

Effects of OCR Errors on Ranking and Feedback Using the Vector Space Model

Kazem Taghva*, Julie Borsack and Allen Condit
Information Science Research Institute
University of Nevada, Las Vegas

June 1995

Abstract

We report on the performance of the vector space model in the presence of OCR errors. We show that average precision and recall is not affected for our full text document collection when the OCR version is compared to its corresponding corrected set. We do see divergence though between the relevant document rankings of the OCR and corrected collections with different weighting combinations. In particular, we observed that cosine normalization plays a considerable role in the disparity seen between the collections. Furthermore, we show that even though feedback improves retrieval for both collections, it can not be used to compensate for OCR errors caused by badly degraded documents.

*Email: taghva@cs.unlv.edu.

1 Introduction

As the technology improves and the costs of devices drop, the applications for optical character recognition (OCR) continue to grow. Recognizing full-text documents is one of OCR's more challenging but applicable objectives. A common subsequent goal is the retrieval of these documents. But what effect will the errors caused by this method of capture have on information retrieval (IR)? The determination of the effects of OCR error on IR is the intent of this project and our previous works (Taghva, Borsack, Condit, & Erva, 1994c) (Taghva, Borsack, & Condit, 1994b) (Croft, Harding, Taghva, & Borsack, 1994).

Our previous experiments show that "noise" produced by OCR misrecognition has little effect on *average* recall and precision when compared to manually corrected ASCII. We believe this result is due to the following factors:

- There is enough redundancy in full text to compensate for the amount of error caused by misrecognized characters.
- Since OCR devices *see* the whole page, nothing is selectively removed (i.e. running headers, references, captions, etc.).
- Errors occur in manually typed text as well as optically recognized text, so no text of significant length can be considered 100% accurate.

Does this imply that OCR can be used for any full text document capture task for information retrieval without negative effects? If average recall and precision is the only measure, probably. But we have also found that certain characteristics of OCR data combined with more sophisticated document ranking techniques may affect individual queries (Taghva et al., 1994b). Individual query evaluation, in some sense, is a more definitive criterion; these are the results by which a user judges an IR system. Further, if the document texts are short, less redundancy exists, and more degradation of recall and precision will be seen (Croft et al., 1994). Our previous works support the theory that, in general, OCR and IR can be applied in succession with little human intervention. But characteristics of the original collection and the IR system to be applied should not be underestimated.

With this in mind, we expand our study to the SMART Retrieval System (Salton, 1971). SMART is based on the well-known vector space model, giving us a new framework for our experimentation. Also, in its implementation, SMART provides most of the standard algorithms tested and applied in IR research. In particular, we investigate the effect of various term/document weighting algorithms. We know from our previous work that a Boolean system, which records only the presence or absence of terms, shows virtually no discrepancies when the OCR and corrected collections are compared (Taghva et al., 1994c). But, we have also discovered that variation in document ranking occurs with statistical-based systems dependent on specific word frequencies (Taghva et al., 1994b). We want to be able to explicitly identify those components that may work well with OCR data and those that may cause unanticipated problems.

Another aspect we considered in this current work with SMART is the application of relevance feedback. Feedback has always been an integral element of IR research and has been applied in operational systems since 1977 (Harman, 1992a). Since relevance feedback is an automatic process that uses information directly from the collection, this process may be hampered by OCR noise. We evaluate these aspects of IR, report on our findings, and review specific problems we have encountered in our various experiments with OCR-generated text.

2 Test Collection

The test collection we use is a subset of a larger collection given to our institute by the Department of Energy (DOE) for continued research in the areas of optical character recognition and information retrieval. Two versions of this single collection (674 documents) are used—a recognized version¹ and the manually corrected version of this same set.² No accuracy rates were calculated for the OCR document set but for the type of input we used, a conservative estimate would be an 80-90% level of character accuracy(Futrelle et al., 1991).

The collection consists of full-text documents. The average document length is forty pages and the median length is sixteen pages. For more information about the complete document collection, see (Nartker, Bradford, & Cerny, 1992).

