Exploiting WWW Resources in Experimental Document Analysis Research

Daniel Lopresti

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

Abstract. Many large collections of document images are now becoming available online as part of digital library initiatives, fueled by the explosive growth of the World Wide Web. In this paper, we examine protocols and system-related issues that arise in attempting to make use of these new resources, both as a target application (building better search engines) and as a way of overcoming the problem of acquiring ground-truth to support experimental document analysis research. We also report on our experiences running two simple tests involving data drawn from one such collection. The potential synergies between document analysis and digital libraries could lead to substantial benefits for both communities.

1 Introduction

In the six years that have passed since a paper at an earlier DAS workshop identified potential synergies between the World Wide Web and the field of document analysis [1], the Web has established itself as the largest distributed collection of documents in the history of civilization. Many researchers are now exploring problems that have arisen out of this phenomenon, including, for example, the extraction and recognition of text embedded in color GIF and JPEG images [2, 3]. Document analysis is being applied to the conversion process of placing archival material on the WWW (*e.g.*, [4]). Moreover, the pervasive impact of the Web has spawned work in related areas, including the use of XML to represent recognition results [5]. Such opportunities and challenges were the subject of a recent workshop [6].

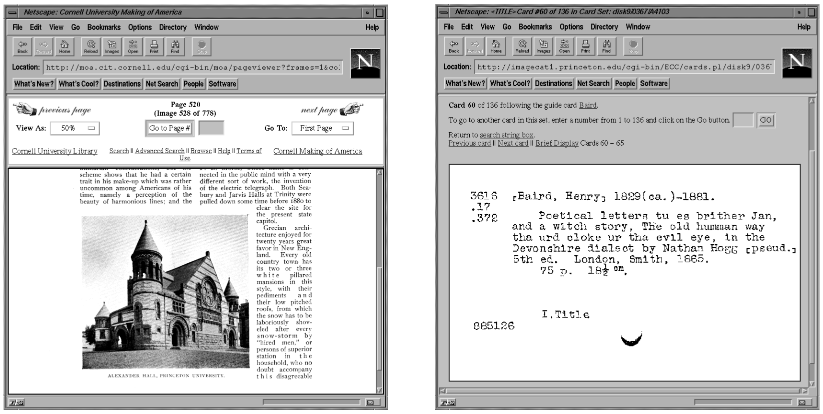
Despite this flurry of activity centered around the Web, there is an important development that appears to have been largely overlooked: that is, the rapidly growing body of traditional scanned documents now being made available online. In retrospect, this should come as no surprise as: (1) the WWW was always touted as a delivery mechanism for multimedia content, (2) documents serve as a basic "quantum" of information in our society, and (3) most users are generally oblivious to the distinction between a page presented in image format versus one encoded in, say, HTML. Often, collections of scanned documents are the product of digital library projects aimed at preserving and disseminating works of historical significance (*e.g.*, [7, 8]).

For example, the *Making of America* collection (part of Cornell University's Prototype Digital Library [7]) comprises 267 monographs (books) and 22 journals (equaling 955 serial volumes) for a total of 907,750 pages, making it almost 1,000 times larger than the dataset offered on the UW1 CD-ROM [9]. The procedures used in creating this digital library match standard methodologies employed in experimental document analysis research:

> "The materials in the MOA collection were scanned from the original paper source, with materials disbound locally due to the brittle nature of many of the items ... The images were captured at 600 dpi in TIFF image format and compressed using CCITT Group 4. Minimal document structuring occurred at the point of conversion, primarily linking image numbers to pagination and tagging self-referencing portions of the text ...
>
> Further conversion included both optical character recognition of the page images, and SGML-encoding of the ensuing textual information." [10]

While OCR results are used for full text retrieval purposes, the default view returned to users of the system is an image of a scanned page.

Fig. 1 shows a snapshot of a browser window displaying a page from *Making of America* on the left [11], and another example of an online document image, a card from the catalog for Princeton University's library, on the right [12].



