A Comparison of Text-Based Methods for Detecting Duplication in Scanned Document Databases*

Daniel P. Lopresti

Bell Labs, Lucent Technologies Inc.
600 Mountain Avenue, Room 2D-447
Murray Hill, NJ 07974, USA
dpl@research.bell-labs.com

November 18, 2002

Abstract

This paper presents an experimental evaluation of several text-based methods for detecting duplication in scanned document databases using uncorrected OCR output. This task is made challenging both by the wide range of degradations printed documents can suffer, and by conflicting interpretations of what it means to be a "duplicate." We report results for four sets of experiments exploring various aspects of the problem space. While the techniques studied are generally robust in the face of most types of OCR errors, there are nonetheless important differences which we identify and discuss in detail.

Keywords: *duplicate detection*, *approximate string matching*, *information retrieval*, *optical character recognition*, *document analysis*.

1 Introduction

An important problem facing large-scale legacy document conversion projects is determining whether duplicates already exist in the database when a new document arrives for processing. For example, one particular collection activity for the U.S. Government's Gulf War Declassification Project accumulated 564,000 pages, the majority of which (292,000 pages) were later found to be duplicates of documents already on hand [3]. Clearly, techniques for detecting duplicates could prove extremely valuable, both in terms of cost savings as well as deepening our understanding of the relationship between documents.

Although this task might seem straightforward at first glance, it becomes quite challenging when one considers the different possible interpretations of what it means to be a "duplicate," and the many types of damage that can be inflicted on a physical medium such as the printed page. Note also that while there is an obvious connection to the known-item

*Appears in *Information Retrieval*, 4(2): 153-173, July 2001.

searching problem, there are some essential differences between traditional information retrieval and duplicate detection. In the latter, queries arise as the result of the automated process of registering a new document with the collection. Hence, both the query and the database entries may contain a substantial number of recognition errors. Moreover, absent an extraordinary event (*i.e.*, detection of a potential duplicate), no human intervention is assumed. Indeed, the goal is to minimize operator involvement. This places severe demands on precision in particular (as usual, the importance of high recall varies with the application). Performance in cases when no match is present, a "negative" result, can be just as interesting as returning true hits, although much more difficult to quantify.

In previous papers [9, 10, 12], we introduced a formalism based on approximate string matching for categorizing duplicates into four classes as illustrated in Figure 1:

Full-layout duplicates are visually identical (*e.g.*, one is a photocopy or fax of the other).

Full-content duplicates have identical textual content, but not necessarily the same layout of text lines (*e.g.*, a Web page printed twice using different margin settings).

Partial-layout duplicates share significant content and have the same layout (*e.g.*, the copy-and-pasting of whole paragraphs from one document to another).

Partial-content duplicates share significant content but not necessarily the same layout (*e.g.*, the copy-and-pasting of individual sentences).

In addition, we described a family of algorithms, one for each of the four classes, which operate on the "raw" (*i.e.*, uncorrected) output of an optical character recognition (OCR) process. We presented exerimental results showing that these methods were robust in the presence of real-world noise, and that the models could successfully distinguish the various types of duplication.

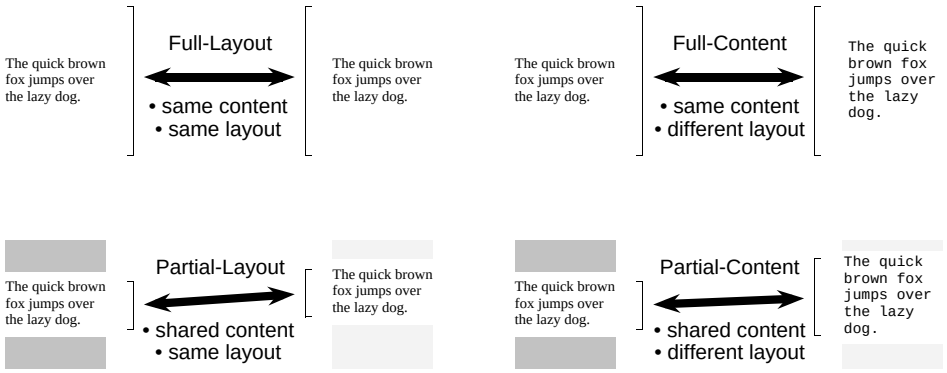

Figure 1: The four duplicate classes.

There are, of course, a number of other well-known approaches to searching textual databases. While schemes predicated on finding long strings of perfect similarity will likely fail when noisy documents are included in the corpus, the traditional vector space metric from information retrieval, for example, has been found to be relatively immune to OCR

errors [21]. Identifying plagiarism and copyright violations, where some degree of difference between the documents in question can naturally be expected, is also a related problem [17]. Such methods could be run on OCR output in an attempt to identify duplicates in document image databases.

The present paper makes use of this work and builds on our previous results in duplicate detection in several ways. We present an experimental evaluation comparing other text-based methods to the string matching techniques mentioned earlier. We also examine three new "hard" scenarios: cases where the queries are severely degraded, the reading order is incorrectly determined, and the database contains numerous false duplicates. These highlight the strengths and weaknesses of the different approaches.

The remainder of this paper is organized as follows. In Section 2, we briefly review related work. The algorithms to be studied are summarized in Section 3. Experimental results are presented in Section 4. Finally, Section 5 gives our conclusions and discusses possible future research.

2 Related Work

A number of researchers have begun to examine the problem of detecting duplicates in the context of document image databases [1, 4, 5, 8, 13, 14, 20]. Still, most previous work on this subject has concentrated on identifying which features to extract and not so much on the different ways they might be compared. This step is typically handled using one or another of the techniques from the literature.

Broadly speaking, these approaches can be classified depending on whether they operate on low-level image features [4, 5, 13, 14] or on the output of a symbolic recognition process such as OCR [1, 8, 20]. The former are more general in the sense they can be applied to non-textual input (*e.g.*, drawings, photographs), but more limiting in that they can only be used to find full-layout duplicates. Our primary interest lies in techniques from the latter category.

Spitz, for example, employs character shape codes as features and compares them using a standard string matching algorithm [20]. In the taxonomy presented in Section 1, this corresponds to the full-content problem. Doermann, et al., also use shape codes, but extract n-grams for a specific text line to index into a table of document pointers [1]. Since this signature is computed from a single line, it does not explicitly measure the similarity of complete pages. The intention, though, is that this is a method for addressing the full-layout problem. Hull, et al., describe three techniques: one based on decomposing the page into a grid and counting connected components within each cell, another using word lengths as a hash key, and one comparing image features (pass codes arising from fax compression) under a Hausdorff distance measure [5]. More details on the last method appear in later paper [4]. The first and third of these fall in the full-layout category, while the second can be classified as searching for full-content duplicates. In a recent paper, Lee and Hull describe a method for performing duplicate detection on symbolically compressed images by solving the text deciphering problem through the use of Hidden Markov Models [8]. They then apply n-gram indexing with term weighting to detect duplicates, addressing the

full-content problem.