The 68 queries run against these test sets are the same queries used in our previous experiments. They were initially constructed for this document collection to test the collection's usefulness for potential users. After extraneous terms were removed, the average number of words per query is five.

The relevancy judgments were not part of the data we received from the DOE. Graduate students in fields related to topics covered in the collection have read each document and assessed its relevance to each query. Excluding seven of the queries having no relevant documents in the collection, the average number of relevant documents per query is 10. Our previous papers cover these elements of our experiment in more detail(Taghva et al., 1994c)(Taghva et al., 1994b).

3 SMART Vector Representation

The most distinctive quality of a vector space IR system is its simple but elegant method of representing both the document and query as vectors. With the same vector representation, a natural measure of correlation between queries and documents is the cosine angle measure. Although other methods exist, this is the similarity measure applied in SMART. The method used to weight document terms though, is not unique. There have been numerous algorithms devised, and it seems from the literature(Salton & Buckley, 1988), that the appropriate scheme to use depends heavily on the characteristics of the collection. SMART includes most of the well-tested weighting schemes, providing a rich environment for experimental testing.

The weighting algorithms in SMART are formulated by combining three parameters: the term frequency component, the collection frequency component, and the vector normalization component. Commonly applied schemes are available in each category. Any combination of these can be applied to weight the document and query vector terms, although particular combinations have proven to be superior for collections with certain characteristics(Salton & Buckley, 1988)(Harman, 1992b). We apply components of selected weighting algorithms singly, helping to identify the effect each may have on query results. The function of each is described as follows:

term frequency component measures the frequency of occurrence of terms in the documents. Some methods normalize term frequencies to between 0 and 1. The most common method divides each document term's frequency by the maximum term frequency of that document.

¹produced by OCR.

²Correctness level of main body text was quoted as 99.8%.

collection frequency component applies collection frequency information (i.e. documents in the collection, documents to which a term has been assigned) magnifying term weights that are concentrated in a few documents. The expected result is to discriminate such documents from the rest of the collection.

vector normalization component equalizes vector length so that no advantage is given to longer document representations. This factor may not be necessary if the collection's document lengths are uniform. Cosine normalization is a commonly applied technique.

Vector representation is also defined by other factors. For example, the stop words that are explicitly not indexed affect this representation. In our testing, we alter only the stopword list and the stemming method (s removal only); all other factors affecting vector representation are left at their default values (Salton, 1971). These settings keep this experiment consistent with our previous projects.

4 Impact of Weighting Parameters

We use the versatility of SMART to understand more fully the effects of these components on the OCR-generated text. We applied several weighting combinations to the collections trying to identify a noticeable variance in average precision. The following weights from each component class were applied to our collections:

- **term frequency component**

Let tf denote the term frequency of a term t , then new_tf weights the terms according to the following schemes:

none, in symbol n: $new_tf = tf$

augmented normalized, in symbol a: $new_tf = 0.5 + 0.5 * (\frac{tf}{max_tf})$
where max_tf is the largest tf value in the vector.

log, in symbol l: $new_tf = \ln(tf) + 1.0$

square, in symbol s: $new_tf = tf^2$

- **Merging of collection frequency component**

Let num_docs , $coll_freq_of_term$, and $coll_freq$ denote the number of documents in the collection, the number of documents to which term t is assigned, and the total number of occurrences of the term t in the collection respectively, then new_wt is defined as follows:

none, in symbol n: $new_wt = new_tf$

inverse document frequency weight (tfidf), in symbol t: $new_wt = new_tf * \log(\frac{num_docs}{coll_freq_of_term})$

probabilistic, in symbol p: $new_wt = new_tf * \log(\frac{num_docs - coll_freq}{coll_freq})$

squared, in symbol s: $new_wt = new_tf * (\log(\frac{num_docs}{coll_freq_of_term}))^2$

