

A Comparison of Automatic and Manual Zoning An Information Retrieval Prospective

Kazem Taghva, Julie Borsack, Steven Lumos, and Allen Condit

Information Science Research Institute, e-mail: isri@isri.unlv.edu

The date of receipt and acceptance will be inserted by the editor

Abstract. In this paper, we study the effects of automatic zoning on retrieval and ranking variability. We will show that OCR generated text from automatic zoning, followed by post processing, produces retrieval results equivalent to OCR generated text from manual zoning. We further show that there is a strong linear association between the ranked query results obtained from these two methods of zoning.

1 Introduction

In both the government and private sector there has been a directed effort to convert paper-based information to electronic form. The Digital Libraries Initiative, Freedom of Information Act, information sharing, mandated state and federal regulations for improved record access, electronic health records, litigation support, and IRS reform is a short list of projects that require this form of conversion. The notion of hard copy conversion though, conjures up a long list of manual processing steps that include scanning, organizing, indexing, re-keying, manual cleanup, etc. The costs associated with these conversion steps for a project of any magnitude is prohibitive. Optical Character Recognition (OCR) was one of the first technologies that reduced the amount of manual intervention into the conversion process, but the generated text still contained errors.

From the retrieval point of view, it has been shown that OCR errors do not affect average precision[2,11,12]. It is also known that for low quality documents (e.g., faxed documents, n^{th} generation photocopies), we may have to work harder in order to retrieve them[12,4]. This brings up a few interesting and open problems that should be addressed. For example, at what character accuracy do we start seeing a degradation in retrieval? Hawking[4] has shown that a 5% character error rate may hinder the retrieval of certain documents. Also, Taghva[12,16] has shown that feedback is not as effective in OCR generated text as it is in clean text. Both

authors report on ranking variability caused by OCR errors: a relevant document may be ranked 20 in the clean text but ranked 100 in the OCR generated text.

One major task in the conversion process is *zoning* which separates document text from non-text while preserving the reading order. In a production environment, the task of zoning can be done manually or automatically. Automatic zoning can introduce errors that may affect the reading order of the text. It may further produce long spurious character strings known as *graphic text* or *garbage strings*. Since these garbage strings become a part of the index for retrieval, it can increase document length[11,12,9]. It is also shown that document length normalization techniques are the main cause of ranking variability. For these reasons, some organizations with large conversion projects employ manual zoning instead of automatic zoning.

This paper addresses the issues of retrieval effectiveness and ranking variability when automatic zoning is applied. We show that post processing of OCR text can help solve the ranking variability problem that may be encountered in the production environment.

2 Background

The Licensing Support Network (LSN), managed by the Nuclear Regulatory Commission (NRC), will provide information that is potentially relevant to the licensing of the high-level radioactive waste repository proposed for Yucca Mountain, Nevada to all interested parties. There are several organizations contributing documents to the LSN, but the Department of Energy (DOE) as the licensee, will submit the vast majority of these documents. Since the document collection will be very large, the identification and retrieval of documents must be timely and effective. What's more, effective retrievability should be of the utmost concern. Whatever capture method is selected, it should produce the expected results.

The retrievability studies we perform and explain in this paper evaluate and compare automatic and manual methods that can be used for document conversion of

Document count	1055
Number of pages	75,236
Query Count	40
Average number of relevant documents/query	100
Median number of relevant documents/query	64
Fewest relevant documents for a query	2
Most relevant documents for a query	608

Table 1. ISRI Collection Statistics for Auto-Zoned vs. Manual-Zoned Test

text from images. These results will enable the DOE, and other organizations with similar projects, to make a well-informed decision for their document conversion processes. We report on optical character recognition’s (OCR) impact on retrieval for the following:

- Manual zoning vs. automatic zoning (Section 3).
- Differences in ranked results between manually-zoned and automatically-zoned OCR text (Section 3.3).
- Post-processing methods applied to improve the OCR text produced (Section 4).
- The automatic document conversion methods compared to nearly perfect (99.8% correct) text (Section 5).

