

Book Search: Indexing the Valuable Parts

Walid Magdy

Cairo Microsoft Innovation Center
Smart Village, Bldg. B115
Km. 28, Cairo/Alex Desert Rd.
Abou Rawash, Egypt
+202 3536-3214
wmagdy@microsoft.com

Kareem Darwish

Cairo Microsoft Innovation Center
Smart Village, Bldg. B115
Km. 28, Cairo/Alex Desert Rd.
Abou Rawash, Egypt
+202 3536-3217
kareemd@microsoft.com

ABSTRACT

With massive book digitization efforts underway, there is a need for developing effective book retrieval strategies. This paper explores the relative contribution of different parts of digitized and OCR'ed books towards effective retrieval. The examined parts include the entire content of books, book headings, book titles, and table of content entries. Results show that indexing the headers and titles of books is nearly as effective as indexing the entire contents of books. These results indicate that certain portions of the books, specifically titles and headers, are more valuable than other parts of books. This is akin to web search where hypertext and page titles are more valuable to index than the rest of the webpage. Also, using a combination of evidence approach provides further improved retrieval effectiveness compared to using any portion of the book in isolation.

Categories and Subject Descriptors

H.3.1 [Content Analysis and Indexing]: *indexing methods*.

General Terms

Algorithms, Measurement, Performance, Experimentation.

Keywords

Book search; OCR retrieval.

1. INTRODUCTION

Many recent initiatives, such as Project Gutenberg¹, Google Print², the Open Content Alliance³, and the Million Book Project, have focused on digitizing and OCR'ing large repositories of legacy books [24], [15], [1]. Such initiatives have been successful in digitizing and OCR'ing millions of books in a variety of languages. One important task associated with the digitization efforts revolves around effectively finding books of interest in response to stated information needs (queries). The task of searching digitized books is potentially complicated by the OCR process, which typically introduces errors in the textual representations of books. The errors are affected by the quality of paper, printing, font, OCR training, and scanning. Arguably, the effect of OCR errors on book search would be expected to be less pronounced due to the sheer size of books (typically hundreds of pages long) leading to the repetition of

potentially valuable terms enough times, hence minimizing the possibility that valuable terms are never recognized correctly [20], [21]. Further, searching for digitized books can potentially be improved by making use of discernable book structures such as book titles, title pages, table of content (TOC) pages, indices, and headers.

This paper explores methods for effectively retrieving digitized and OCR'ed books in response to user-issued queries. Beside generalized non-book specific retrieval strategies, which are used as a baseline, the paper examines the contribution of different discernable book parts including book headings, book titles, and table of content pages towards effectively retrieving books. Like the case of web search where hypertext and titles are more valuable to index than the rest of the text in a webpage [18], one of the goals of this work is to identify the more valuable portions of books to improve retrieval effectiveness. Also, the paper examines a combination of evidence approach that combines the contribution of different parts. The experiments for this paper were performed on the 2007 INEX Book Search⁴ collection, which is described later.

Section 2 provides an overview on previous work relating to the retrieval of OCR'ed documents; Section 3 describes the experimental setup, the data collection, the IR engine, and a description of the book retrieval task runs; Section 4 reports and discusses the results; and Section 5 concludes the paper and provides possible future directions.

2. BACKGROUND

Retrieval of OCR degraded text documents has been reported on for many languages, including English [7], [10], [19], [21], [23]; Chinese [25]; and Arabic [4]. For English, Doermann [5] reports that retrieval effectiveness decrease significantly for OCR'ed documents with a word error rate between 5-20%. Taghva reported experiments which involved using English collections of scanned and OCR'ed documents that ranged in number between 204 and 674 documents and were about 38 pages long on average [20], [21]. His results show negligible decline in retrieval effectiveness due to OCR errors. Taghva's work was criticized for being done on very small collections of very long documents [8][25]. Smith reported results similar to those of Taghva [16] in which there was no significant drop in retrieval effectiveness with simulated OCR degradation. Contradicting their results, Hawking reported a significant drop in retrieval effectiveness at a 5%

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

BooksOnline '08, October 30, 2008, Napa Valley, California, USA.
Copyright 2008 ACM 978-1-60558-249-8/08/10...\$5.00.