- **Merging of vector normalization**

Let m denote the number of entries in the vector, then $norm_wt$ is defined as follows:

weight applied	corrected	OCR	% of difference
nnn.atn	0.2457	0.2497	1.63
ann.atn	0.3222	0.3308	2.67
lnn.atn	0.3311	0.3421	3.32
snn.atn	0.2053	0.2006	-2.29
ntn.atn	0.2898	0.2881	-0.59
npn.atn	0.3110	0.3153	1.38
nsn.atn	0.3305	0.3269	-1.09
nns.atn	0.2913	0.2873	-1.37
nnc.atn	0.3500	0.3381	-3.40
atc.atn	0.2228	0.2235	0.31

Table 1: Results show no significant difference in average precision between the corrected and OCR collection using different weighting schemes

none, in symbol n: $norm_wt = new_wt$

sum, in symbol s: $norm_wt = \frac{tf}{\sum_m new_wt}$

cosine, in symbol c: $norm_wt = \frac{tf}{\sqrt{\sum_m new_wt^2}}$

We exploit the notation used in SMART to describe the combined schemes: *xxx.xxx*. The three characters to the left of the period refer to the document weighting combination while the characters to the right refer to the query weighting combination. For example, **lpc.atn** would apply log term frequency weighting and probabilistic collection frequency weighting with cosine normalization to the document collection, and apply augmented normalized term frequency weighting and **tfidf** collection frequency weighting with no normalization to the queries. For all runs, we hold the query weighting constant at **atn**, the weighting suggested for the type of queries in our collection (Salton & Buckley, 1988). The 11-pt average precision for each collection and the percentage difference between them appears in Table 1. Note that with some weighting schemes, the OCR version returned slightly higher precisions. We attribute this seemingly contrary result to human error and accidental and purposeful exclusion of text by the editors of the corrected collection.

We vary each weighting element singly to determine if any of the components might have consequential impact. But none of the average precision results were significantly different³ when the averages are compared for the two collections. We apply **atc.atn**, a combined scheme, since this is the suggested weighting for our collection type. From these results, we conclude that none of the commonly applied weighting algorithms significantly affect average precision when applied to OCR generated text. It seems inherent qualities of this collection were a stronger influence on results than whether documents had been optically recognized or manually typed.

³Significant here translates to more than 5% difference in average precision.

query	document	<i>nnn.atn</i>		<i>atc.atn</i>	
		corrected rank	OCR rank	corrected rank	OCR rank
4	13	216	226	20	204
5	13	278	275	75	263
5	24	220	271	7	128
6	264	324	316	15	47
7	22	134	129	20	72
7	53	236	242	34	110
10	253	79	77	60	215
10	573	184	179	18	47
11	504	14	15	44	178
16	403	52	57	46	141
16	618	54	52	73	42

Table 2: Ranking variability between corrected and OCR collections increases with more complex weighting schemes

5 Ranking Variability

In our previous experiment (Taghva et al., 1994b) with INQUERY (Callan, Croft, & Harding, 1992), certain OCR data characteristics caused unstable results for individual queries. This consequence was also apparent in this experiment, and was expected. The term weighting algorithms for IR systems that incorporate ranking use similar weighting techniques. Still, with SMART, we were able to compare different weighting algorithms at the query level and evaluate the ramifications of weighting components. We noticed discrepancy in the relevant document rankings between the corrected and OCR collections for all the weighting schemes, but for some weighting combinations, the difference was more extreme. For example, portions of the relevancy ranking for two weighting schemes appear in Table 2. This table displays the query number, the relevant document id, and the rankings for the two weighting schemes, *nnn.atn* and *atc.atn*, for the corrected and OCR collections respectively. Notice the relative closeness of the corrected document rank to the OCR document rank for the *nnn.atc* weighting scheme. For those same query-document pairs using the *atc.atn* weighting, the relevant document ranks became much more divergent. Intuitively, we felt there was a greater disparity in ranking between the corrected and the OCR document rankings as the weighting schemes became more complex.

To confirm the significance of the disparity observed and determine its cause, a more thorough analysis was done to investigate the effects of the three factors: *a*, *t*, and *c*. We first found the difference (OCR document rank – corrected document rank) for each of the 617 query-document pairs (relevant documents only); we then calculated the mean and standard deviation for this set of 617 values. Table 3 shows the results of these calculations for the weighting scheme *nnn*, where no factors are applied, and for each weighting combination adding each of the three factors of the *atc* document weighting.