Elsewhere, Taghva, et al., observed that traditional vector-space techniques from the field of information retrieval (IR) are for the most part unaffected by OCR errors when the input is relatively clean [21]. Kantor and Voorhees' report on the TREC-5 Confusion Track presents the results of running 49 queries ("topics") against three parallel collections (baseline, 5% OCR error rate, and 20% OCR error rate) derived from the *1994 Federal Register* [6, 7]. Five research groups participated, proposing equally-many different ranking strategies. Performance was found to suffer noticeably at the highest error rate (precision in some cases dropped by a factor of 10), and methods occasionally failed to return the intended document anywhere in the top $1,000$.

Also seemingly related is the general copy detection problem. Shivakumar and Garcia-Molina have developed efficient methods for searching large online databases for signs of copyright infringement [17]. A later paper of theirs considers the task of identifying near-replicas on the World Wide Web (WWW) to improve the performance of Web crawlers, archivers, and search engines [18]. All of these approaches are text-based, employing character or word n-grams or longer syntactic entities (sentences, paragraphs, etc.), and must allow for the fact that two documents need not be identical for the results of their comparison to qualify as "interesting." There are, however, significant differences between a typical IR query and a complete document, and between the kinds of errors that arise during OCR and the steps taken to conceal an attempt at plagiarism. One question we seek to answer in this paper is how well these methods work for detecting duplicates in the presence of large amounts of "noise" and under the different models described in the Introduction.

3 Algorithms

In this section, we briefly summarize the algorithms under study. The reader is referred to the original works in question for more details.

3.1 Approximate String Matching Applied to Duplicate Detection

The string measures are based on the well-known concept of *edit distance*. In the simplest case, the following three operations are permitted: (1) delete a symbol, (2) insert a symbol, (3) substitute one symbol for another. Each of these is assigned a cost, c_{del}, c_{ins}, and c_{sub}, and the edit distance is defined as the minimum cost of any sequence of basic operations that transforms one string into the other.

As it relates to full-content duplicates, this optimization problem can be solved using a dynamic programming algorithm [22]. Let $Q = q_1 q_2 \ldots q_m$ be the query document, $D = d_1 d_2 \ldots d_n$ be the database document, and define $dist1_{i,j}$ to be the distance between the first i symbols of Q and the first j symbols of D. The initial conditions are:

$$\begin{array}{lcll} dist1_{0,0} & = & 0 & \\ dist1_{i,0} & = & dist1_{i-1,0} + c_{del}(q_i) & 1 \le i \le m,\ 1 \le j \le n \\ dist1_{0,j} & = & dist1_{0,j-1} + c_{ins}(d_j) & \end{array} \qquad (1)$$

and the main dynamic programming recurrence is:

$$dist1_{i,j} = \min \begin{cases} dist1_{i-1,j} & + & c_{del}(q_i) \\ dist1_{i,j-1} & + & c_{ins}(d_j) \\ dist1_{i-1,j-1} & + & c_{sub}(q_i, d_j) \end{cases} \qquad 1 \le i \le m,\ 1 \le j \le n \qquad (2)$$

The computation builds a matrix of distance values working from the upper left corner ($dist1_{0,0}$) to the lower right ($dist1_{m,n}$).

The other three string algorithms, *sdist1* for partial-content duplicates, *dist2* for full-layout duplicates, and *sdist2* for partial-layout duplicates, reflect various adaptations of this approach to allow for partial matchings and 2-D document structure [9, 10, 12].

For the partial duplicate problem, what is needed is the best match between any two substrings of Q and D. Fortunately, the original computation can be modified so that shorter regions of similarity can be detected in two longer documents with no increase in time complexity [19]. The edit distance is made 0 along the first row and column of the matrix, so the initial conditions become:

$$sdist1_{0,0} = sdist1_{i,0} = sdist1_{0,j} = 0 \qquad 1 \le i \le m,\ 1 \le j \le n \qquad (3)$$

In addition, another term is added to the inner-loop recurrence capping the maximum distance at any cell to be 0. This has the effect of allowing a match to begin at any position between the two strings. The recurrence is:

$$sdist1_{i,j} = \min \begin{cases} 0 & & \\ sdist1_{i-1,j} & + & c_{del}(q_i) \\ sdist1_{i,j-1} & + & c_{ins}(d_j) \\ sdist1_{i-1,j-1} & + & c_{sub}(q_i, d_j) \end{cases} \qquad 1 \le i \le m,\ 1 \le j \le n \qquad (4)$$

Finally, the resulting distance matrix is searched for its smallest value. This reflects the end-point of the best substring match. The starting point can be found by tracing back the sequence of optimal editing decisions.

For the 2-D models (*i.e.*, layout duplicates), another level is added to the optimization. The problem is still one of editing, but at the higher level the basic entities are now strings (lines). Say that $Q = Q^1 Q^2 \ldots Q^k$ and $D = D^1 D^2 \ldots D^l$, where each Q^i and D^j is itself a string. For full-layout duplicates, the inner-loop recurrence takes the same general form as the 1-D case:

$$dist2_{i,j} = \min \begin{cases} dist2_{i-1,j} & + & C_{del}(Q^i) \\ dist2_{i,j-1} & + & C_{ins}(D^j) \\ dist2_{i-1,j-1} & + & C_{sub}(Q^i, D^j) \end{cases} \qquad 1 \le i \le k,\ 1 \le j \le l \qquad (5)$$

where C_{del}, C_{ins}, and C_{sub} are the costs of deleting, inserting, and substituting whole lines, respectively. The initial conditions are defined analogously to Equation 1.

Since the basic editing operations now involve full strings, it is natural to define the new costs as:

$$\begin{array}{lcl} C_{del}(Q^i) & \equiv & dist1(Q^i, \phi) \\ C_{ins}(D^j) & \equiv & dist1(\phi, D^j) \\ C_{sub}(Q^i, D^j) & \equiv & dist1(Q^i, D^j) \end{array} \qquad 1 \le i \le k,\ 1 \le j \le l \qquad (6)$$

where ϕ is the null string. Hence, the 2-D computation is defined in terms of the 1-D computation.