¹ Project Gutenberg website, <http://www.gutenberg.org>

² Google Print website, <http://books.google.com>

³ Open Content Alliance website, <http://www.opencontentalliance.org>

⁴ <http://inex.is.informatik.uni-duisburg.de/2007/bookSearch.html>

character error rate on the TREC-4 “confusion track” [9]. In the TREC-4 confusion track, approximately 50,000 English documents from the federal registry were degraded by applying random edit operations to random characters in the documents [10]. The contradiction might be due to the degradation method, the size of the collection, the size of the documents, or a combination of these factors. Generally retrieval effectiveness is adversely affected by increased degradation and decreased search term redundancy in documents [6].

Several studies reported the results of using n-grams. A study by Harding, Croft, and Weir [7], compared the use of different length n-grams to words on 4 English collections, in which errors artificially introduced. The documents were degraded iteratively using a model of OCR degradation until retrieval effectiveness of using words as index terms started to significantly deteriorate, leading to unknown error rates. Indexing using a combination of 2 and 3 grams and a combination of 2, 3, 4, and 5 grams outperformed indexing using words. Similarly, Tseng and Oard experimented with different combinations of n-grams on a Chinese collection of 8,438 document images and 30 Chinese queries [25]. Although ground-truth was not available for the image collection to conclude the effect of degradation on retrieval effectiveness, the effectiveness of different index terms were compared. The results of the experiments show that a combination of unigrams and bigrams consistently and significantly outperform character bigrams, which in turn consistently and significantly outperforms character unigrams. Chinese words were not segmented and bigrams crossed word boundaries. For Arabic, [4] reported that character 3-gram and 4-grams were the best index terms for searching OCR degraded text. They conducted their experiments on a small collection of 2,730 scanned documents. In general, blind relevance feedback does not help for the retrieval of OCR degraded documents [3], [11], [22], [25]. For some IR documents, weighting different portions of documents differently may lead to improved retrieval effectiveness, as was demonstrated in web search [18] and was attempted for book search [26].

3. EXPERIMENTAL SETUP

3.1 Data Collection and Used Search Toolkit

The collection used for all experiments was the collection used for the INEX 2007 Books Search Track. The collection, provided to the track by Microsoft Live Book Search and the Internet Archive, consisted of 42,049 digitized out-of-copyright books, with books typically being printed before the 1930’s. The actual number of books used was 41,825, where 224 books were missed due to extraction errors or empty content books. The OCR’ed content of the books was stored in djvu.xml format, which provides xml fields specifying pages, paragraphs, lines, and words along with their coordinates in the page. Associated with each book were other metadata such as name(s) of author(s), publisher, publication data, Library of Congress classification, etc.

The focus of the experiments was to help users identify books of interest based on stated information needs. There were 250 queries, with associated relevance judgments, about general subjects: typically consisting of 1 word and commonly containing named entities. Two sample queries were: “Botany” and “Rigveda.” The relevance judgments were done on a scale from 0 to 4, with 0 = bad, 1 = fair, 2 = good, 3 = excellent, and 4 = perfect. The figure of merit used was Normalized Discounted Cumulative Gain (NDCG), which is a metric that is becoming

increasing popular for evaluating web search engines [26]. The computation of NDCG was done as described by [12]. NDCG attempts to compute the information gained by the user when (s)he reads the top n results, with documents with higher scores portraying more information gain. NDCG was computed at using top n results, where n were selected to be 1-10, 100, and 1000 top results. For all runs, Indri search toolkit was used for indexing and searching the collection of books. Indri was used with stop-word removal and no stemming or blind relevance feedback. A number of previous studies showed that blind relevance feedback is ineffective for OCR’ed documents. Indri combines inference network model with language modeling [13]. A series of experiments were performed as follows:

Non-book specific setups (baseline):

1. Book Content (BC): Each document corresponded to the entire content of a single book. The documents were subsequently indexed and searched using the provided queries. This run assumes no structure in the book.

