

The UNLV-ISRI Document Collection for Research in OCR and Information Retrieval

Kazem Taghva*, Tom Nartker, Julie Borsack, and Allen Condit
Technical Report 99-01

Information Science Research Institute
University of Nevada, Las Vegas

October 1999

Abstract

We report on the UNLV-ISRI document collection history, composition, and characteristics. We further provide a short summary of research projects that were conducted using subsets of this collection. These projects were designed to address the retrieval effectiveness from OCR generated collections. Along with this report, ISRI is making this collection available to researchers for further study on the topic of OCR and Information Retrieval.

*Email: taghva@cs.unlv.edu

1 Background

The Information Science Research Institute (ISRI) has been involved with research of the interaction between optical character recognition (OCR) and information retrieval (IR) since 1989. This research has focused on issues associated with the construction of a large document database containing over 20 million pages of scientific, legal, and official memoranda. This collection will be used for online legal discoveries by the Department of Energy (DOE), its contractors, and other interested parties. Many, if not most of the documents to be included in this collection, will be obtained in their hard copy form. Therefore, a prototype conversion process was studied on a small representative subset to determine the capture and retrieval methodology. This study produced the Licensing Support System (LSS) Prototype consisting of approximately 2,600 documents (104,000 pages). A copy of this prototype database was given to ISRI for further experimental research related to LSS. Unfortunately, for various reasons, the complete prototype collection of hardcopy, images, and typed ASCII was not recoverable. Of this set, 1056 documents were identified as potentially useful by ISRI (see Section 3 below). These documents comprise the UNLV-ISRI Collection (the collection) that ISRI has used for many experiments to study OCR accuracy and retrieval effectiveness from OCR generated text. Along with this paper, ISRI is making this collection available to researchers for further studies.

2 Collection Characteristics

The collection is comprised of documents dealing with scientific, technical, and legal issues. Many of these documents are full of maps, scientific formulas, and other types of graphical material. Although most of the documents are scientific, within this domain, the collection is diverse. These documents cover topics from rock mining to safety issues for the transportation of nuclear waste.

Characteristics of this collection that make it useful for demonstrating the ramifications of using OCR data include its lack of uniformity in page format, its diverse set of font styles, and its variation in hard copy quality. There are almost as many different authoring sources as there are documents. This document collection consists of full-text documents. The average document length is 44 pages and the median length is 22 pages.

There are four versions of the documents in the collection, of which three will be made available for research. They are: a set of document images, the manually typed correct text (which we refer to hereafter as the ASCII text), and the OCR text. The recoverability, consistency, and quality of these versions determined whether a document was included in our test collection. The document verification process used to determine document inclusion, *Minimum Document Verification* (MDV), is described in Section 3.

3 Minimum Document Verification

The documents used for our research, and consequently for this collection, are those documents from the LSS prototype that met MDV. MDV can be defined as the verification of correspondence between the set of hard copy documents and their images, and the set of hardcopy documents and their ASCII text. One would assume that such correspondence would be implicit in such a collection, but even a cursory scan of the document set shows that differences exist. Of course, perfect correspondence between the hard copy and the

ASCII text should not be expected since rich text and images cannot be represented in ASCII. But some level of resemblance should reasonably be assumed so that measures of comparison between the ASCII text and optically recognized text are meaningful. Therefore, the purpose of this verification was to ensure each ASCII document file corresponded to its hardcopy/image counterpart, that no gross errors were encountered in either the images or the ASCII, and that the collections created from these texts were suitable for experiments.

There are two separate tasks required for MDV: image verification and ASCII verification. In Sections 3.1 and 3.2 we describe the efforts devoted to analyzing the collection and bringing as many documents as possible to the MDV standard. Any critical information with regard to image or ASCII correspondence is noted and filed with the hardcopy.

