

The Impact of Running Headers and Footers on Proximity Searching

Kazem Taghva, Julie Borsack, Tom Nartker, Jeffrey Coombs, Ron Young
Information Science Research Institute
University of Nevada, Las Vegas
Las Vegas, NV 89154-4021, USA

ABSTRACT

Hundreds of experiments over the last decade on the retrieval of OCR documents performed by the Information Science Research Institute have shown that OCR errors do not significantly affect retrievability. We extend those results to show that in the case of proximity searching, the removal of running headers and footers from OCR text will not improve retrievability for such searches.

Keywords: OCR, running header, running footer, information retrieval, proximity search

1. INTRODUCTION

The Information Science Research Institute (ISRI) has performed hundreds of experiments and a thorough analysis of the effects of OCR error on information retrieval. All of our tests have shown that average recall and precision is unaffected by OCR error.^{1,2} One notable side effect that we discovered in our experimentation is that OCR document collections tend to have *more* correctly spelled words than a re-keyed “correct” version.³ Although this seems like a contradiction, if you look at a page of a document, a re-keyed version typically includes just main body text. The OCR version contains every recognizable string, including figure captions, text in tables, page numbers, and running headers and footers. In general, these recognized strings increase recall. So for a particular query, our studies show that an OCR collection will return more hits than its corresponding correct version.

But can words that are not part of the main body text cause problems for certain types of queries? This paper addresses this issue for running headers and footers with respect to *proximity queries*. Proximity queries are exact match queries where word distance can usually be indicated. For example, DEPARTMENT within 3 ENERGY is a proximity query that should return all documents where the word DEPARTMENT is within 3 words of ENERGY, e.g., DEPARTMENT OF ENERGY. We describe our test design, comparison of overall results, comparison results specific to the header/footer/proximity query question, and our conclusion.

2. TEST ENVIRONMENT

2.1. Test Design

The design of our test required addressing the effect of headers and footers included in OCR text on the result sets of proximity queries. In most retrieval experiments, precision and recall is used for measuring the quality of results to a query. But for this experiment, these metrics would not give us pertinent answers. To make this clear, one needs to understand how “proximity queries” work. For most retrieval systems, proximity queries default to an *exact match boolean query*. This means that a document will *only* be returned for a query if the terms appear exactly as they appear within the indicated proximity. This of course may or may not equate to document relevance. But what we want to discover in our tests is,

if proximity query terms cross page boundaries (where headers and footers occur), will these documents still be returned by the retrieval system?

So the design of this test required a different approach than most of our other retrieval experiments. Our test strategy follows these basic steps:

1. Select a set of documents for which we have:
 - unzoned (no header/footer removal) OCR text
 - correct text with headers and footers removed.
2. For each page of a correct document, select a reasonable proximity query distance at the top and bottom of the page to generate potential proximity queries. We selected 10 since this is the default used by Autonomy, the test retrieval system used in this study. Pair the last five non-stopwords on a page with first five non-stopwords of the next page. Generate all ordered combinations.
3. Filter this list for “sensible queries.” We used two filters to obtain our final query list:
 - a list of co-occurrence terms from a superset of our test document collection
 - a list of manually assembled phrases generated from a superset of our collection.
4. Index the correct collection (used to generate the unfiltered phrase list) and the OCR collection using Autonomy, the Licensing Support Network (LSN) retrieval system.
5. Finally, run phrases as proximity queries against both collections and compare result sets for differences.

Once the test design was established, implementation was straightforward.

2.2. Test Element Description

The elements of our test collection are mentioned briefly in Section 2.1. In this Section, we give a more detailed description of these components.

2.2.1. Test Collection

Our test collection consists of a set of documents for which we have both OCR text and the corresponding correct text version. These documents are a subset of the Licensing Support Network (LSN) prototype collection⁴ that we have used in many of our previous experiments. the OCR’d version was generated using Scansoft Developers Kit 2000 (SDK2000) v10.0.⁵ No manual preprocessing of the scanned images was applied, in particular, the pages were automatically zoned.