Location: http://moa.cit.cornell.edu/cgi-bin/moa/pageviewer?frames=1&co

Page 520
(Image 528 of 778)

View As: 50%

Go To: First Page

scheme shows that he had a certain trait in his make-up which was rather uncommon among Americans of his time, namely a perception of the beauty of harmonious lines; and the

nected in the public mind with a very different sort of work, the invention of the electric telegraph. Both Seabury and Jarvis Halls at Trinity were pulled down some time before 1880 to clear the site for the present state capitol.

Grecian architecture enjoyed for twenty years great favor in New England. Every old country town has its two or three white pillared mansions in this style, with their pediments and their low pitched roofs, from which the snow has to be laboriously shoveled after every snow-storm by "hired men," or persons of superior station in the household, who no doubt accompany this disagreeable

ALEXANDER HALL, PRINCETON UNIVERSITY.

Location: http://imagecat1.princeton.edu/cgi-bin/ECC/cards.pl/disk9/036

Card 60 of 136 following the guide card Baird.

To go to another card in this set, enter a number from 1 to 136 and click on the Go button.

Return to search string box.
Previous card || Next card || Brief Display Cards 60 – 65

```
3616   ₍Baird, Henry₎ 1829(ca.)-1881.
.17
.372        Poetical letters tu es brither Jan,
       and a witch story, The old humman way
       tha urd cloke ur tha evil eye, in the
       Devonshire dialect by Nathan Hogg ₍pseud.₎
       5th ed.  London, Smith, 1865.
            75 p.  18½ cm.


            I.Title
 885126
```

Fig. 1. Examples of documents delivered in image format on the Web

As a result of such efforts at bringing scanned documents online, several intriguing opportunities present themselves to researchers working in document analysis. The most obvious of these would be to apply state-of-the-art techniques to build higher quality and/or more powerful indices for information retrieval and presentation. This notion of crafting a better third-party search engine for digital libraries has an analog on the Web as a whole, where competing search engines vie for users by indexing documents encoded in HTML, PDF, PostScript, and other "easy" formats. It is certainly possible to imagine doing a better job on the MOA collection; for example, a search for the keyword "modem" returns 1,364 hits in documents published between 1815 and 1926, even though the word was first coined in the early 1950's.[1] Two of the librarians in the project write:

> "Our attention to retaining pagination and document structure will allow us to selectively insert improved OCR as it is completed. As we insert the more accurate OCR over time, we expect that the greatly improved OCR will make the searching tools even more effective." [14]

Beyond this relatively straightforward improvement, it seems conceivable that higher-level document analysis methods could provide useful new paradigms for retrieval from digital libraries.

Another thought-provoking possibility would be to use existing online collections of scanned images as a way of overcoming the problem of acquiring sufficient training and testing data to support experimental document analysis research. This matter is regarded as so pressing that it was one of the prime motivations behind the creation of the Open Mind Initiative [15], a project to enlist Web users around the world to assist in the labeling of ground-truth data for algorithm development. But while Open Mind deals with this one aspect of the problem, it does not address where the raw data comes from, or what qualifies it as "relevant." These issues will be a focus of this paper.

2 Traditional Approaches

Typically, document analysis researchers either assemble their own collections of scanned images and/or use pre-existing datasets, such as those disseminated by UW [9], NIST [16], UNLV [17], and CEDAR [18]. The former approach allows the corpus to be targeted to the task under study, but the acquisition process can be time-consuming and perhaps expensive. On the other hand, standard datasets distributed on CD-ROM, once purchased, are easy to use and provide a convenient basis for comparison, although they may not cover the precise application of interest, potentially introduce copyright issues, and could become overused to the point that techniques are developed specific to the test set, which is usually relatively small.

Another methodology designed to replace or supplement the previous two approaches involves synthesizing training or testing data. There are, for example, models for generating noisy page images [19] and for creating random instances of tables [20]. While it is possible to produce an endless stream of data in this way, there is always the question of whether such data is truly representative.

[1] This test was inspired by a discussion in Baker's book *Double Fold*, p. 71 [13].