The ordering of the weighting algorithms in Table 3 is not random. Beginning with *nnn*, if you group the algorithms by two, you will notice that each pair only differs by *c*, the cosine normalization measure. Note that when cosine is added, the mean increases substantially. To analyze this increase in mean we use the 2³ factorial experiment on the three factors, *a*—term frequency component, *t*—collection frequency component, and *c*—vector normalization

weighting scheme	mean	standard deviation
nnn	-1.063	24.534
nnc	8.760	46.256
ann	1.437	26.910
anc	6.600	43.752
ntn	-0.575	22.104
ntc	4.076	34.070
atn	1.217	26.054
atc	8.648	50.115

Table 3: Differences between OCR and correct document ranking for each permutation of the `atc` document weighting scheme reported in mean and standard deviation

component. The 2^3 factorial experiment is a standard statistical method for testing all possible combinations of factors with only two levels. This test examines the effect of each factor, or *source variable*, singly and it also examines the interactions of the source variables in combination, on the dependent variable—difference in ranking. The analysis of variance from the 2^3 factorial experiment appears in Table 4. For a complete discussion of this method of evaluation and how to interpret the results, see (Anderson & McLean, 1974).

These results show that the cosine normalization component, with a significance probability (Pr) value of 0.0001, has a highly significant effect. But since the variances of the response (difference in ranking) for the treatment combinations are dissimilar, we rerun the factorial test on the log-transformation to stabilize the variance. The minimum value of the response (difference in ranking) was -359 , so we take each difference, add 360, and take the natural log. The results for this analysis for the log-transformed variable are given in Table 5. This table shows that even on a log-scale, the cosine factor is highly significant.

Although not always the case, the most common difference between the `atc.atn` rankings was the improvement in document ranking for the corrected collection over the OCR collection (see Table 2); from our variance test, cosine normalization is the significant factor causing this discrepancy.

We know that “garbage strings” increase the length of the OCR document vectors. This is especially true for documents with many misrecognized terms and/or graphic text. We found that the OCR vector length was approximately three times the length of the corresponding corrected-text document vector for those documents with a high discrepancy in ranking. Aggravating this fact, “garbage strings” tend to have high term weights.⁴ When these terms are normalized, their significance to the document increases in relationship to other terms in the collection.

6 Post-processing

Our post-processing system (pps) is an automatic error detection and correction program designed for OCR text. The system incorporates knowledge specific to the kinds of problems one would encounter when OCR text is used directly with an IR system (i.e. misrecognized terms and graphic text). We explain this system in detail in (Taghva, Borsack, & Condit,

⁴TREC experiments showed similar problems with misspellings in their collections.

factor	DF	Sum of Squares	Mean Square	F Value	Pr > F
a	1	3,514.429	3,514.429	2.75	0.0973
t	1	449.173	449.173	0.35	0.5532
c	1	56,697.617	56,697.617	44.38	0.0001
at	1	2,757.034	2,757.034	2.16	0.1419
ac	1	260.066	260.066	0.20	0.6519
tc	1	670.332	670.332	0.52	0.4689
atc	1	4,218.187	4,218.187	3.30	0.0693
error	4928	6,295,489.89	1277.49		

Table 4: Analysis of variance for the 617 difference values (OCR document rank – correct document rank)

factor	DF	Sum of Squares	Mean Square	F Value	Pr > F
a	1	0.0414	0.0414	2.50	0.1137
t	1	0.0073	0.0073	0.44	0.5065
c	1	0.4092	0.4092	24.73	0.0001
at	1	0.0158	0.0158	0.96	0.3273
ac	1	0.0222	0.0222	1.34	0.2465
tc	1	0.0007	0.0007	0.04	0.8345
atc	1	0.0112	0.0112	0.68	0.4106
error	4928	81.5720	0.0165		

Table 5: Log-scale analysis of variance for same set of 617 difference values

1994a)(Taghva, Borsack, Bullard, & Condit, 1993). After running the OCR text collection through the post-processing system, we reload the collection into SMART using the `atc.atn` weighting scheme; there was no significant improvement in average precision. It seems the corrections made by the post-processing system could not stabilize the effects of cosine normalization described in Section 5. This lack of improvement is due in part to our system’s conservative approach to the removal of “garbage strings.” These types of strings are difficult to characterize since they syntactically resemble real words (e.g. balance of consonants and vowels). The vector length decreased by some margin for problem documents after running pps, but not enough to improve the high variability found in document ranking between the post-processed and the corrected collections.