These tests, in conjunction with the results reported in [5], will provide an in-depth analysis of the performance of our automatic document conversion system.

3 Manual Zoning vs. Automatic Zoning

3.1 Environment

The zoning experiments described in this paper duplicate the DOE production environment. The documents and queries are a good representation of the expected LSN collection and its anticipated use. Our environment consisted of the following software:

OCR: Scansoft v10 with the MTX OCR module[7]
 Post Processing: MANICURE v1.7[15]
 Retrieval Engine: Autonomy Server v2.2.0[1]

The test collection that we use to compare manual vs. automatic zoning consists of 1055 documents that were selected from documents in the RIS with the document type “Report,” “Plan,” “Design Document,” or “Correspondence.” Manual zoning information for each page was made available by the Yucca Mountain Project management and operations contractor. Forty queries with relevancy judgments were developed by UNLV Geology students who were familiar with the DOE collection. Table 1 shows some statistics for this dataset.

3.2 Recall/Precision Results

We produced two versions of the 1055 document collection described in Section 3.1. For one version, we applied the manual zoning information described above (call it

	Manual-Zoned	Auto-Zoned
Recall	Precision	Precision
0.00	0.78	0.84
0.10	0.58	0.61
0.20	0.50	0.54
0.30	0.44	0.46
0.40	0.41	0.42
0.50	0.36	0.35
0.60	0.31	0.31
0.70	0.27	0.26
0.80	0.22	0.22
0.90	0.18	0.18
1.00	0.12	0.12
Average	0.379	0.392

Table 2. 11-Point Precision for Manual-Zoned Vs. Auto-Zoned Sets

manual-zoned). We also produced another version applying automatic zoning performed by Scansoft (call it *auto-zoned*). Both collections were loaded and indexed into Knowledge Server. All 40 queries were batch run against these two datasets in exactly the same way.

The objective is to compare these two result sets against each other. *Recall* and *precision* are the accepted measures applied in the Information Retrieval (IR) community for comparing retrieval results[6][3]. In Table 2, recall percentages are shown in the left most column with the corresponding precision values at these recall points. For example, when the system has returned 20% (0.20 recall) of the relevant documents in the collection, 50% of those returned in the manual-zoned set were relevant and 54% of the auto-zoned were relevant. The precision values represent *averages* across all queries. The last row of Table 2 is the average of the precision values in the columns. Table 2 shows precision at 11 recall points for both the manual-zoned and the auto-zoned versions of this collection.

Note that the 11-point average for auto-zoned is 3.5% better than for manual-zoned. This higher average return for auto-zoned indicates that running these queries against this data set gives better results from automatic zoning than one could expect from manual zoning. This difference is not statistically significant though. To be statistically significant for this size collection, the difference would have to be 5% or greater. What we can learn from these results is that in general, automatic zoning can give results as good as those obtained from manual-zoned OCR.

3.3 Ranked Query Analysis

The 11-point precision average indicates that there is no difference in query results for the methods used for collection preparation. On the micro level, we felt it was important to investigate what exactly happens to individual query rankings. In other words, if a relevant document was ranked 25 in the automatic results, what would be the rank of the same document in the manual version? By reviewing the query-by-query results for