Lastly, the extension for partial-layout duplicates combines the modifications for the partial (Equation 4) and layout (Equation 5) problems:

$$sdist2_{i,j} = \min \begin{cases} 0 & \\ sdist2_{i-1,j} & + C_{del}(Q^i) \\ sdist2_{i,j-1} & + C_{ins}(D^j) \\ sdist2_{i-1,j-1} & + C_{sub}(Q^i, D^j) \end{cases} \qquad 1 \le i \le k,\ 1 \le j \le l \qquad (7)$$

Note that C_{del}, C_{ins}, and C_{sub} are defined as before in terms of *dist1* (*i.e.*, Equation 6).

At this point four different algorithms have been presented, one for each of the models described in the Introduction. For exact duplicates, the distance returned by any of the algorithms will either be 0 or a negative number that grows smaller as the lengths of the documents increase. For dissimilar documents, the maximum distance grows larger as the lengths increase. It is always the case that, for a given query, a smaller distance corresponds to a better match. In order for the results for different queries to be comparable, however, it is necessary to normalize the distances.

If the target interval is $[0, 1]$, where 0 represents a perfect match and 1 a complete mismatch, then the following formula provides an appropriate mapping:

$$EDITdist(Q, D) = \frac{dist - mindist}{maxdist - mindist} \qquad (8)$$

where *dist* is one of the four edit distance measures and *mindist* and *maxdist* are, respectively, the minimum and maximum possible distances for the comparison in question.

Assuming a full-duplicate computation, and making certain reasonable assumptions about the cost functions, the minimum is obtained when all of the characters in the query match the database document and there are no extra, unmatched characters. If the query is $Q = q_1 q_2 \ldots q_m$, then:

$$mindist = \sum_{i=1}^{m} c_{sub}(q_i, q_i) \qquad (9)$$

Or, more simply, $mindist = m \cdot c_{mat}$ when the costs are constant and c_{mat} is the cost of an exact match.

The maximum distance, on the other hand, is determined by the query and the set of all strings with the same length as the database document. If the cost functions are unconstrained, this in itself becomes an optimization problem. Fortunately, for constant costs there is a simple closed-form solution. Without loss of generality, let the query be the shorter of the two strings (*i.e.*, $m \le n$). There are two possible "worst-case" scenarios: either all of the symbols of the query are substituted and the remaining symbols of the database string are inserted, or all of the query symbols are deleted and the entire database string is inserted. Thus:

$$maxdist = \min \begin{cases} m \cdot c_{sub} + (n - m) \cdot c_{ins} \\ m \cdot c_{del} + n \cdot c_{ins} \end{cases} \qquad (10)$$

The partial-duplicate computations are normalized similarly.

3.2 The Vector Space IR Method

The vector space model first proposed by Salton, et al., is extremely popular in the field of information retrieval [15, 16]. This approach assigns large weights to terms that occur frequently in a given document but rarely in others because such terms are able to distinguish the document in question from the rest of the database. Let tf_{ik} be the frequency of term t_k in document D_i, n_k be the number of documents containing term t_k, T be the total number of terms, and N be the size of the database. Then a common weighting scheme ($tf \times idf$) defines w_{ik}, the weight of term t_k in document D_i, to be:

$$w_{ik} = \frac{tf_{ik} \cdot log(N/n_k)}{\sqrt{\sum_{j=1}^{T}(tf_{ij})^2 \cdot (log(N/n_j))^2}} \tag{11}$$

The summation in the denominator normalizes the length of the vector so that all documents have an equal chance of being retrieved.

Given query and document vectors $Q_i = (w_{i1}, w_{i2}, \ldots, w_{iT})$ and $D_j = (w_{j1}, w_{j2}, \ldots, w_{jT})$, a vector dot product is computed to quantify the similarity between the two:

$$VSdist(Q_i, D_j) = \sum_{k=1}^{T} w_{ik} w_{jk} \tag{12}$$

3.3 The SCAM Algorithm for Plagiarism Detection

SCAM, the work of Shivakumar and Garcia-Molina, attempts to overcome some of the limitations of the vector space model when the lengths of the documents differ significantly [17]. They define a *closeness set* for two documents Q_i and D_j, $c(Q_i, D_j)$, to contain those terms that have a similar number of occurrences in both documents. That is, a term t_k is in $c(Q_i, D_j)$ if it satisfies the following condition:

$$\epsilon - \left(\frac{tf_{ik}}{tf_{jk}} + \frac{tf_{jk}}{tf_{ik}} \right) > 0 \tag{13}$$

where $\epsilon = (2^+, \infty)$ is a user-settable parameter.

Next, they define an asymmetric measure:

$$subset(Q_i, D_j) = \frac{\sum_{t_k \in c(Q_i, D_j)} \alpha_k^2 \cdot tf_{ik} \cdot tf_{jk}}{\sum_{k=1}^{T} \alpha_k^2 \cdot (tf_{ik})^2} \tag{14}$$

that reflects the degree to which Q_i is contained within (*i.e.*, is a partial-content duplicate of) D_j. The similarity between two documents is then computed to be:

$$SCAMdist(Q_i, D_j) = \max\{subset(Q_i, D_j), subset(D_j, Q_i)\} \tag{15}$$

4 Experimental Results

To investigate the performance of the algorithms described in this paper, four sets of experiments were designed. The first studied various real-world noise sources and their effects, the second examined a common document layout analysis error (decolumnization), the third looked at the four duplicate models and how they relate, and the last considered the impact of including multiple false duplicates in the database.

The base test collection consisted of 1,000 professionally written news articles gathered from Usenet and was used as-is (*i.e.*, no attempt was made to inject OCR errors, either real or synthetic). The query documents, however, and the intended duplicates were "authentic," pages that had been printed, scanned, and OCR'ed.

We analyzed the four string algorithms, the vector space measure using both word unigram tokens (with stopword removal) and character trigram tokens (without stopword removal), and SCAM. All of the algorithms were coded in C and run on an SGI O2 workstation. To make the results directly comparable, the outputs were normalized to the interval $[0, 1]$, with 0 corresponding to a perfect match (this entailed subtracting the vector space results from 1). For the full-duplicate computations, the edit costs were set to be $c_{del} = c_{ins} = c_{sub} = 1$ and $c_{mat} = 0$. For the partial-duplicate computations, the match cost was changed to $c_{mat} = -1$. In our experiments using SCAM, we set $\alpha = 1$ and $\epsilon = 3$.