3 Exploiting WWW Resources

As we have noted, there is an enormous quantity of page image data now available on the Web. How might this be used to support document analysis research? Consider the basic steps involved in building datasets for either training or testing purposes: (1) collecting and scanning representative pages, (2) labeling the ground-truth, and (3) distributing the dataset. While the last step might not seem strictly necessary, good scientific practice requires describing experiments in sufficient detail that it is possible to reproduce them. With that in mind, it clearly becomes important that the test data be accessible to other researchers.

With digital libraries, the first and last of these steps are already taken care of. The pages have been scanned and are freely available online. The developer of the library presumably has dealt with any copyright issues connected to the works in question. Furthermore, it is easy to argue that such pages must be representative because they are, in fact, real documents of definite value to some target audience. Still, there remains the question of what to do about labeling the ground-truth. What are the available options?

One solution would be to make use of the existing ground-truth provided by the digital library itself (*e.g.*, the OCR results in the case of the *Making of America* collection). Another would be to develop protocols for using truth produced and/or maintained by a third party (previous researchers who have used the same test documents, or an Open Mind-like entity). A third approach would be to study evaluation techniques that do not depend on having an explicit ground-truth (*e.g.*, comparing retrieval effectiveness relative to what is obtained when using the source library's tools).

4 Proof of Concept: Analysis of a Digital Library

To explore the ideas outlined in this paper, we have performed two simple "proof of concept" exercises: the first using the *pagereader* system developed by Baird at Bell Labs [21] to OCR a set of pages randomly chosen from the *Making of America* digital library [7], and the second examining an algorithm we have proposed with colleagues for table detection [22]. This sort of evaluation is fundamentally different from the kinds typically described in the literature. Because the selection of test images is unbiased and completely automatic, the pages in question are never seen in advance by the researcher(s) involved in running the tests; there can be no attempt, explicit or subconscious, to discard images that do not fit the model or to tune an algorithm to the dataset.[2] As a result, this criterion is almost certainly more demanding than what is normally encountered.

Most research systems for document analysis, including *pagereader* and our table detection code, assume the input image will be in TIF format, however TIF is not a native encoding for current Web browsers. In the case of *Making of America*, the pages are delivered in one of three possible formats: a "50% size"

[2] It is, of course, quite acceptable to maintain a record of the test documents that were used for an after-the-fact analysis.

GIF image, a “100% size” GIF image, and a PDF document containing the original scanned TIF. The GIF forms have relatively low spatial resolution, making use of grayscale (image depth) to compensate, and hence would be difficult to use without a significant amount of extra work. Hence, we chose to implement a process pipeline that first converts the PDF version of the page into PostScript and then extracts the image directly from the PostScript. In addition to the various image “flavors” of the page, the OCR output used to create the searchable index for *Making of America* is available. We can use this text for evaluation purposes, but must be careful about making assumptions concerning its quality or the way that it is formatted.[3]

Lacking our own complete index of the digital library, our approach to retrieving a random page image from *Making of America* is to issue a query by choosing a term from the Unix *spell* dictionary, which contains 24,259 words including a number of proper names. From the results of this search, we randomly choose one of the works (book or journal) that is returned, and from that work we select a specific page that contains a match. The implementation of the Web interface is programmed in Tcl/Tk using the Spynergy Toolkit [23].

It takes a total of six HTTP “round-trips” to get the data we need:

1. First, issue a search request using a randomly chosen keyword and retrieve the results.
2. The results are presented in “slices” of 50 works per HTML page. Randomly select a slice and retrieve it.
3. Within the slice, determine one of the works at random and retrieve it.
4. Within the work, randomly choose one of the matches and retrieve it.
5. Based on the HTML for the final target page, retrieve the PDF file that contains the embedded TIF image (which is then extracted locally).
6. In the same way, retrieve the OCR text corresponding to the target page.

The last step is skipped in the table detection experiment as it is unnecessary. We have developed a set of simple “wrappers” to extract the required information from the HTML code returned by the MOA server.