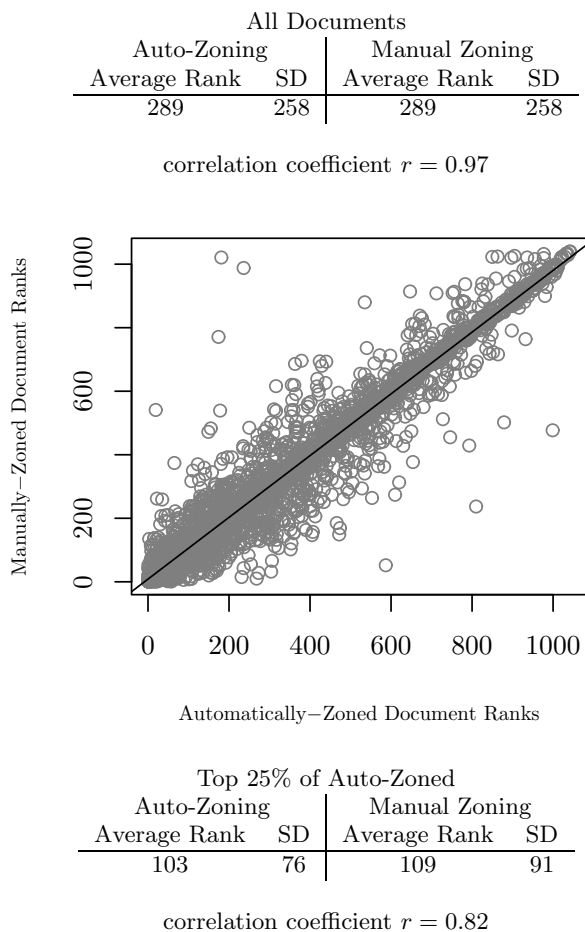


Fig. 1. Auto-Zoned Ranks vs. Manual-Zoned Ranks scatter plot and summary for all documents, and high ranking auto-zoned documents.

both versions, there seemed to be no significant variation between rankings. The following paragraphs report the technical details supporting our observation.

The collection has 1055 documents. Hence, for a specific query, a relevant document can be ranked between 1 and 1055. Obviously, we would like to see the relevant documents ranked as close to 1 as possible. Now consider all the relevant documents that were ranked 3 for the automatic version. We can find the corresponding rank of the same documents in the manual version and represent the resultant pairs as points: (3, 6), (3, 1), (3, 135), etc.

If we continue this process for all the ranks and plot these points, we will discover the scatter plot in Figure 1. This graph exhibits the relationship between the corresponding rankings. In other words, for a fixed rank m on the x axis, the auto-zoned ranks, then the y values, the manual-zoned ranks, represent the corresponding ranks for the same documents for the manual-zoned.

Formally, we are interested in the relationship between the quantitative random variable y (Manual-Zoned Rank) to the quantitative variable x (Auto-Zoned Rank). Intuitively, we want to know if we employ automatic zoning in our production environment, then how will ranking of the relevant documents be affected. One measure

All Documents	
Auto-Zoned Rank	Predicted Manual-Zoned Rank
200	203
150	154
100	106
50	56
10	19
350	349
400	397

Top 25% of Auto-Zoned	
Auto-Zoned Rank	Predicted Manual-Zoned Rank
50	57
25	32
5	13
150	155

Table 3. Predicted Manual-Zoned Ranks using the Regression Method

of the strength of the relationship between two variables is the correlation coefficient. Figure 1 also contains a summary of the plot by average ranks, standard deviation, and correlation.

We can use method of *Least Squares* to obtain the *regression line* and predict the manual ranks from the automatic ranks. The solid line in Figure 1 represents the regression line.

As can be seen in the scatter plot, the points in this graph are tightly clustered around the regression line. This clustering indicates a strong linear association between the two variables. In general, we use the equation of the regression line for calculating the predicted manual ranking of the same document in the manual version.

We can use the above analysis to extrapolate the rank performance of manual-zoned from the auto-zoned version. As can be seen from Table 3, the further away we are from the point of average the bigger difference we see in the ranks. This is the way the regression method works. To lend more credence to our analysis, we did the same calculation for the highest ranking documents (i.e., only the ranks between 1 and 261), since these are the documents that the user will most likely evaluate. The correlation when considering only these documents is given in Figure 1 and Table 3 under ‘Top 25% of Auto-Zoned’.

The regression method is one way of comparing the ranking correspondence between the two collections or in this case, between two versions of the same collection. In our experiments, it can be seen that there is no significant difference in the ranking between the two methods of document preparation.