3.1 Image Verification

Image verification ensures that a document's images duplicate the document's hardcopy pages. Image verification in most cases is straight forward, but complications can arise. The procedure for image verification follows these basic steps:

1. Display the image.
2. Compare the image to its corresponding hard copy.
 - (a) If the image does not match the hard copy, the image must be re-scanned.
 - (b) Notice general formatting (margin widths, number of paragraphs, tables, or figures).
 - (c) Read a few words or sentences on the top and bottom of the image and compare them to the hard copy.
 - (d) Does the image need to be rotated?
3. Rotate the image if necessary. An image must be rotated if the orientation of a majority of text on the page is not horizontal. In some cases, this is a best guess decision (see Figure 1).
4. Make any general notes or remarks if necessary about this document.

In many cases, the hard copy pages needed to be re-scanned. An image is re-scanned for the following reasons:

- The image has skew of greater than 5 degrees.
- The image is of poor quality with respect to the hard copy. Sometimes an image will have poor quality because the hard copy from which the image originated is poor quality; these images should not be re-scanned.
- The image is missing or corrupt. Many TIFF files received were written in an invalid TIFF format.

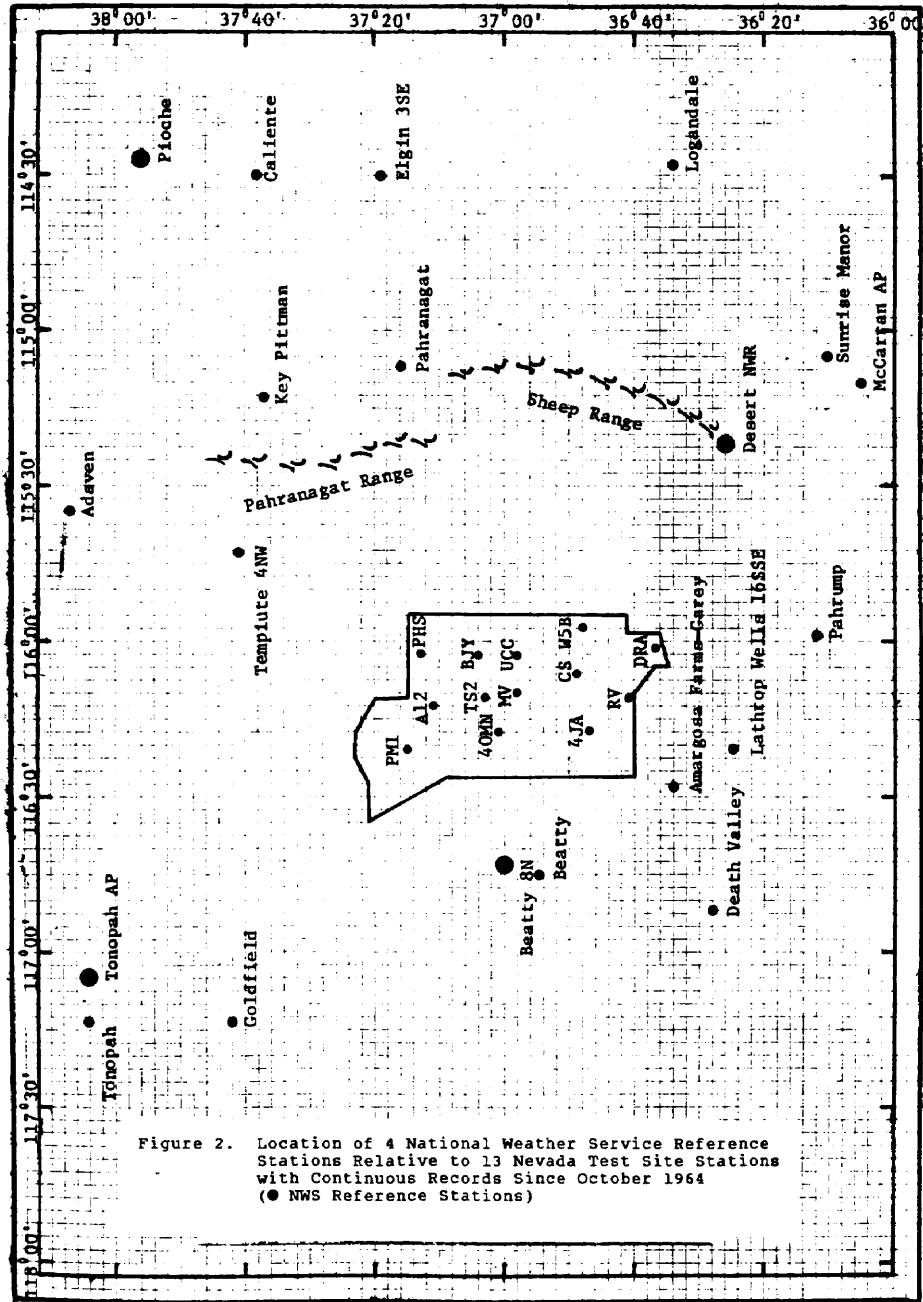


Figure 1: Page orientation is sometimes difficult to determine.

3.2 ASCII Verification

Text verification ensures that the ASCII text is an acceptable ASCII reproduction of the hard copy's (original) text. In some of our experiments, the output of the OCR devices is compared and measured against this text. The correspondence between the ASCII and the document's images is crucial to the validity of any experiments' results. Verifying an ASCII file requires a simple program to view the text. Each page of the text file (pages are delimited with a `Control-L`) is compared to the corresponding hardcopy page. Information about the typed ASCII documents is recorded. The following steps outline the procedure for text verification.

1. Record the docid, the document's page count, whether or not references are present and the number of text columns.
2. Display the ASCII file.
3. Compare the ASCII text to the hard copy.
 - Read a few words or sentences on the top and bottom of the image and compare them to the hard copy.