4.1 Optical Character Recognition

For the OCR experiment, we retrieved 250 pages from the digital library. On the occasions when an HTTP fetch timed-out (after 30 seconds for the initial connection, and 5 seconds for each subsequent buffer), the search was attempted again using a different term.[4] This situation seemed to arise most often when the original query generated an extremely large number of hits (tens of thousands); it is likely that the machine serving *Making of America* builds data structures that grow with the size of the result. The 10 most- and least-frequent matches are listed in Table 1. Note that there is a wide distribution and even arcane terms arise occasionally in the collection.

[3] Generally, we assume that the text may contain a modest number of OCR errors, but that any severe problems will have been detected and corrected by those responsible for building the digital library.

[4] We also re-ran searches that returned no matches.

Table 1. Most- and least-frequent matches in the OCR experiment

Most-Frequent			*Least-Frequent*		
Search Term	*Matches*	*Works*	*Search Term*	*Matches*	*Works*
enemy	236,021	15,000	psychopathic	4	4
science	103,160	31,956	gumdrop	4	2
edge	46,007	20,291	glamorous	3	3
sold	44,467	18,834	pentagram	3	3
empire	42,054	14,677	uninominal	3	3
taught	39,429	20,574	constructible	2	2
request	35,812	12,803	saddlebag	2	2
guide	35,667	17,192	dressmake	1	1
base	34,123	17,139	godparent	1	1
virtue	31,952	16,175	riverfront	1	1

The times need to retrieve and process the pages are graphed, in order of decreasing total time, in Fig. 2 (note that the y-axis uses a log scale).[5] The four components of the total are the times need to: (1) fetch the data, (2) convert the PDF to TIF, (3) OCR the image, and (4) compare the output from *pagereader* to the ground-truth. The minimum total time was 93 seconds, the maximum was 1,966 seconds, and the average was 376 seconds. These values are dominated by the time it took to perform OCR (minimum 42 seconds, maximum 1,890 seconds, average 323 seconds). In other words, OCR was responsible for 86% of the computation time, on average. On the other hand, processing the HTTP requests and retrieving the page images over the Internet amounted to only about 6% of the total. This ratio is likely to hold true for any kind of sophisticated document analysis, so overhead due to network delay should not be an issue.

Given the output from OCR and a suitable ground-truth, we would ordinarily apply techniques from approximate string matching to classify errors and provide a quantifiable measure of the accuracy of the recognition process. Such an approach will not work here, however. Although we presume the ground-truth contains a reasonably reliable representation of the text on the page (a “bag of words,” if you will), we cannot be certain of the precise layout standards used by those who built the digital library. For example, a two-column page could be represented that way in the ground-truth, or it may be de-columnized. Arbitrary conventions might be employed for unrelated articles appearing on the same page. The fact that we have no guarantee there will be a correspondence between the reading orders for the OCR output and the truth, combined with the potential for large numbers of OCR errors and the need for the evaluation to be fully automated, means that string matching methods must be ruled out.

Instead, we have chosen to perform evaluation by applying a well-known measure developed in the context of information retrieval. The vector space model, first proposed by Salton, *et al.* [24], assigns large weights to terms that

[5] All tests were performed on an SGI O2 workstation (200 MHz MIPS R5000 CPU, 64 MB RAM).