As stated in Section 2.1, the initial superset of queries was produced from the correct text version of the collection. At every page break (marked by ctrl-L), the last five non-stopwords prior to a page break and the first five non-stopwords following the page break were combined to generate 25 ordered query pairs for each page boundary in the collection. This process produced 774,775 possible queries. Not only was this set too large, many of the queries were nonsensical. For example, “area area” or “ma fig” would not be typical user queries. We wanted to pare down our query list but without any bias. From unrelated experiments we had both co-occurrence “phrases” and manually selected phrases from a superset of this collection. These were used as a filter to reduce the query set to something manageable.

2.2.2. Proximity Query Selection

Although the phrases we used to filter our initial query list were not particular to this project, it was a satisfactory and unbiased means of preparing our final proximity query list. The procedures used for creating these two lists follow.

- *Co-occurrence Phrases* The first list of terms consisted of terms determined to be “similar” and was created in the following way.
 1. Documents from the OCR text collection were indexed.
 2. For each non-stopword t which occurred in 3 or more documents, the **mutual information** (defined below) of t and every other distinct term t' was computed.

3. For every non-stopword t , construct a list of five terms with highest mutual information in relation to t .

Mutual information (I) is a measure of the amount of information a term t provides about another t' .⁶ It is defined by the formula:

$$I(x_t, y_{t'}) = \sum_{x_t \in \{1,0\}} \sum_{y_{t'} \in \{1,0\}} p(x_t, y_{t'}) \log \frac{p(x_t, y_{t'})}{p(x_t)p(y_{t'})}$$

The expressions x_t and $y_{t'}$ stand for random variables which map term, document pairs to the set $\{1,0\}$, that is, they represent the presence (1) or absence (0) of a term within a document. In the case where $x_t = 1$, $p(x_t)$ represents the probability that t occurs in a document and where $x_t = 0$, $p(x_t)$ is the probability that t fails to occur. When $x_t = 1$ and $y_{t'} = 1$, $p(x_t, y_{t'})$ is the probability that t and t' occur together in a document.

Two terms with high mutual information will tend to be similar in the sense that they are distributed within a document collection in the same way. For example, the following list was generated for the term **etiolog**.

etiolog: bromu, mycoplasma, pathogen, blacktail, black-tail

These terms have been stemmed, which means that morphological variants of terms are removed. For example, the stem **etiolog** replaces **etiology**, **etiologies**, **etiological**, etc. In the example, the term **etiology** was found in the same three documents in the collection as the terms **bromus**, **mycoplasma**, **pathogen**, and **black-tail**. The example illustrates that within our document collection, these concepts will appear in roughly the same groups of documents.

- *Manually Selected Phrases* These phrases were selected from three different sources.
 - Phrases from a domain specific thesauri
 - Phrases associated with Guidance specific to the LSN
 - Phrases associated with 1054 documents known to be relevant to the LSN

In all, there were 32,578 unique two-word phrases generated from these for the manually selected phrases.

Once both lists were created, queries were generated in the following way:

1. In the case of the co-occurrence terms, if the first term of the co-occurrence list matched the first term in the initial superset query list, and one of the five similar co-occurrence terms matched the second, the term pair was selected as a query. There were 223 proximity queries generated from the co-occurrence phrase list.
2. In the case of the manually selected phrases, the two-word phrases were compared with the proximity term pairs, and if there was a match, the pair was kept as a query. There were 115 proximity queries matched from this list.

Table 1 gives statistics about the collection and the queries.

2.2.3. Autonomy Retrieval System

Both the correct and the OCR'd versions of the collection were loaded into the Autonomy Retrieval System.⁷ We applied Autonomy for no other reason than this is the retrieval system selected for the Licensing Support Network (LSN). There are though several parameters that an administrator can set that may affect the way Autonomy handles proximity searching. Our decision on these settings was guided by the settings that will be used for the LSN. For completeness, we list them in table 2.

Number of documents	720
Total number of pages	33,468
Number of unique queries	338
Total number of expected “hits”	594

Table 1. Test Collection Statistics

STOPLIST	./langfiles/english.dat
STRIPLANGUAGE	0
PROPERNAMES	1
PROXIMITY	1
MAXPROX	10
COMBINE	1

Table 2. Autonomy Settings

3. COMPARISON OF CORRECT VS. OCR RESULT SETS

As mentioned in Section 2.1, we did not use recall and precision to compare the correct results to the OCR results. Our objective was to discover if headers and/or footers would cause documents not to be returned in the OCR result sets. One must assume then that since the original queries were generated from the correct collection, what is returned for the correct run is ground-truth. The results in table 3 are based on this assumption.