4 Post Processing of Auto-Zoned Text

Based on ISRI’s OCR and information retrieval research, several post-processing routines were built to improve

Recall	Auto-Zoned with <code>rmgarbage</code>	Auto-Zoned w/o <code>rmgarbage</code>
0.00	0.84	0.83
0.10	0.61	0.60
0.20	0.54	0.53
0.30	0.46	0.44
0.40	0.42	0.40
0.50	0.35	0.35
0.60	0.31	0.31
0.70	0.26	0.26
0.80	0.22	0.22
0.90	0.18	0.18
1.00	0.12	0.12
Average	0.392	0.385

Table 4. 11-Point Precision for Auto-zoned with/without `rmgarbage`

the quality of OCR text loaded into a retrieval system. This set of processes eventually was streamlined into a system we call MANICURE[15].

The major components of MANICURE include `ppsys`[10], a process that automatically corrects misspellings in the text, and `rmgarbage`, a process which removes “graphic text” and other non-retrieval strings from an automatically-zoned and OCR’d document. Previous experiments [11] have proven the effectiveness of `ppsys` but the level of improvement together with `rmgarbage` had not been measured. A simple means of testing its effects was to run the same experiment discussed in Section 3 except that the tested sets would be two versions of auto-zoned: auto-zoned with `rmgarbage` and auto-zoned without `rmgarbage`. All other processing steps remained the same. The results of this test appear in Table 4.

Reviewing these results, we see slightly increased precision at the highest recall values and average precision is nearly 40% when `rmgarbage` is used. Again, although not statistically significant for this set of documents, a small but definite improvement is apparent with the complete MANICURE system.

These experiments show that with automatic zoning and MANICURE, users will obtain retrieval results equivalent to what could be expected with manual zoning. Further, since in many cases non-stopwords [5] are corrected when MANICURE is applied, retrievability can potentially be improved.

5 Automatic Zoning vs. 99.8% Correct Text

One might believe that the closer we get to 100% character accuracy, the better the retrieval results we will obtain from a search engine like Autonomy. In fact, one of the goals specified by the NRC is that collections submitted for the LSN should try to reach 99.5% character accuracy across the collection and 98.5% for any particular page. What this next experiment shows (and several other experiments performed at ISRI have shown)[14, 11, 2, 12, 13] is that close to 100% character accuracy may not be necessary for good retrieval performance.

Document count	1058
Number of pages	46,731
Query Count	62
Average number of relevant documents/query	17
Median number of relevant documents/query	9
Fewest relevant documents for a query	1
Most relevant documents for a query	99

Table 5. LSS Prototype Collection Statistics

Recall	99.8% Correct	Auto-Zoned with MANICURE
0.00	0.55	0.54
0.10	0.46	0.45
0.20	0.35	0.34
0.30	0.29	0.30
0.40	0.26	0.26
0.50	0.22	0.22
0.60	0.18	0.18
0.70	0.14	0.14
0.80	0.12	0.11
0.90	0.08	0.07
1.00	0.06	0.05
Average	0.245	0.242

Table 6. 11-Point Precision for 99.8% Correct Text vs. Automatically-zoned and Recognized Text

We have a collection of 1058 documents, 62 queries, and 1104 relevancy judgments that we will use to answer this question. This collection is particularly well-suited for determining how character accuracy may affect retrieval performance for a couple of reasons. First, we have two versions of the collection: one version with 99.8% character accuracy[8] and another version that has been recognized and processed as described in Section 3.1. Second, the documents and the queries in this collection were part of the original LSN Prototype Collection and so they should have similar characteristics and topic content as the planned LSN. Collection statistics appear in Table 5.

Again, as in our tests comparing manual vs. automatic zoning, we report on retrieval results using the standard measures of recall and precision. We loaded and indexed both collections into Knowledge Server and the 62 queries were batch run against these two datasets in exactly the same way. The recall/precision results appear in Table 6.

The difference between average precision for the two runs is less than 0.3%. As we pointed out for the manually-zoned vs. automatically-zoned runs, the difference is too small to be considered statistically significant. This test tells us that the process used by DOE to prepare the documents for the LSN will return results equivalent to a collection that was corrected to meet 99.8% character accuracy. With respect to retrievability, an artificially high character accuracy does not guarantee better results for the end user.