Another problem we encountered was with `t`—the collection component of `atc` document weighting. Terms that appear in only a single document, especially if these terms appear more than once in a given document, have very high weights. Unfortunately this characterizes misrecognized terms and graphic text strings. If an OCR device is having trouble with a particular document’s font and consistently misrecognizes certain characters, an erroneous term will occur frequently, but only in that particular document. Table 6 shows some of these problem terms, their weights, and their ranked order based on this weight within the document.

document id	problem terms	atc weight	term rank
253	soilratio	0.02707	10th out of 2831
13	hi5	0.08939	4th out of 300
377	liball	0.02501	1st out of 2882
377	verily	0.02478	4th out of 2882
38	wnw	0.17328	8th out of 100
38	florjan	0.25540	1st out of 100
504	dcon	0.04152	2nd out of 2013
504	tm32	0.03759	6th out of 2013

Table 6: Useless terms with high atc weight

7 Effect on Relevance Feedback

The notion of relevance feedback was introduced in the mid 1960’s and its effectiveness has been proven under experimental conditions (Salton & Buckley, 1990). Feedback is an automatic process that uses information derived from known relevant and non-relevant documents to reformulate queries. There are two commonly employed modification techniques: query term reweighting and query expansion (Frakes & Baeza-Yates, 1992). These can be used singly or in conjunction with each other. A number of papers have been written on the subject of relevance feedback. For a complete explanation of these techniques see (Harman, 1988)(Harman, 1992a)(Harper, 1980)(Salton & Buckley, 1990).

Two feedback methods implemented in SMART are *Ide dec-hi* (Ide, 1969) (Formula 1 in Figure 1), and *Rocchio* (Rocchio, 1971) (Formula 2 in same figure). Both use a combination of query expansion with term reweighting. The vectors for the retrieved documents, \mathcal{R}_i and \mathcal{S} or \mathcal{S}_j , are merged with the original query vector, \mathcal{Q}_0 . The weights of the original query terms are adjusted accordingly, based on their occurrence in the relevant and nonrelevant documents. Additional terms are added to the query from the relevant documents while the terms in the nonrelevant documents are used to balance the weight of these newly added terms.

Intuitively, term reweighting should not be heavily affected by using an OCR generated collection. We have shown in (Taghva et al., 1994b) that in OCR text, *correct* word frequencies rarely decrease by a substantial amount. Further, since both the relevant and non-relevant documents used to determine distribution were highly ranked, the initial weight of these terms was not affected by other OCR text complications (Taghva et al., 1994b). What one would more readily question is query expansion. New words extracted from the text are added to the query based typically on their frequencies within the relevant documents’ text. With this in mind, if misrecognized terms extracted from OCR text are added to the query, at best they may have no effect or worse, they may decrease a query’s effectiveness.

To determine if the advantage of relevance feedback is hindered when an OCR collection is used, we applied the described feedback methods to each collection using the following suggested parameters (Salton & Buckley, 1990)(Harman, 1992a):

atc.atc document/query weighting for all runs.

partial query expansion with 20, 50, 100, 250, 500, 750, and 1000 additional terms.

$$\mathcal{Q}_1 = \mathcal{Q}_0 + \sum_{i=1}^{n_1} \mathcal{R}_i - \mathcal{S} \quad (1)$$

$$\mathcal{Q}_1 = \mathcal{Q}_0 + \beta \sum_{i=1}^{n_1} \frac{\mathcal{R}_i}{n_1} - \gamma \sum_{j=1}^{n_2} \frac{\mathcal{S}_j}{n_2} \quad (2)$$

where

- \mathcal{Q}_0 = the vector for the initial query
- \mathcal{R}_i = the vector for relevant document i
- \mathcal{S} = the vector for topmost nonrelevant document
- \mathcal{S}_j = vectors for nonrelevant documents
- n_1 = the number of relevant documents
- n_2 = the number of nonrelevant documents (Frakes & Baeza-Yates, 1992)