4.1 Experiment 1

The goal of this experiment was to study duplicate detection under a range of realistic noise conditions. Working from the corpus, 10 documents were randomly chosen to provide a focus. The minimum document length in this set was 364 characters (65 words), the maximum was 2,404 characters (379 words), and the average was 1,274 characters (200 words). Seven "batches" of these 10 pages were then created, the first six to be inserted into the database as the duplicates, and the remaining batch to serve as the queries. For the duplicates, one set of pages was used as-is (*i.e.*, "clean") while the others were subjected to one of five different degradations: faxing, noticeably light or dark or third generation photocopying, or handwritten markup (annotations) that obscured a random 20% of the lines on the page. In addition, the original ASCII text for the documents was left in the database. The query batch was degraded twice: first the pages were photocopied light, then the light copies were faxed. All of the resulting pages were then scanned and OCR'ed. Hence, each of 10 queries was run against a database of 1,000 documents containing seven intended duplicates, six that had been OCR'ed plus one error-free version.

Except for the few lines that had been obliterated by the marker pen, all of the documents were still easily legible to a human reader. Optical character recognition systems are not yet as adept, however. Table 1 shows the OCR accuracies for the duplicates, and Table 2 the accuracies for the queries.[1] Note that the rates range widely, dropping as low as 56.3% in the former case and 45.1% in the latter. Faxing and annotation were the worst offenders,

[1]These figures are computed by using string edit distance (*e.g.*, algorithm *dist1*) to compare the OCR output to the original ASCII text for the document in question. Provided there have been no higher-level errors in the document analysis process, this is an accepted way of quantifying OCR accuracy [2].

at least in terms of impairing accuracy, while third generation photocopying displays the highest variance. The queries clearly show the consequences of having been twice-degraded. As expected, a large variety of OCR errors were encountered. Beyond this, other complications arose as well. For example, the standard headers prepended to faxes were transcribed (albeit with numerous mistakes), and the lines that had been crossed-out were completely missing from the annotated pages.

Table 1: OCR accuracies for the duplicates used in Experiment 1.

		Degradation							
Document	Clean	Fax	3rdGen	Light	Dark	Note	Min	Max	Ave
737	93.9%	74.9%	92.1%	83.0%	92.6%	56.3%	56.3%	93.9%	82.1%
8161	95.7%	83.8%	84.9%	84.8%	95.4%	73.0%	73.0%	95.7%	86.3%
9837	96.7%	67.9%	81.7%	84.5%	96.6%	74.7%	67.9%	96.7%	83.7%
9877	96.2%	62.0%	71.0%	78.8%	96.1%	80.0%	62.0%	96.2%	80.7%
15233	96.0%	88.4%	70.9%	85.5%	95.9%	76.1%	70.9%	96.0%	85.5%
15317	96.3%	69.9%	93.2%	89.8%	96.1%	79.1%	69.9%	96.3%	87.4%
15334	96.3%	83.6%	63.4%	80.7%	96.1%	76.3%	63.4%	96.3%	82.7%
16697	96.5%	65.5%	83.7%	86.0%	96.2%	83.3%	65.5%	96.5%	85.2%
16884	95.1%	80.7%	78.3%	86.2%	95.1%	72.5%	72.5%	95.1%	84.7%
19962	95.3%	60.7%	88.3%	86.9%	94.9%	77.2%	60.7%	95.3%	83.9%
Min	93.9%	60.7%	63.4%	78.8%	92.6%	56.3%	56.3%	93.9%	
Max	96.7%	88.4%	93.2%	89.8%	96.6%	83.3%	83.3%	96.7%	
Ave	95.8%	73.7%	80.7%	84.6%	95.5%	74.9%			84.2%

Table 2: OCR accuracies for the queries used in Experiment 1.

Document	Light-Fax
737	58.4%
8161	65.0%
9837	54.2%
9877	63.6%
15233	65.0%
15317	75.7%
15334	77.0%
16697	57.5%
16884	53.7%
19962	45.1%
Min	45.1%
Max	77.0%
Ave	61.5%

We then ran the *dist2* approximate string matching algorithm, as well as the vector space measure using single-word tokens and character trigrams and the SCAM method. The average precision at 100% recall for each of the 10 queries is given in Table 3. As can be seen, all of the methods performed quite well; *dist2* was perfect, and the other measures

nearly so. This demonstrates that, on the whole, the four techniques are robust when faced with the sorts of OCR errors seen in practice. The one anomaly, perhaps, is the low score obtained by SCAM for query `19962`. This particular document was relatively short: a listing of worldwide gold markets and the current prices per ounce at the time. SCAM ranked similar listings for other days higher than several of the true matches. The issue of near-duplicates will be examined further in subsection 4.4.

Table 3: Average precision at 100% recall for Experiment 1.

	Measure			
		Vector Space		
Document	*dist2*	Word Unigram	Char Trigram	SCAM
`737`	1.000	1.000	1.000	1.000
`8161`	1.000	1.000	1.000	1.000
`9837`	1.000	1.000	1.000	1.000
`9877`	1.000	1.000	1.000	1.000
`15233`	1.000	0.778	0.778	1.000
`15317`	1.000	1.000	1.000	1.000
`15334`	1.000	0.875	0.875	1.000
`16697`	1.000	1.000	1.000	1.000
`16884`	1.000	1.000	0.778	1.000
`19962`	1.000	0.875	0.875	0.538
Ave	1.000	0.953	0.931	0.954

Figure 2 plots, for the 10 queries, the normalized distances computed by each of the four approaches for every document in the database. Note that there is usually a clear distinction between true duplicates and everything else, although this appears more evident in the case of *dist2* than, say, SCAM. This suggests that it may be easier to set a predetermined threshold for the former.

Qualitatively, certain other interesting effects can be seen in the charts. The various forms of degradation seem to be handled somewhat differently by the four methods. For example, *dist2* does worse than the others on the annotated duplicates. Recall that approximately 20% of the content of these documents was completely obscured. The vector space measures (including SCAM) are tuned to tolerate this kind of scenario. To be fair, however, the *dist2* algorithm is not designed to capture partial matches; *sdist2* or *sdist1* would be a more appropriate choice here. On the other hand, the faxed, third generation, and light duplicates look to be more troublesome for the vector space techniques. Many of the non-duplicates (the hollow circles) that are assigned relatively low distance scores correspond to follow-up postings to the original article used as the query. While sharing some of the same unique terminology (*e.g.*, proper names), they are by no means true duplicates.