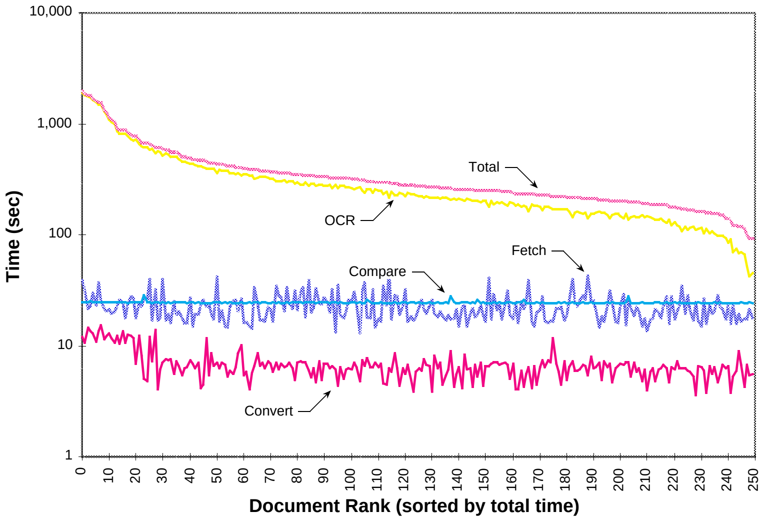


Fig. 2. Times to retrieve and process the 250 test pages used in the OCR experiment

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 collection. 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 collection. Then a common weighting scheme ($tf \times idf$) defines w_{ik}, the weight of term t_k in document D_i, to be:

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

The summation in the denominator normalizes the length of the vector so that all documents have an equal chance of being retrieved. Given query vector $Q_i = (w_{i1}, w_{i2}, \ldots, w_{iT})$ and document vector $D_j = (w_{j1}, w_{j2}, \ldots, w_{jT})$, a dot product is computed to quantify the similarity between the two. In our analysis, we apply this measure using word unigram tokens with stopword removal.

The similarity scores for the 250 test documents relative to their ground-truths are graphed in Fig. 3, sorted in order of decreasing similarity. The maximum was 0.916, the minimum 0.030, and the average 0.520. While these values may seem low, one must keep in mind several important mitigating factors: (1) the severity of the test, (2) the "ground-truth" may itself contain OCR errors, and (3) vector space similarity is not identical to OCR accuracy. A more detailed examination of the results for the 5 best and 5 worst documents, as listed in Table 1, provides subjective confirmation that this paradigm makes useful distinctions between "easy" and "hard" pages for the system under study.

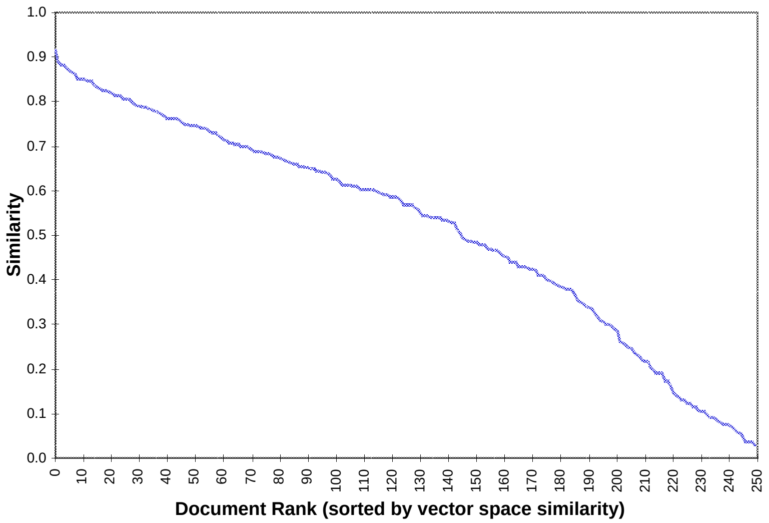


Fig. 3. Vector space similarity scores for the 250 test pages used in the OCR experiment

4.2 Table Detection

Our past work on table detection considered input in both ASCII and image format. In the latter case, we tested our techniques on a relatively small number of pages that we knew in advance contained tables. The focus was on whether the algorithm could correctly delimit the boundaries of a table and its various component regions. Another important aspect of the detection problem, however, is deciding when tables are present in an unknown input. Indeed, for many real applications this must be the first step and hence becomes a key issue.