 - Note any text omissions, table omissions, or text in figures that are not in the ASCII file.
 - Note any other text omissions from the ASCII file.
4. Make any general notes or remarks necessary concerning the document verified.

Table 1 records some of the image and text verification problems that surfaced during MDV.

3.3 MDV Inclusion

Many verified documents had errors ranging from removed tables and figure captions to deletion of main body text. These errors must be documented, and in some cases corrected, to ensure a complete and usable database for our experiments. By noting the differences between typed ASCII and the original hardcopy, we provide a basis against which results from our OCR experiments can be compared. To meet MDV, we must have:

- The original DOE hardcopy document.¹
- Good quality images (with respect to the hardcopy) for the complete document.
- An acceptable representation of the hardcopy.

We classify problems with either the images or the ASCII into two groups: major errors and minor errors. Major errors will disqualify a document and exclude it from the database. The following errors are considered major errors.

- Hardcopy pages are missing.

¹These may be nth generation copies given to us by the DOE.

| Docid | Page(s) | Error |
|-------|---------------------------|---|
| 0637 | 1-4 | images have many broken characters (due to the quality of the original) |
| 0641 | all | This document is double sided. The ASCII file page order is out of order. The pages in ASCII file are in the order: 1,3,5, etc. The images were scanned in the order to match the ASCII file. |
| 0648 | all | This document is doubled sided. Only one side of the page is entered in the ASCII file. Images were scanned to correspond to ASCII file. |
| 0660 | 12 | Docid page count should be 12 as opposed to 11. |
| 0660 | 8 | table is missing from ASCII file |
| 0663 | 1 | Image has text continued from another page continued on the present page. This text is not in the ASCII file. |
| 0663 | 3-7 | text in figure not included in ASCII file |
| 0663 | 3-8 | page header is missing from ASCII file |
| 0663 | 1-2 | there is some text in the beginning of the ASCII file with no corresponding images or hard copies. It is page 1 and 2 of the ASCII file. |
| 0687 | 1 | all text is missing from ASCII file. |
| 1852 | 135-143 | computer code is not included in ASCII file |
| 1852 | 163-164, 228-229, 234,244 | text near mathematical formulas is missing from ASCII file |

Table 1: Sample errors found during MDV.