It is perhaps instructive to examine more closely the best and worst cases. Table 4 lists the minimum-distance duplicate for each measure/query combination, and Table 5 the maximum-distance duplicate. Note that the annotated documents often turn up on the former list for the vector space methods (especially SCAM), and the latter list for *dist2*. It should come as no surprise that the error-free and clean duplicates are generally ranked

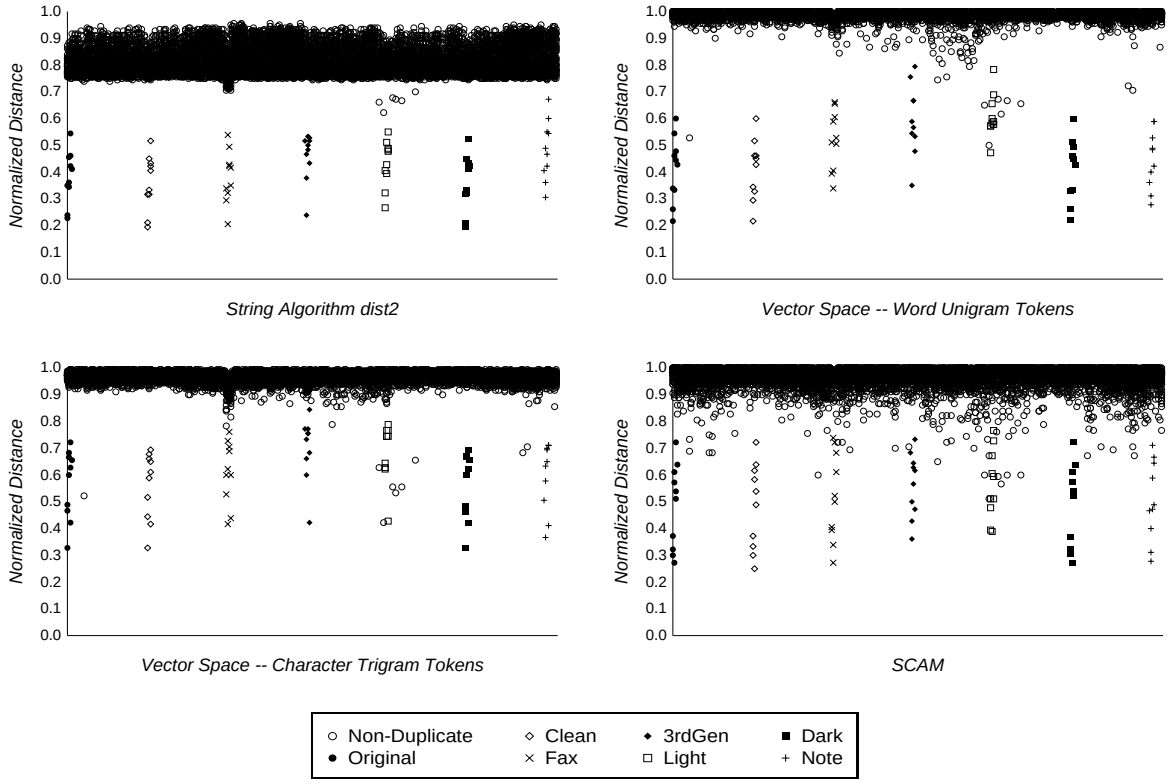

Figure 2: Results for Experiment 1.

near the top. The third generation photocopies provide an obvious challenge across all of the measures. Upon closer examination, it was determined that the machine used to generate the database copies introduced significant optical distortion, which no doubt had an adverse impact on OCR and ultimately the distance values computed during retrieval.

4.2 Experiment 2

A fundamental difference between the string-based models and those derived from the vector space approach is the dependence of the former on determining a canonical reading order for the text. (This reading order need not conform to the way a human would interpret the text, it need only be consistent from one document to the next.) While finding a reading order is generally not hard for single column documents, it can become challenging for complex, multicolumn layouts.

The intent of this experiment was to examine duplicate detection when reading order is a problem. We took query `15233` from the previous experiment and formatted it in two-column mode, forcing the inter-column spacing to be so tight that the column break was lost on some, but not all, lines (this is known as "decolumnization"). The page was photocopied light and then scanned and OCR'ed. In this case it was impossible to compute a value for the OCR accuracy, but visual inspection suggested it was somewhat higher than the queries in Experiment 1. Figure 3 gives the results for this test.

Table 4: Minimum-distance duplicates for Experiment 1.

	Measure			
		Vector Space		
Document	*dist2*	Word Unigram	Char Trigram	SCAM
737	Fax (0.354)	Original (0.430)	Fax (0.604)	3rdGen (0.617)
8161	Clean (0.322)	Clean (0.334)	Clean (0.595)	Clean (0.254)
9837	Clean (0.451)	Original (0.461)	Clean (0.666)	Original (0.614)
9877	Dark (0.334)	Dark (0.512)	Dark (0.676)	Note (0.473)
15233	Fax (0.295)	Dark (0.331)	Clean (0.447)	Dark (0.368)
15317	Clean (0.215)	Original (0.262)	Dark (0.489)	Note (0.313)
15334	Clean (0.199)	Original (0.218)	Dark (0.331)	Note (0.278)
16697	Clean (0.410)	Original (0.449)	Clean (0.613)	Fax (0.527)
16884	Fax (0.431)	Note (0.589)	Fax (0.694)	Note (0.667)
19962	Fax (0.418)	Note (0.424)	Note (0.414)	3rdGen (0.476)

Table 5: Maximum-distance duplicates for Experiment 1.

	Measure			
		Vector Space		
Document	*dist2*	Word Unigram	Char Trigram	SCAM
737	Note (0.677)	Light (0.689)	Light (0.768)	Light (0.729)
8161	Note (0.490)	Light (0.586)	Light (0.771)	3rdGen (0.429)
9837	Note (0.551)	Light (0.602)	3rdGen (0.756)	Fax (0.741)
9877	Fax (0.495)	3rdGen (0.670)	3rdGen (0.777)	Light (0.609)
15233	3rdGen (0.522)	3rdGen (0.759)	3rdGen (0.775)	3rdGen (0.684)
15317	Fax (0.344)	Fax (0.513)	Light (0.626)	Fax (0.410)
15334	3rdGen (0.471)	3rdGen (0.592)	3rdGen (0.665)	3rdGen (0.500)
16697	3rdGen (0.500)	3rdGen (0.671)	3rdGen (0.760)	3rdGen (0.631)
16884	Note (0.549)	3rdGen (0.795)	3rdGen (0.846)	Light (0.771)
19962	Note (0.604)	Light (0.594)	Fax (0.442)	Fax (0.616)

As expected, the string technique nearly loses its ability to distinguish duplicates from non-duplicates. This is the penalty paid for not maintaining the same reading order. The vector space charts, on the other hand, look more like those for the first experiment. Hence, they are immune to this sort of failure in the document analysis process.

4.3 Experiment 3

The purpose of this experiment was to determine how the different duplicate models and comparison measures relate empirically (recall Figure 1). The same source document was used as in the previous experiment. Duplicates were constructed from the query by changing the line breaks and/or deleting roughly half of the text from the beginning of the document and appending an equal amount of unrelated text to the end.