Figure 1: Formulas for feedback methods implemented in SMART.

reduced document weighting of $\beta = 0.75$ and $\gamma = 0.25$ used in the Rocchio method (see Formula 2)

The top fifteen documents retrieved by SMART were used in these feedback formulas. All other adjustable parameters were left at their default values.

Table 7 shows the average precision results for the Ide dec-hi method at each specified expansion length and the Rocchio feedback method using only the suggested partial expansion of 20 terms. These results show the 11-pt average precision for each feedback method, the percentage of change over its *residual* collection’s 11-pt average precision, and in the last column, the percentage of difference between the precision of the corrected and OCR feedback runs. A residual collection can be defined as all documents in the original collection minus all items previously seen by the user. Recall and precision measured on this reduced collection is one accepted method for evaluation. Since retrieved documents are removed from residual collections the one-to-one mapping between relevant documents in the corrected and OCR collections no longer exists. Still, we feel the comparison of the feedback runs are a valid measure of what should be expected for an OCR collection.

Since in previous feedback experiments and in these results from our feedback runs, the Ide dec-hi method has proven superior, we selected this method to analyze more closely. Table 7 shows a consistent increase in average precision as more terms are added to the queries from the corrected collection.⁵ This same increase is not apparent in the OCR collection; in fact, the OCR collection’s precision values level off after the 20 term expansion. This divergence is apparent by the two collections increasing difference shown in the last column of Table 7.

Since the only change in each run was the augmented queries, our first guess for the discrepancy was query degradation by the addition of meaningless terms. But after studying the feedback queries, we found that those generated from the OCR text were not heavily affected by misspellings or “garbage strings.” In fact, as a whole they were quite clean;⁶ out

⁵This improvement has been seen in previous feedback experiments as well.

⁶We believe this is partially due to the method of term selection for the queries.

Ide dec-hi					
	correct		OCR		
partial expansion	feedback	% change over residual	feedback	% change over residual	% of difference
20	0.2000	30.1	0.1928	36.6	-3.6
50	0.2211	43.8	0.2045	44.8	-7.5
100	0.2164	40.7	0.2089	48.0	-3.5
250	0.2240	45.6	0.2071	46.7	-7.5
500	0.2328	51.4	0.2031	43.9	-12.8
750	0.2422	57.5	0.2092	48.2	-13.6
1000	0.2536	64.9	0.2095	48.4	-17.4
Rocchio					
20	0.1770	15.1	0.1593	12.8	-10.0

Table 7: Average precision improves for the correct collection as more terms are added to the feedback queries but remains fairly constant for the OCR collection

of the 48,753 feedback query terms in the 1000 term expansion, 7% qualify as misspellings or “garbage strings.”

So what caused the increasing difference? Referring again to Table 7, note that the difference is a consequence of the continued improvement in average precision of the corrected collection not the deterioration of the OCR results. Some of the documents that moved to a better rank (i.e. top fifteen) in the corrected collection were not influenced in the same way by query expansion in the OCR collection. We noticed that a good number of the relevant OCR documents that did not improve in rank for the feedback runs had also been the OCR documents that had poor rank in comparison to the corrected documents in our initial retrieval runs. These documents which caused problems with cosine normalization, in general, had repetitive, tabular data or had numerous OCR-generated errors. Table 8 shows some of the rankings of these documents for both the corrected and OCR collections. We believe from this analysis, that the ranking differences for both cosine normalization and feedback are a result of a few difficult-to-retrieve documents within the OCR collection.

Also, with less impact, it seems the OCR collection did not realize the same gains from feedback as the corrected collection. When additional terms are added to a query with feedback, the more frequently occurring terms (and it seems the most useful), receive less weight. In the OCR collection, some of the less frequent terms added when increased expansion is applied, are erroneous and indeed useless. These two phenomena reduce an OCR feedback query’s effectiveness.