- If a document is made up primarily of tables, and tables have been deleted from the ASCII version.
- If a document is made up primarily of figures, and text from figures have been removed from the ASCII version.
- Text is written in a foreign language.
- The hardcopy document was in double or triple column format but has not been decolumnized in the ASCII.
- Large amounts of main body text was missing from the ASCII.
- Text sequence of a majority of the document did not match the original hardcopy document.

Minor errors are errors which are not threatening to a document's use in experimentation, but should be noted. Noting these errors provides a means of verifying why there are differences in expected results. The following are classified as minor errors.

- Computer output or programs are missing from the ASCII file.
- Page headers and footers are not in the ASCII file.
- Table of contents formatting errors occurred in the ASCII file.
- Some images contained a slight skew.
- Some images needed to be rotated.
- Tables are missing from the ASCII, but the document is primarily main body text.
- Text in figures is missing from the ASCII, but the document is primarily main body text.

We did not anticipate the need for such an extensive verification when we initially received the LSS prototype collection. But the electronic database was itself inherently difficult to create from hardcopy. The documents are primarily scientific and technical; the texts are filled with graphs, tables, maps, formulas and special symbols. Some document qualities were difficult or impossible to represent in ASCII so rules were devised to resolve these problems. Unfortunately, with several contractors using different methods for conversion, interpretation of these rules was not consistent. Further, whenever human intervention is introduced some errors and inconsistencies should be expected.

We also found that the procedures used for storing data were not standardized or documented. The tapes we received in some cases were unreadable or corrupt. Of the tapes we were able to read, file formats and naming conventions were not consistent making automatic extraction difficult. For all these reasons, the number of recoverable and usable documents was greatly reduced from the original prototype version.

4 Queries and Relevancy Judgments

There are 68 queries in this collection that are a subset of the LSS prototype test questions. These queries were artificially constructed to evaluate how well users were able to retrieve needed information from the database. Many of the queries were written to retrieve information from the structured fields of records, not the actual text. Some of these structured fields are: author name, title, and document type. We have excluded the original queries that only dealt with structured fields.

The relevancy judgments were not part of the collection we received from the DOE. We have been collecting this information since we began our original accuracy experiments. The queries were examined and the complete set of documents was divided among a group of geology graduate students to determine relevance. The students made binary relevancy judgments, classifying documents as either relevant or not relevant. No relevancy ranking was assessed.

5 Prior Research

To date, we have completed five experiments to better understand and partially answer the fundamental question: *To what extent do OCR errors affect retrieval effectiveness?*

5.1 Simulated OCR Collections

We chose four standard collections for our simulation experiment: CACM[3], NPL[5], WEST[2], and WSJ[4] collections. These collections are used in a variety of IR experiments and they represent a wide range of sources and varying document and query lengths.

In order to simulate the OCR databases, IR test collections were indexed by randomly assigning the text of a document to a page group and then randomly discarding index words according to the error rates for that page group and word length.

5.2 Exact Match Model and OCR Text

Our experiment with an exact match system was our first of the five experiments[9]. We chose a subset of our best database (204 documents) and ran 71 queries on this collection. We then ran the same set of queries on the corresponding corrected set and compared the overlap of the documents retrieved. For these 71 queries, there were 632 documents returned in the correct database and 617 in the OCR database. Fifteen documents were missing from the OCR result sets.

These experiments led us to believe that, at least in an environment where 100% accuracy is not imperative, OCR and full-text information retrieval based on exact match technology can be applied in succession with little or no human intervention.

5.3 Probabilistic Model and OCR Text

For this experiment, we loaded four collections (correct, worst, middle, best) with varying levels of character accuracy into a probabilistic retrieval system[7].