As reported elsewhere (*e.g.*, [22]), our approach to table detection is to formulate the task of partitioning the input into tables as an optimization problem that can be solved using dynamic programming. Say that $tab[i,j]$ is a measure of our confidence when text lines i through j are interpreted as a single table. Let $merit_{pre}(i,[i+1,j])$ be the merit of prepending line i to the table extending from line $i+1$ to line j, and $merit_{app}([i,j-1],j)$ be the merit of appending line j to the table extending from line i to line $j-1$. Then:

$$tab[i,j] = \max \begin{cases} merit_{pre}(i,[i+1,j]) \; + \; tab[i+1,j] \\ tab[i,j-1] \; + \; merit_{app}([i,j-1],j) \end{cases} . \tag{2}$$

The merit functions are based on white space correlation. This defines an upper triangular matrix with values for all possible table starting and ending positions.

The partitioning of the input into tables can then be expressed as an optimization problem. Let $score[i,j]$ correspond to the best way to interpret lines i through j as some number of (*i.e.*, zero or more) tables. The computation is:

$$score[i,j] = \max \begin{cases} tab[i,j] \\ \max_{i \le k < j} \{score[i,k] \; + \; score[k+1,j]\} \end{cases} . \tag{3}$$

Table 2. Highest and lowest vector space similarity scores for the OCR experiment

Score	*http://cdl.library.cornell.edu/ cgi-bin/moa/moa-cgi?notisid=*	*Note*
Highest Vector Space Similarity Scores:		
0.916	ABP2287-0047-195	p. 766: two column layout.
0.889	ABK2934-0016-34	p. 192: two column layout.
0.883	ABR0102-0171-4	p. 97: two column layout with ruling line down gutter.
0.882	ABP2287-0042-55	p. 251: two column layout.
0.874	ABP2287-0056-192	p. 929: two columns headed by centered title and abstract, text starts with ornate drop-cap.
Lowest Vector Space Similarity Scores:		
0.045	ABR0102-0045-13	p. 661: two column layout, scan looks light, ground-truth also noisy.
0.038	ABS1821-0024-102	p. 46: three columns (newspaper format), page looks slightly skewed, irregular line spacing.
0.036	ABK4014-0008-45	p. 285: two columns, obvious skew, small font, tight spacing.
0.036	ABS1821-0006-20	p. 6: three columns, line drawing in middle of page, scan skewed and light, ground-truth also noisy.
0.030	ANU4519-0130	p. 881: two columns (index page from pension records including many proper names), sparse text, obvious skew.

The precise decomposition can be obtained by backtracking the sequence of decisions made in evaluating Eq. 3. Any region on the optimal path whose *tab* value is higher than a predetermined threshold is considered a table region.

Since our table detection procedure assumes single-column input, we used Nagy and Seth's X-Y cut algorithm [25] to segment the page images recursively, from the level of logical columns down to individual word bounding boxes.

The vast majority of pages in the *Making of America* collection contain no tables. Rather than begin with a completely random document as in the OCR experiment, we chose to search for pages that held a match for the query term "table." This yielded 103,176 hits in 33,595 works. Note that most of these still do not possess what we would call a table, since the term has many other, unrelated meanings (*e.g.*, it is an article of furniture). From this sub-collection, we selected 250 random pages and ran the X-Y cut and table detection algorithms, saving the 100 highest scores. For the steps shared with the first experiment, fetching the PDF file and converting it to TIF, the average times were comparable at

17 and 7 seconds, respectively (recall Fig. 2). The time to segment a page using X-Y cut was 31 seconds, and table detection required a little over 6 seconds.

For evaluations such as this, the familiar concepts of precision and recall are appropriate performance measures. While the former is relatively easy to compute after-the-fact (we simply need to examine each instance where the algorithm claims to have found a table), the latter requires knowing something about every document in the corpus which is not feasible when the collection is large. Hence, for now we must limit ourselves to precision measurements; these results are presented in Fig. 4. Ultimately, however, as knowledge is acquired working with the digital library, it should be possible to accumulate it for use in future tests. This "meta-data" (*e.g.*, which pages in MOA contain tables) could perhaps be published on the WWW as a supplemental index into the collection.