6 Conclusion

The aggregation of the experiments explained in this paper tells us in no uncertain terms that using automatic-zoning followed by OCR post processing will give retrieval results equivalent to what one could expect from manually-zoned pages or even from a 99.8% correct collection. We have also shown that there is a strong linear association between ranked results. This association implies that for all practical purposes, the two ranked query result sets are statistically equivalent as well.

We believe these retrieval tests, in conjunction with the results reported in [5], demonstrate that, in general, OCR technology coupled with post processing provides an economical solution to the migration problem from paper documents to electronic form.

References

1. Autonomy, Inc., San Francisco, CA. *Autonomy Knowledge Server*, 2.2.0 edition, 1999.
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. D. Harman. *Information Retrieval, Data Structures and Algorithms*, chapter Ranking Algorithms, pages 363–392. Prentice Hall, Englewood Cliffs, NJ 07632, 1992.
4. D. Hawking. Document retrieval in ocr-scanned text. In *Proc. Sixth Parallel Computing Workshop*, Kawasaki, Japan, 1996. paper P2-F.
5. Tom Nartker and Ron Young. OCR accuracy produced by the current DOE document conversion system. Technical Report 2002-06, Information Science Research Institute, University of Nevada, Las Vegas, April 2002.
6. G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw Hill, New York, 1983.
7. Scansoft, Inc., Peabody, MA. *Recognition API Manual*, v10 edition, 2000.
8. Science Applications International Corporation. Capture station simulation lessons learned. Final report for the Licensing Support System, 1990. Final report for the Licensing Support System prepared under contract DE-AC01-87RW00084 for the U.S. Department of Energy, Office of Civilian Radioactive Waste Management, Washington D.C.
9. Amit Singhal, Gerard Salton, and Chris Buckley. Length normalization in degraded text collections. In *Fifth Annual Symposium on Document Analysis and Information Retrieval*, pages 149–162, University of Nevada, Las Vegas, 1996.
10. Kazem Taghva, Julie Borsack, and Allen Condit. An expert system for automatically correcting OCR output. In *Proc. IS&T/SPIE 1994 Intl. Symp. on Electronic Imaging Science and Technology*, pages 270–278, San Jose, CA, February 1994.
11. 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.
12. 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.
13. Kazem Taghva, Julie Borsack, and Allen Condit. Evaluation of model-based retrieval effectiveness with OCR text. *ACM Transactions on Information Systems*, 14(1):64–93, January 1996.
14. 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.
15. Kazem Taghva, Allen Condit, Julie Borsack, John Kilburg, Changshi Wu, and Jeff Gilbreth. The MANICURE document processing system. In *Proc. IS&T/SPIE 1998 Intl. Symp. on Electronic Imaging Science and Technology*, San Jose, CA, January 1998.
16. Kazem Taghva and Jeffrey Coombs. Hairetes: A search engine for OCR documents. In *Proc. of 5th IAPR Intl. Workshop on Document Analysis Systems*, Lecture Notes in Computer Science, pages 412–422, Princeton, NJ, August 2002. Springer-Verlag.

Kazem Taghva received his Ph.D in 1980 from the University of Iowa. He is currently Associate Director of the Information Science Research Institute (ISRI) and Professor of Computer Science at the University of Nevada, Las Vegas. Prior to joining UNLV, he was Chairman of the Computer Science Department at New Mexico Tech. His current research interest is on the interaction of OCR and IR. He has authored papers published in journals such as ACM Transactions on Information Systems, Information Processing and Management, Theoretical Computer Science, Journal of Information Processing, Information Processing Letters, and the Journal of the American Society for Information Science.

Julie Borsack is a Research Associate at ISRI. Her expertise is the fields of Information Retrieval and Optical Character Recognition. She has her B.S. and M.S. in Computer Science and has numerous publications in related journals and professional conferences.

Steven Lumos has been a Research Programmer at ISRI since receiving his M.S. in Computer Science in 2000. His research interests include Text Classification and Data Mining.

Allen Condit has been a Software Engineer at ISRI for 12 years and has written software for many of the institute's IR and OCR research projects. His research led to the implementation of the MANICURE document processing system for OCR generated text. He has his B.A. and M.S. in Computer Science and has co-authored several significant journal and conference papers.