The most noteworthy anomaly of the OCR collections was not something that was immediately obvious by examining statistics or comparing recall and precision results. Since

probabilistic systems employ term weighting, term frequencies greatly impacted document-to-query relevance. Through query-by-query examination, we found remarkable variability in document ranking between the OCR collections and the corrected set, and among the OCR collections themselves. We found that the formula used by this particular probabilistic system to assign concepts (in our collection, document terms) to documents and queries may not have given an accurate representation of the documents due to characteristics of the OCR text (e.g. “garbage” strings, misspellings).

Irregularities in document ranking produced unanticipated results in the recall and precision tables. Although the average differences seem insignificant, individual queries in a probabilistic system can be greatly affected by OCR text.

5.4 Vector Space Model and OCR Text

Our previous experiments implied that in an inclusive sense, OCR text had little effect on IR results. But we still felt our testing was incomplete without applying this experiment in a vector-space environment[8].

Using the same LSS collection as in our previous experiments, we applied several different weighting combinations. We selected the most commonly used term, collection, and normalization techniques and applied them singly. But none of the average precision results were significantly different when the averages were compared.

As with our experiment with the probabilistic system, we found with the vector space model, certain characteristics of OCR data caused unstable results for individual queries. We noticed variation in relevant document ranking for all weighting schemes applied when comparing the retrieved document order, or document rank, between the correct and the OCR collections. We attribute the root of the problem to the increased vector size of OCR text documents due to “garbage strings” and misrecognized words. 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.

5.5 Feedback and OCR Text

We also tested the effects of OCR errors on relevance feedback. Feedback is an automatic process that uses information derived from known relevant and non-relevant documents to reformulate queries[8].

The results we show between the correct and OCR collections give a good indication of what should be expected with respect to the use of feedback as a means of improving possibly corrupt data: that the difference is a consequence of the continued improvement in average precision of the correct collection—not the deterioration of the OCR results. The documents that moved to a better rank (i.e. top fifteen) in the correct collection were not influenced in the same way by query expansion in the OCR collection. We believe from our analysis, that the complications are a result of a few difficult-to-retrieve documents within the OCR collection; a majority of these had contributed to the high variability in rank in our initial experimental runs.

Documents that caused problems with cosine normalization, those with repetitive tabular data or those with numerous OCR-generated errors, cannot be retrieved even after adding feedback. It is apparent from these experiments that feedback cannot be applied with the hope of fixing the few shortcomings of an OCR generated collection.

5.6 OCR'd Short Documents

With several IR models tested using OCR-generated full text, one question that remained open was how well short documents (abstract length) could be retrieved in an un-simulated setting[10].

Our experiment with short documents included 830 abstracts. The documents used in this experiment were members of the UNLV-ISRI collection. They intersect this collection but do not encompass it.

One prominent difference between this experiment and our other OCR experiments was the use of manual zoning. SAIC noted in a report that manual zoning resulted in higher output accuracy[6]. Further, non-text data will not be present in the document records, improving character classification and recognition[1] and minimizing "garbage strings" in the index.

We found little difference in average precision for the correct collection when we compare it to the OCR'd version. We looked closely at individual queries. There were eighteen relevant documents where the ranking differed by more than 50. The most common reason for ranked differences was poor OCR-generated text caused by poor original hard copy. Other explanations for ranked differences include: typing errors in correct text;² part of title omitted from correct text; query terms (including acronyms) misrecognized by the device; and one abstract could not be completely recognized. Note that, unlike our vector-space experiment using full text, cosine has no adverse affect on ranking.

6 The UNLV-ISRI Collection

The collection accompanying this paper has the following components. Table 2 shows some general information about the collection.

Correct Text Consists of one file for each document containing the manually corrected text for that document. Files are named *ddd.txt*, where *ddd* is a document ID. The files are ASCII text files with lines terminated by newlines (Control-J). Pages within the file are delimited by form feeds (Control-L).