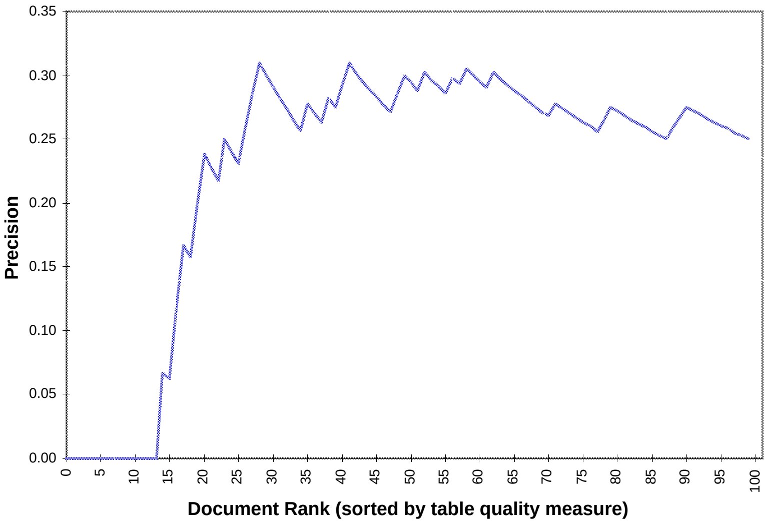


Fig. 4. Precision for the table detection experiment (top 100 hits)

Turning to the results, the 14 pages with the highest table quality measures (the value of $score[i, j]$ in Eq. 3) are false positives. In examining the documents in question (*e.g.*, http://cdl.library.cornell.edu/cgi-bin/moa/moa-cgi?notisid=ABS1821-0013-90, p. 30), we found that this was due to engraved line drawings with fine cross-hatching. While we had tuned our implementations of X-Y cut and table detection to ignore small components as noise, these made it through and generated extremely high white space correlations, thus fooling our system. Other problems were caused by page headings that use table-like spacing but are not really tables. Such scenarios might have been easy to overlook if not for the random page selection process used in the experiment.

5 Conclusions

In this paper, we have suggested that the recent phenomenon of digital libraries serving vast collections of scanned page images can be exploited by document analysis researchers, both as a target application (building better search engines) and as a way of overcoming the problem of acquiring ground-truth to support experimental investigations. We discussed some of the protocols and system-related issues involved, and offered solutions in the specific case of using the *Making of America* collection to exercise Baird's *pagereader* system and an algorithm we have developed for table detection.

It is important to reiterate that these evaluations were performed with no *a priori* knowledge of the test images or their ground-truths. The selection process was completely random, working from a very large collection. In principle, there is nothing preventing much more comprehensive experiments from being performed fully automatically, with no human intervention. At 323 seconds per page image, the entire *Making of America* collection (907,750 pages) would be sufficient to exercise *pagereader* for over 9 years running 24 hours a day and, at 62 seconds per page, our table detection code for almost 2 years.

Note that we are not advocating "attacking" digital libraries with the intention of consuming their resources and/or appropriating their content. Rather, our observation is that document analysis applied to page images drawn from such online repositories is analogous to what current search engines do when indexing the World Wide Web. The potential synergies between document analysis research and digital libraries could lead to substantial benefits for both communities.

References

1. D. Lopresti and J. Zhou. Document analysis and the World Wide Web. In *Proceedings of the Second IAPR Workshop on Document Analysis Systems*, pages 651–669, Malvern, PA, Oct. 1996.
2. D. Lopresti and J. Zhou. Locating and recognizing text in WWW images. *Information Retrieval*, 2(2/3):177–206, May 2000.
3. A. Antonacopoulos and D. Karatzas. An anthropocentric approach to text extraction from WWW images. In *Proceedings of the Fourth IAPR International Workshop on Document Analysis Systems*, pages 515–525, Rio de Janeiro, Brazil, Dec. 2000.
4. A. C. Downton, A. C. Tams, G. J. Wells, A. C. Holmes, S. M. Lucas, G. W. Beccaloni, M. J. Scoble, and G. S. Robinson. Constructing web-based legacy index card archives – architectural design issues and initial data acquisition. In *Proceedings of the Sixth International Conference on Document Analysis and Recognition*, pages 854–858, Seattle, WA, Sept. 2001.
5. O. Hitz, L. Robadey, and R. Ingold. An architecture for editing document recognition results using XML technology. In *Proceedings of the Fourth IAPR International Workshop on Document Analysis Systems*, pages 385–396, Rio de Janeiro, Brazil, Dec. 2000.