Book specific setups:

1. Book Headings (BH): Each document was composed of all the chapter and section headings in a book, which were assumed to be the first line in each page not composed entirely of digits (to skip page numbers). The documents were indexed and searched using the provided queries. Headers are believed to reflect the main topics with the proper term frequencies.
2. Table of Content (TOC): Each document was composed of the TOC and index pages in a book. A page was deemed a TOC or index page if any of the following conditions are met:
 - i. Presence of the key phrase “Table of Contents.”
 - ii. Presence of ordinary key words such as “contents”, “page”, or “index”, with moderate number of lines that end with digits.
 - iii. Absence of keywords indicating a TOC or index page, but the presence of a large number of lines ending with digits.
 - iv. Presence of keywords such as contents, page, or index in a page that was immediately preceded by a page that was deemed as a TOC or index page.

In case no pages met these conditions, the first 3,000 characters from the OCR output and last 10 pages of a book were used instead, as they are likely to contain TOC and index pages. The first 3,000 characters were used instead of a fixed number of pages because many books contained empty pages in the beginning. Detecting TOC and index pages in this manner was not very accurate and improved detection techniques is warranted. The intuition behind using the TOC and index pages is that they generally summarize book topics.

3. Book Title (BT): Each document corresponded to the title of a single book, where titles were obtained from the manually entered book metadata. The book titles were specifically chosen, because the book titles are actual parts books not externally added, and they are often telling of book contents.

3.2 Combination of Evidence Setups:

Given the scores obtained from the search engine from each experiment for each query-document pair in the results, a new score S_{Total} was computed for each query-document pair using the following formula: $S_{Total} = w_0S_0 + w_1S_1 + \dots + w_nS_n$

Where $S_0 \dots S_n$ are the scores from different experiments for each pair and the $w_0 \dots w_n$ are the weights assigned to each, where the

sum of all w 's is 1 and the values of w 's was changed in increments of 0.1. The tested combinations are:

1. BH+TOC+BT: Combining score from BH, BT, and TOC experiments.
2. BC+BH+TOC+BT: Combining scores from BH, BT, and TOC along with those from BC.

The weights $w_0 \dots w_n$ were obtained using 2-fold cross validation where the weights were optimized for the training half of queries and the obtained weights were applied on the other test half of the queries. Luckily, the weights obtained from each half were nearly identical. For significance testing, a paired 2-tailed t-test was used with p-value < 0.05 to indicate significance testing. Although, a t-test assumes a parametric distribution, different studies suggest that the t-test aptly distinguishes between IR runs [14], [17].

Table 1. p -values of t-test comparing NDCG @ [1, 5, 10, 100] for BC vs. other experiments. Black and grey cells indicate statistically significantly better or worse than BC respectively.

	NDCG @			
	1	5	10	100
BH	0.62	0.00	0.00	0.00
BT	0.97	0.01	0.00	0.00
BC+BH+TOC+BT	0.00	0.02	0.12	0.69
BH+TOC+BT	0.64	0.37	0.02	0.00

4. RESULTS AND DISCUSSION

Figure 1 shows the average NDCG scores for different retrieval experiments. Tables 1, 2, and 3 report the p-value of the t-tests comparing different experiments. Results show that indexing the table of content pages yields the worst retrieval effectiveness. Further, the best weighted w associated with the TOC scores for the combination of evidence experiments was consistently 0, meaning any inclusion of TOC scores hurt retrieval effectiveness. This result seems peculiar especially that TOC pages generally contain all the topics in a book. The authors' intuition is either that the employed method for finding TOC pages was inadequate or TOC pages list topics once, which leads to losing term frequency information. This requires further investigation. Indexing the entire book content (BC) was statistically significantly better than using the just the titles (BT) or just the book headings (BH) except for NDCG @ 1. Drawing firm conclusions from experiments using NDCG @ 1 scores, even when using significance testing, would probably require more queries than 250. The relative retrieval effectiveness between BT and BH experiments on one hand and the BC experiments on the other is 80-100% and the resulting indices are much smaller. Perhaps, the book headings did well because the redundancy of headers provides more accurate term frequency estimates.