OCR Text Consists of one directory for each document, named after the document ID. Each directory contains one file for each page of the document. Each file contains the unedited output from the Xerox TextBridge version 4.5 OCR system for that page. Files are named *ddd-*nnn*.txt* where *ddd* is the document ID number and *nnn* is the page number. Note in Table 2 that 626 pages of the OCR generated text are empty files. This occurs when the OCR software was not able to find any recognizable text on the page.

Images Consists of one directory for each document, named after the document ID. Each directory contains one file for each page of the document. Each file is a TIFF file using group 4 compression. The images were scanned at 300 dpi using 1 bit per sample. Files are named *ddd-*nnn*.tiff* where *ddd* is the document ID number and *nnn* is the page number. Note that in Table 2 there are 7 empty image files. These files represent either blank pages or pages with figures that were excluded from the hard copy document. We have added corresponding empty OCR files as place holders.

²OCR documents were ranked higher than the correct.

| | |
|--|---------------|
| Number of documents | 1,056 |
| Number of pages | 46,586 |
| Average number of pages per document | 44 |
| Median number of pages per document | 22 |
| Empty OCR text files | 626 |
| Empty image files | 7 |
| Number of queries | 68 |
| Average number of relevant documents per query | 16 |
| Average correct text file size | 75 kilobytes |
| Average OCR text file size | 2.5 kilobytes |
| Average page image file size | 45 kilobytes |
| Total size of all components | 2.4 gigabytes |

Table 2: Some statistics about the collection.

Queries and Relevancy Judgments Consists of two files, the queries and relevancy judgments. For each query listed in the query file, the query ID is given, followed by a newline, then the text of the query, followed by another newline. The relevancy judgment file contains lines of the form *qqqq-qq-qq: dddd* where *qqqq-qq-qq* is a query ID and *dddd* is a document ID. Each line in the file indicates that the given document is relevant to the given query. Some queries do not appear in the file, and so they have no relevant documents in the collection.

References

- [1] Richard G. Casey and Kwam Y. Wong. *Image Analysis Applications*, chapter 1, pages 1–36. Marcel Dekker, 1990.
- [2] W. B. Croft, S. Harding, K. Taghva, and J. Borsack. An evaluation of information retrieval accuracy with simulated OCR output. In *Proc. 3rd Symposium on Document Analysis and Information Retrieval*, pages 115–126, Las Vegas, NV, April 1994.
- [3] Edward A. Fox. Characterization of two new experimental collections in computer and information science containing textual and bibliographic concepts. Technical Report 83-561, Department of Computer Science, Cornell University, Ithaca, NY, September 1983.
- [4] D. Harman. Overview of the first TREC conference. In *Proc. 16th Intl. ACM/SIGIR Conf. on Research and Development in Information Retrieval*, pages 36–47, Pittsburgh, PA, June 1993. ACM Press.
- [5] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Proc. and Management*, 24(5):513–523, 1988.
- [6] Science Applications Intl. Corp. Capture station simulation: Lessons learned, Final Report, for the Licensing Support System, November 1990.

- [7] Kazem Taghva, Julie Borsack, and Allen Condit. Results of applying probabilistic IR to OCR text. In *Proc. 17th Intl. ACM/SIGIR Conf. on Research and Development in Information Retrieval*, pages 202–211, Dublin, Ireland, July 1994.
- [8] Kazem Taghva, Julie Borsack, and Allen Condit. Effects of OCR errors on ranking and feedback using the vector space model. *Inf. Proc. and Management*, 32(3):317–327, 1996.
- [9] Kazem Taghva, Julie Borsack, Allen Condit, and Srinivas Erva. The effects of noisy data on text retrieval. *J. American Soc. for Inf. Sci.*, 45(1):50–58, January 1994.
- [10] Kazem Taghva, Julie Borsack, Allen Condit, and Padma Inaparthi. The effects of OCR errors on short documents. Technical Report 94-10, Information Science Research Institute, University of Nevada, Las Vegas, February 1995.