125-137.

Taghva, K., Borsack, J., Nartker, T., & Condit, A. (2004). The Role of Manually-Assigned Keywords in Query Expansion. *Information Processing and Management*, 40(3):441-458.

Tong, X., Zhai, C., Milić-Frayling, N., & Evans, D. A. (1997). OCR Correction and Query Expansion for Retrieval on OCR Data - CLARIT TREC-5 Confusion Track Report. In: Voorhees, E. amd Harman, D. (Eds.) *Proceedings of the Fifth Text REtrieval Conference (TREC-5)*, NIST Special Publication 500-238, pp. 477-487.

Zobel, J., & Dart, P. (1996). Phonetic String Mathing: Lessons from Information Retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 96)*, Zurich, pp. 30-38, ACM.