The results of the single feedback expansion run with the Rocchio method appears in the second part of Table 7. Here again, a drop in average precision occurs between the OCR text and its corrected counterpart. Although this method was not analyzed in as much detail, we did find the following:

- The same terms that augment the queries in the Ide dec-hi method augment the queries in the Rocchio method.
- The reduced document term weighting causes high weight to be assigned to original query terms while document term weights are markedly lower. This is true for both collections, although it is more pronounced in the OCR collection.

query	document	corrected		OCR	
		original	feedback	original	feedback
5	25	27	3	40	114
10	573	18	3	47	32
16	604	40	8	36	53
28	468	16	1	45	41
28	469	83	14	99	79
31	377	38	1	122	189
32	327	40	1	48	44
33	348	30	11	147	132
33	504	21	5	178	164
33	513	20	3	56	43
35	8	53	13	56	58
35	90	33	9	36	55
40	9	33	1	154	82
57	53	96	3	185	160
58	377	38	1	104	42
60	328	52	4	45	41
60	376	94	2	343	399
68	376	29	14	113	189

Table 8: The corrected collection’s improved ranking over the OCR collection for some documents

We do not have enough evidence to state that the percentage of difference between the collections in the Rocchio run is a direct result of the use of OCR text. But from the combined results of both methods, it is apparent that feedback cannot be applied with the hope of fixing the few shortcomings of an OCR generated collection.

8 Conclusion

Extending the accuracy experiment with SMART brought the interaction between OCR text and IR techniques into focus. We showed, in this experiment as well as in our previous works, that average precision was not heavily affected. But we established that cosine normalization is a potential problem for some recognized documents with particular features (i.e. documents which include tabular data, repetitive data, or poor hard copy). Although these problems can arise when OCR text is used directly with an IR system, we believe these complications were partially induced by the uniform methods used in our experimentation. We made no efforts to handle documents with special peculiarities (for example, by changing resolution or threshold at scan time or by zoning out tabular data).

We also tested a number of weighting algorithms available in SMART. Although too time consuming to analyze all of them in detail, we were able to evaluate `atc.atn` more fully. It was suggested that full-text documents in the TREC experiments worked best with `inc.ltn`. This scheme may be a better choice for full, OCR-text documents as well (Buckley, 1994).

Feedback was a new IR technique tested in this experiment. Although some improvement over the initial runs did take place for the OCR collection, continued improvement

through increased query expansion did not return certain difficult-to-retrieve documents. This implies that some form of special document processing may be necessary to retrieve documents with these characteristics. Also since we found that the cosine seemed to have a negative effect in our initial runs, it may have also had some effect in our feedback experiments. Experiments that altered document and query weights prior to running the feedback query could explain its effect (e.g. initial run `ltn.ltc`, feedback run `ltc.ltc`)(Buckley, Salton, & Allan, 1994).

The implications of using *unedited* OCR-generated text are significant. There are certainly thousands of planned tasks to create on-line collections. By eliminating or drastically reducing the human aspect of correction in these projects, costs and completion time can be reduced while maintaining quality. Testing the accuracy of the text generated by an OCR device though is only a small part of a bigger picture. There are other ways in which document collections can be exploited that we have not considered in our experimentation. Certain algorithms may not produce the same results on unedited OCR-generated text that they produced on clean text.

From our experience in working with OCR text, we are convinced that average precision is unaffected. Complications may arise though when more information is expected to be extracted from a recognized document collection; for example, in the case of passage retrieval, paragraphs or sections may need to be delimited. This is a fairly simple process in clean formatted text, but becomes more difficult when spacing and/or punctuation has been misrecognized by an OCR device. Even manual processing tasks could be less straightforward, for example, including hypertext links. As document processing becomes more sophisticated, these concerns will become more consequential and their significance should be considered.

9 Acknowledgment

We would like to thank Chris Buckley, Jamie Callan, Donna Harman, George Nagy, Gerard Salton and Ashok Singh for reading an earlier draft of this paper. We would also like to thank the anonymous referees for their thorough reading of this paper. Their comments and suggestions have greatly improved the quality of this work.