6. *International Workshop on Web Document Analysis*, Seattle, WA, Sept. 2001. http://www.csc.liv.ac.uk/~wda2001/.
7. Cornell University Prototype Digital Library. http://moa.cit.cornell.edu/.
8. Library of Congress: Digital Library Initiatives. http://memory.loc.gov/ammem/dli2/index.html.
9. I. Phillips, S. Chen, and R. Haralick. CD-ROM document database standard. In *Proceedings of Second International Conference on Document Analysis and Recognition*, pages 478–483, Tsukuba Science City, Japan, Oct. 1993.
10. About Making of America: The conversion process. http://moa.cit.cornell.edu/moa/moa_conversion.html.
11. Search result for Making of America, page 520 of *The Development of College Architecture in America* by Ashton R. Willard. http://cdl.library.cornell.edu/cgi-bin/moa/moa-cgi?notisid=AFJ3026-0022-73.
12. Search result for Princeton University Electronic Card Catalog, card 60 following the guide card Baird. http://imagecat1.princeton.edu/cgi-bin/ECC/cards.pl/disk9/0367/A4103?d=f&p=Baird&g=2000.500000&n=60&r=1.000000.
13. N. Baker. *Double Fold: Libraries and the Assault on Paper.* Random House, New York, NY, 2001.
14. E. J. Shaw and S. Blumson. Making of America: Online searching and page presentation at the University of Michigan. *D-Lib Magazine*, July/Aug. 1997. http://www.dlib.org/dlib/july97/america/07shaw.html.
15. D. G. Stork. The Open Mind Initiative. http://www.openmind.org/index.shtml.
16. M. D. Garris, S. A. Janet, and W. W. Klein. Federal Register document image database. In *Proceedings of Document Recognition and Retrieval VI (IS&T/SPIE Electronic Imaging)*, volume 3651, pages 97–108, San Jose, CA, Jan. 1999.
17. S. V. Rice, J. Kanai, and T. A. Nartker. Preparing OCR test data. Technical Report TR-93-08, UNLV Information Science Research Institute, Las Vegas, NV, June 1993.
18. CEDAR Databases. http://www.cedar.buffalo.edu/Databases/.
19. H. S. Baird. Document image defect models. In H. S. Baird, H. Bunke, and K. Yamamoto, editors, *Structured Document Image Analysis*, pages 546–556. Springer-Verlag, New York, 1992.
20. Y. Wang, I. T. Phillips, and R. Haralick. Automatic table ground truth generation and a background-analysis-based table structure extraction method. In *Proceedings of the Sixth International Conference on Document Analysis and Recognition*, pages 528–532, Seattle, WA, Sept. 2001.
21. H. S. Baird. Anatomy of a versatile page reader. *Proceedings of the IEEE*, 80(7):1059–1065, 1992.
22. J. Hu, R. Kashi, D. Lopresti, and G. Wilfong. Medium-independent table detection. In *Proceedings of Document Recognition and Retrieval VII (IS&T/SPIE Electronic Imaging)*, volume 3967, pages 291–302, San Jose, CA, Jan. 2000.
23. H. Schroeder and M. Doyle. *Interactive Web Applications with Tcl/Tk.* AP Professional, Chestnut Hill, MA, 1998.
24. 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, Nov. 1975.
25. G. Nagy and S. Seth. Hierarchical representation of optically scanned documents. In *Proceedings of the Seventh International Conference on Pattern Recognition*, pages 347–349, Montréal, Canada, July 1984.