The pages were then printed, scanned, and OCR'ed. In this case, the OCR accuracies were all fairly close, ranging from 94.9% to 96.1%, as indicated in Table 6. As before, the

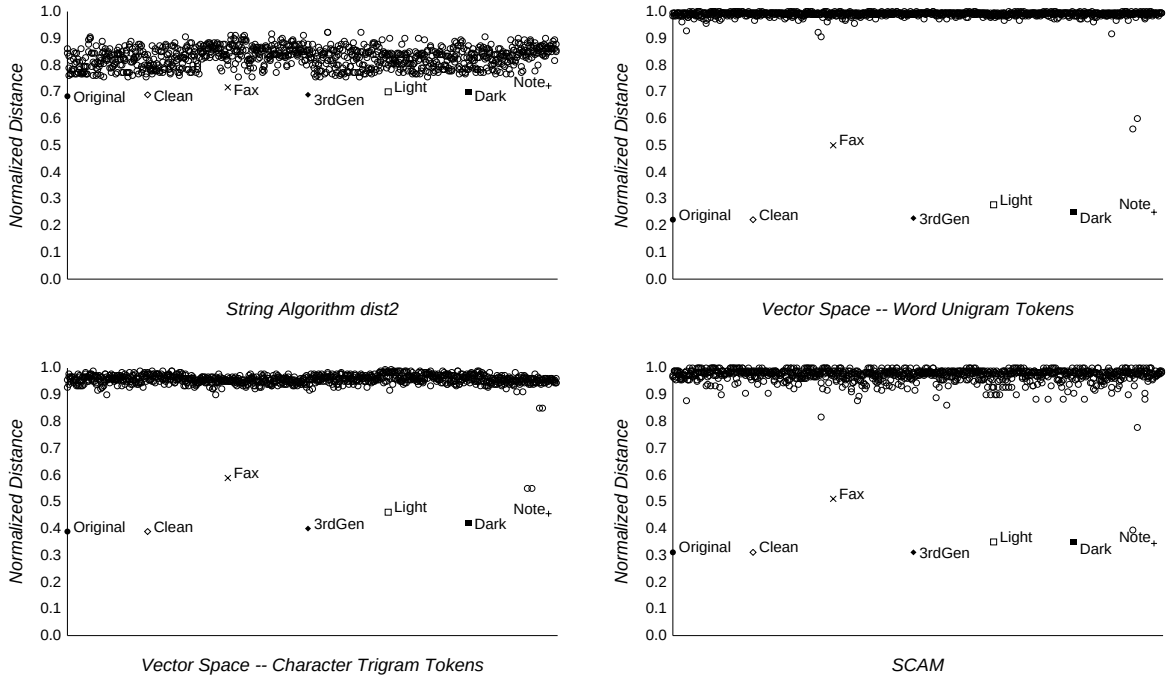

Figure 3: Results for Experiment 2 (two-column query with improper reading order).

original source text was left in the database to serve as a second full-layout duplicate of the query. Thus, there were between two and five duplicates in the database, depending on the model.

Table 6: OCR accuracies for Experiment 3.

Document Type	Duplicate	Query
Full-Layout	96.0%	95.9%
Full-Content	96.1%	–
Partial-Layout	94.9%	–
Partial-Content	96.0%	–

The results for this experiment are shown in Figure 4 for the string techniques, and in Figure 5 for the vector space and SCAM methods. Note that various string matching algorithms are capable of distinguishing different types of duplicates, while the vector space and SCAM measures are all quite similar, producing results most like the *sdist1* algorithm. This suggests that, in certain applications, the string-based approaches may yield higher precision (*e.g.*, locating only duplicates that are photocopies of the document in question and not other types also present in the database).

4.4 Experiment 4

While the vector space and SCAM techniques are more robust with respect to variations in reading order (as demonstrated in Experiment 2), this same attribute could prove to be a

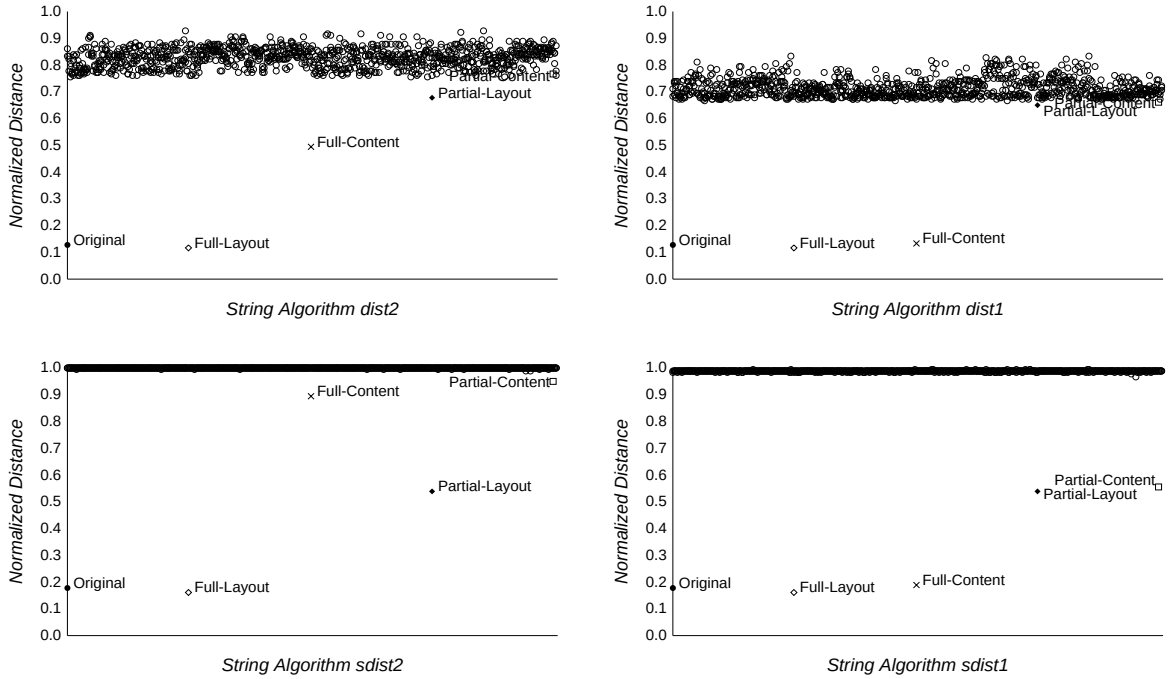

Figure 4: Approximate string matching results for Experiment 3 (the duplicate models).

disadvantage when the precise ordering of terms is crucial to identifying true duplicates.