References

- Anderson, V. L., & McLean, R. A. (1974). *Design of Experiments: A Realistic Approach*. Marcel Dekker, Inc., New York.
- Buckley, C. (1994). Personal Communication..
- Buckley, C., Salton, G., & Allan, J. (1994). The Effect of Adding Relevance Information in a Relevance Feedback Environment. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 292–301.
- Callan, J. P., Croft, W. B., & Harding, S. M. (1992). The INQUERY Retrieval System. In *Proceedings of the Third International Conference on Database and Expert Systems Applications*, pp. 78–83.

- Croft, W. B., Harding, S., Taghva, K., & Borsack, J. (1994). An Evaluation of Information Retrieval Accuracy with Simulated OCR Output. In *Proc. 3rd Symposium on Document Analysis and Information Retrieval*, pp. 115–126 Las Vegas, NV.
- Frakes, W. B., & Baeza-Yates, R. (Eds.). (1992). *Information Retrieval, Data Structures & Algorithms*. Prentice Hall, Englewood Cliffs, NJ 07632.
- Futrelle, R. P., et al. (1991). Document Analysis, Understanding, and Knowledge Access. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pp. 101–111 St. Malo, France.
- Harman, D. (1988). Towards Interactive Query Expansion. In *Proceedings of the Eleventh Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval* Grenoble, France. ACM Press.
- Harman, D. (1992a). *Information Retrieval, Data Structures and Algorithms*, chap. Relevance Feedback and Other Query Modification Techniques, pp. 241–263. Prentice Hall, Englewood Cliffs, NJ 07632.
- Harman, D. (1992b). *Information Retrieval, Data Structures and Algorithms*, chap. Ranking Algorithms, pp. 363–392. Prentice Hall, Englewood Cliffs, NJ 07632.
- Harper, D. J. (1980). *Relevance Feedback in Document Retrieval Systems: An Evaluation of Probabilistic Strategies*. Ph.D. thesis, Jesus College, Cambridge, England.
- Ide, E. (1969). Relevance Feedback in an Automatic Document Retrieval System. Tech. rep. ISR-15, Cornell University.
- Nartker, T. A., Bradford, R. B., & Cerny, B. A. (1992). A Preliminary Report on UNLV/GT1: A Database for Ground-truth Testing in Document Analysis and Character Recognition. In *Proc. 1st Symp. on Document Analysis and Information Retrieval*, pp. 300–315 Las Vegas, NV.
- Rocchio, J. J. (1971). *The SMART Retrieval System*, chap. Relevance Feedback in Information Retrieval, pp. 313–323. Prentice Hall, Inc, Englewood Cliffs, NJ 07632.
- Salton, G. (1971). *The SMART Retrieval System, Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Englewood Cliffs, NJ.
- Salton, G., & Buckley, C. (1988). Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*, 24(5), 513–523.
- Salton, G., & Buckley, C. (1990). Improving Retrieval Performance by Relevance Feedback. *Journal of the American Society for Information Science*, 41(4), 288–297.
- Taghva, K., Borsack, J., Bullard, B., & Condit, A. (1993). Post-Editing through Approximation and Global Correction. Tech. rep. 93-05, Information Science Research Institute, University of Nevada, Las Vegas.
- Taghva, K., Borsack, J., & Condit, A. (1994a). An Expert System for Automatically Correcting OCR Output. In *Proc. IS&T/SPIE 1994 Intl. Symp. on Electronic Imaging Science and Technology*, pp. 270–278 San Jose, CA.

- Taghva, K., Borsack, J., & Condit, A. (1994b). Results of Applying Probabilistic IR to OCR Text. In *Proc. 17th Intl. ACM/SIGIR Conf. on Research and Development in Information Retrieval*, pp. 202–211 Dublin, Ireland.
- Taghva, K., Borsack, J., Condit, A., & Erva, S. (1994c). The Effects of Noisy Data on Text Retrieval. *J. American Soc. for Inf. Sci.*, 45(1), 50–58.