Note that even the correct collection did not return all the expected hits. This is difficult to explain since the queries were generated from this set. One explanation could be the way Autonomy “counts terms” for proximity querying. Since we do not have access to Autonomy source code, we can only speculate. Note also that the OCR result sets returned seven of the expected hits where the correct collection did not. Again, without a complete understanding of Autonomy’s search technology, we cannot attempt to explain why. The manual analysis shown below however, gives some explanation of why these documents were not returned.

So for the 31 documents returned in the correct but not in the OCR results, there is potentially one document from one query that may not have been returned due to the non-removal of headers and footers. We say potentially because there was also a graphic near the top of the page that generated some graphic text which interceded the main body text on consecutive pages. Note that manual intervention (the re-keying of the correct text) caused six of the 31 errors. For these, text, both missing and inserted, created erroneous proximity queries. This situation would be comparable to getting non-relevant hits to real-world queries.

It may seem significant when 31 “hits” were returned in the correct collection but not in the OCR even though the misses were not caused by headers and footers. One might believe that a correct collection will return superior results. Keep in mind though that this test is somewhat biased to the correct collection since the proximity queries were generated from the correct set, and it was used as ground truth for the results in table 3. Further, a comparison of the complete results were not considered in the above analysis. Table 5 presents more comprehensive results.

	<i>expected hits</i>	594
	correct collection hits	555
	OCR collection hits	531
	in correct result sets but not in OCR result sets	31
	in OCR result sets but not in correct result sets	7

Table 3. Correct vs. OCR Results

<i>missing from OCR result sets</i>	31
OCR error unrelated to header/footer	22
correct text error, missing text	4
correct text error, text inserted	2
hyphenation error	1
OCR inserted graphic text and header	1
undetermined	1

Table 4. Reasons for Non-retrieval of OCR Documents

	<i>Correct</i>	<i>OCR</i>
Number of documents returned	15,881	16,555
Number of documents <i>not</i> in compared set	333	1007
Number of subset queries (76 identical)	121	167

Table 5. Comprehensive Results

Note that for these statistics the OCR returns more hits than the correct collection. This result was expected. In all our comparison tests with correct and OCR collections this has been the case.¹⁻³ But note also the second row of the table; there were 1007 documents returned for queries in the OCR result sets that were not returned in the correct result sets. What’s more, the third row indicates that of the 338 queries run, in 167, the correct result sets were a proper subset of the OCR result sets; the opposite was only true for 121. This says that in general, recall is higher for the OCR collection than for the correct collection.

4. CONCLUSION

Based on this study, we believe that eliminating headers and footers from OCR’d documents will not improve retrievability for proximity queries. In fact, if important text is removed, recall could be negatively affected. Comparison of correct vs. OCR text has proved to us time and time again that manual intervention can cause more problems than it fixes. We know from previous tests that average recall and precision is unaffected when OCR text is used. We have also shown that in most cases, OCR recall is higher. This test re-establishes our previous results and demonstrates that headers and footers are a non-issue for proximity searching.

REFERENCES

1. K. Taghva, J. Borsack, A. Condit, and S. Erva, “The effects of noisy data on text retrieval,” *J. American Soc. for Inf. Sci.* **45**, pp. 50–58, January 1994.
2. K. Taghva, J. Borsack, and A. Condit, “Evaluation of model-based retrieval effectiveness with OCR text,” *ACM Transactions on Information Systems* **14**, pp. 64–93, January 1996.
3. K. Taghva, J. Borsack, and A. Condit, “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), July 1994.
4. Science Applications Intl. Corp., “Capture station simulation: Lessons learned, Final Report, for the Licensing Support System,” November 1990.
5. Scansoft, Inc., Peabody, MA, *Recognition API Manual*, v10 ed., 2000.
6. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley, 1991.
7. Autonomy, Inc., San Francisco, CA, *Autonomy Knowledge Server*, 2.2.0 ed., 1999.