The goal of this final experiment was to study duplicate detection using a database that also contained a number of false (*i.e.*, near) duplicates. We injected 49 issues of a daily electronic newsletter into the same database of 1,000 documents used in our previous tests. While the content of the newsletter varies significantly from day to day, most issues share common section headings as well as company- and person-specific references (*e.g.*, the names of executives, products, partners, competitors). One issue was selected for use as the query, printed, photocopied light, and then scanned and OCR'ed. The OCR accuracy for the query was estimated to be 74.2%. All of the documents in the database were "perfect" (*i.e.*, electronic) text, including both the intended match and the 48 false duplicates. Figure 6 presents the results for this experiment.

As can be seen in the charts, the vector space methods and SCAM produce less separation between the intended duplicate and the remainder of the documents than does the *dist2* string algorithm. Indeed, the presence of so many false duplicates significantly impairs the ability of the character trigram implementation to rank the true match appropriately. This confirms that the performance of a given technique depends not only on the condition of the query and the duplicates in the database, but on the existence of potential near (but otherwise uninteresting) matches as well.

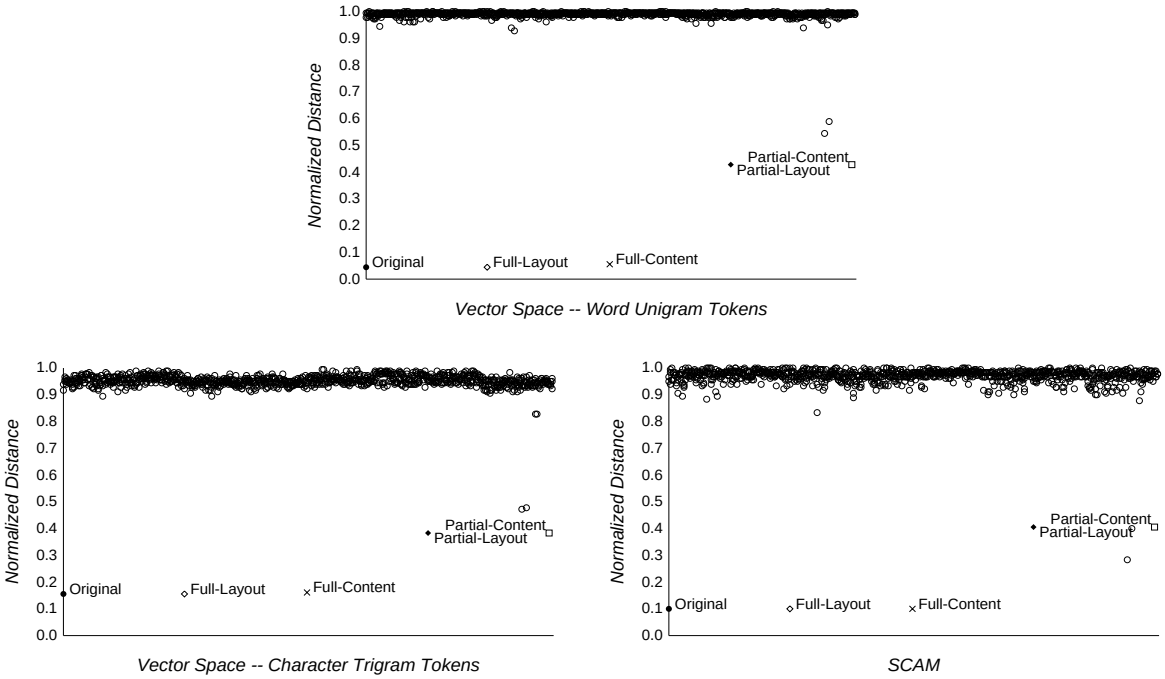

Figure 5: Vector space and SCAM results for Experiment 3 (the duplicate models).

5 Conclusions

In this paper we have presented an experimental evaluation of several text-based methods for detecting duplication in document image databases. All were tested using uncorrected OCR output for documents that had been subjected to a variety of real-world degradations. While the techniques under study are generally robust in the face of most types of OCR errors, there are nonetheless important differences, as we have shown. It appears likely that the most effective, efficient solution to this problem will be to combine several of the methods discussed as well as perhaps approaches based on lower-level image features not considered in the present paper.

Table 7 summarizes the algorithms once again. Here a solid bullet (•) indicates the broadest class for which a given method will work, while a hollow bullet (○) indicates more restricted kinds of duplicates it will also locate. Since some of the models subsume others, an obvious question is "Why bother with the less general ones?" The answer lies in increased precision for those situations where admitting a larger class of duplicates is undesirable (*e.g.*, when the targeted duplicates are known to be photocopies). Special cases may also make it possible to develop more efficient algorithms.

Possible topics for future research include examining the resource requirements (time, space) for each of the implementations since this may prove more critical than small differences in ranking ability. The test collection used in the current study was small, and the range of document analysis errors, while suggestive, was by no means comprehensive. Hence, the question of handling much larger, more realistic databases is open and will undoubtedly

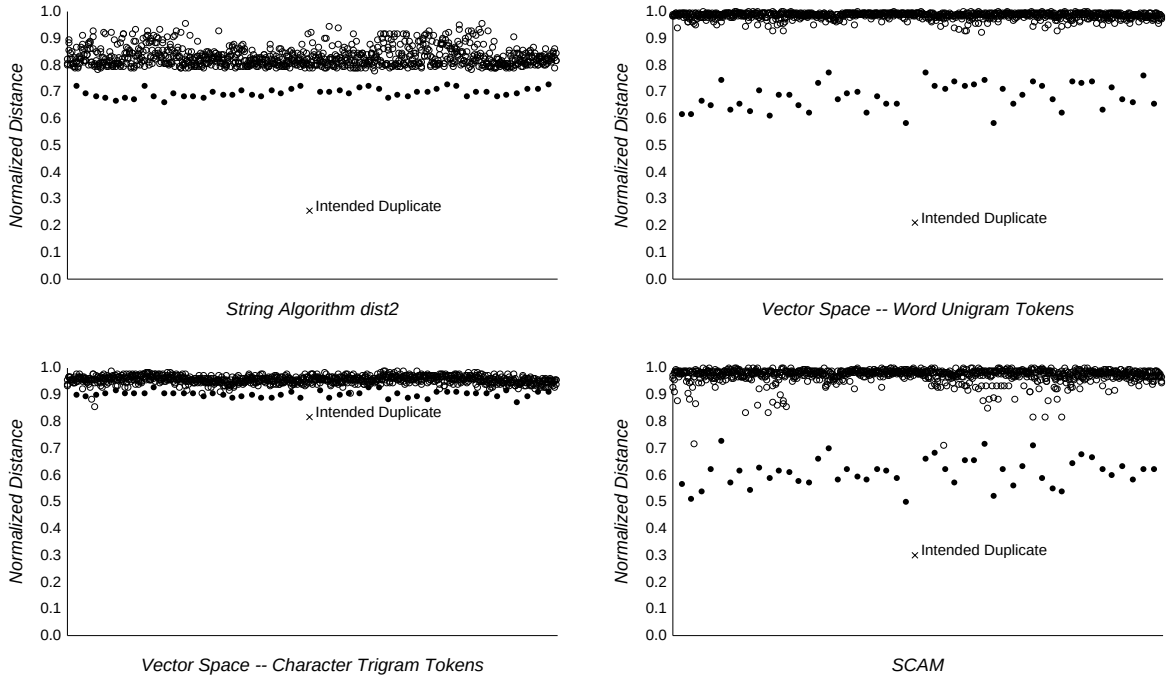

Figure 6: Results for Experiment 4 (database containing numerous false duplicates).

raise new issues as well as further opportunities for model and algorithm development.

Finally, as noted earlier a "negative" result, determining that no duplicate exists in the database, may be just as important for this application as the more common kind of query analysis presented in this paper. Establishing a methodology for measuring this aspect of performance would be beneficial.

6 Acknowledgements

An earlier version of this paper was presented at *Document Recognition and Retrieval VII (IS&T/SPIE Electronic Imaging 2000)* [11]. The author would like to thank the anonymous reviewers for their helpful suggestions. The trademarks mentioned in this paper are the properties of their respective companies.

References

[1] D. Doermann, H. Li, and O. Kia. The detection of duplicates in document image databases. In *Proceedings of the Fourth International Conference on Document Analysis and Recognition*, pages 314–318, Ulm, Germany, August 1997.

[2] J. Esakov, D. P. Lopresti, J. S. Sandberg, and J. Zhou. Issues in automatic OCR error classification. In *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*, pages 401–412, Las Vegas, NV, April 1994.

Table 7: Comparison of the methods studied in this paper.

Measure	Duplicate Type			
	Full-Layout	Full-Content	Partial-Layout	Partial-Content
Edit Distance				
dist2	●			
dist1	○	●		
sdist2	○		●	
sdist1	○	○	○	●
Vector Space				
Word Unigram	○	○	○	●
Char Trigram	○	○	○	●
SCAM	○	○	○	●

[3] G. G. Gilmore. Former Army operations officers assist DoD's search. *Army Link News*, December 1997.
http://www.dtic.mil/armylink/news/Dec1997/a19971209moredata.html.

[4] J. J. Hull. Document image similarity and equivalence detection. *International Journal on Document Analysis and Recognition*, 1(1):37–42, February 1998.

[5] J. J. Hull, J. Cullen, and M. Peairs. Document image matching and retrieval techniques. In *Proceedings of the Symposium on Document Image Understanding Technology*, pages 31–35, Annapolis, MD, April-May 1997.

[6] P. B. Kantor and E. Voorhees. Report on the TREC-5 confusion track. In *NIST Special Publication 500-238: The Fifth Text Retrieval Conference (TREC-5)*, pages 65–74, 1996.
http://trec.nist.gov/pubs/trec5/t5_proceedings.html.

[7] P. B. Kantor and E. M. Voorhees. The TREC-5 confusion track: Comparing retrieval methods for scanned text. *Information Retrieval*, 2(2/3):165–176, May 2000.

[8] D.-S. Lee and J. J. Hull. Duplicate detection for symbolically compressed documents. In *Proceedings of the Fifth International Conference on Document Analysis and Recognition*, pages 305–308, Bangalore, India, September 1999.

[9] D. P. Lopresti. Models and algorithms for duplicate document detection. In *Proceedings of the Fifth International Conference on Document Analysis and Recognition*, pages 297–300, Bangalore, India, September 1999.

[10] D. P. Lopresti. String techniques for duplicate document detection. In *Proceedings of the Symposium on Document Image Understanding Technology*, pages 101–112, Annapolis, MD, April 1999.

[11] D. P. Lopresti. A comparison of text-based methods for detecting duplication in document image databases. In *Proceedings of Document Recognition and Retrieval VII*

(IS&T/SPIE Electronic Imaging), volume 3967, pages 210–221, San Jose, CA, January 2000.

[12] D. P. Lopresti. String techniques for detecting duplicates in document databases. *International Journal on Document Analysis and Recognition*, 2(4):186–199, June 2000.

[13] F. Prokoski. Database partitioning and duplicate document detection based on optical correlation. In *Proceedings of the Symposium on Document Image Understanding Technology*, pages 86–97, Annapolis, MD, April 1999.

[14] R. Rogers, V. Chalana, G. Marchisio, T. Nguyen, and A. Bruce. Duplicate document detection in DocBrowse. In *Proceedings of the Symposium on Document Image Understanding Technology*, pages 119–127, Annapolis, MD, April 1999.

[15] G. Salton and J. Allan. Text retrieval using the vector processing model. In *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*, pages 9–22, Las Vegas, NV, April 1994.

[16] G. Salton, A. Wong, and C. Yang. A vector space model for information retrieval. *Communications of the Association for Computing Machinery*, 18(11):613–620, November 1975.

[17] N. Shivakumar and H. Garcia-Molina. SCAM: A copy detection mechanism for digital documents. In *Proceedings of the Second International Conference on Theory and Practice of Digital Libraries*, Austin, TX, 1995.
http://www.csdl.tamu.edu/DL95/papers/shivakumar.ps.

[18] N. Shivakumar and H. Garcia-Molina. Finding near-replicas of documents on the Web. In *Proceedings of the Workshop on Web Databases*, March 1998.

[19] T. F. Smith and M. S. Waterman. Identification of common molecular sequences. *Journal of Molecular Biology*, 147:195–197, 1981.

[20] A. L. Spitz. Duplicate document detection. In *Proceedings of Document Recognition IV (IS&T/SPIE Electronic Imaging)*, pages 88–94, San Jose, CA, February 1997.

[21] K. Taghva, J. Borsack, A. Condit, and P. Inaparthy. Effects of OCR errors on short documents. In *Annual Report of UNLV Information Science Research Institute*, pages 99–105, Las Vegas, NV, 1995.

[22] R. A. Wagner and M. J. Fischer. The string-to-string correction problem. *Journal of the Association for Computing Machinery*, 21:168–173, 1974.