Combining the scores for BH ($w_{BH} = 0.3$), TOC ($w_{TOC} = 0.0$), and BT ($w_{BT} = 0.7$) produced statistically indistinguishable retrieval effectiveness for NDCG @ 1 and 5 compared to BC, but not for NDCG at 10 and 100. Given, that most users tend to look at the top few results, this result is significant because indexing just BH and BT yields a significantly smaller index compared to indexing the entire book and would lead to reduced indexing and retrieval times. The ratio in index size between the two is roughly 1:20. Lastly combining BC ($w_{BC} = 0.2$), BH ($w_{BH} = 0.1$), TOC ($w_{TOC} = 0.0$), and BT ($w_{BT} = 0.7$) which had the effect of giving more weight to certain portions of the book lead to statistically

significantly improved retrieval effectiveness compared to using BC alone (for NDCG @ 1 and 5). In fact, the best weighting scheme was achieved when most of the weight was given to the BT scores.

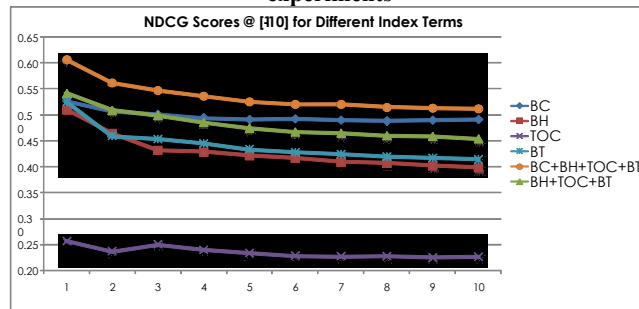
Table 2. p -values of t-test comparing NDCG @ [1, 5, 10, 100] for BH vs. other combination of evidence experiments. Black and grey cells indicate statistically significantly better or worse than BC respectively.

	NDCG @			
	1	5	10	100
BC+BH+TOC+BT	0.00	0.00	0.00	0.00
BH+TOC+BT	0.12	0.00	0.00	0.00

Table 3. p -values of t-test comparing NDCG @ [1, 5, 10, 100] for BT vs. other combination of evidence experiments. Black and grey cells indicate statistically significantly better or worse than BC respectively.

	NDCG @			
	1	5	10	100
BC+BH+TOC+BT	0.00	0.00	0.00	0.00
BH+TOC+BT	0.55	0.02	0.01	0.00

Figure 1. Average NDCG @ [1-10] for different experiments



5. CONCLUSION AND FUTURE WORK

This paper examined the effect of indexing different parts of digitized books on the retrieval of such books in response to information needs. The paper was concerned with determining the effect of indexing different parts of the book compared to indexing the entire book and with indexing entire book contents but while giving more weight to particular portions (via a combination of evidence approach). Although portions of books, namely book headings and titles, generally led to deterioration in retrieval effectiveness, the drop in retrieval effectiveness was less than 20% while the reduction in index size was more than 95% and 99.9% for using book headings and book titles respectively. Contrary to the initial intuition of the authors; indexing TOC and index pages only led to significant degradation in retrieval effectiveness. This result could be due to poor TOC and index page identification, poor term frequency estimates from TOC's, or both. This matter warrants further investigation. When indexing entire book contents, giving more weight to book titles and headings yields statistically significant improvement in retrieval effectiveness. This affirms what was discovered for other IR applications where giving more weight to certain portions of documents yields improved retrieval effectiveness. This result

was observed for web search where retrieval effectiveness is improved when hypertext is given more weight.

For future work, the effect of other portions of books on retrieval needs to be examined including reexamining the TOC and index pages. Better identification of TOC and index pages can determine conclusively if the poor results for indexing TOC and index pages were due to the TOC and index page identification technique or some other factors. Further, incorporation of relative weights to different portions of the books inside the ranking formula [26] warrants attention. Lastly, search for sub-book units, such as pages, is perhaps an important problem as users may want to search for a particular piece of information and not necessarily a book to read.

6. REFERENCES

- [1] Barret, W., L. Hutchison, D. Quass, H. Nielson, and D. Kennard. Digital Mountain: From Granite Archive to Global Access," Proc. of International Workshop on Document Image Analysis for Libraries, Palo Alto, January 2004, pp. 104-121, (2004).
- [2] Croft, W. B., S. Harding, K. Taghva, and J. Andborsak. An evaluation of information retrieval accuracy with simulated OCR output. In Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval, University of Nevada, Las Vegas, Nev., 115-126, (1994).
- [3] Darwish, K. and O. Emam. The Effect of Blind Relevance Feedback on a New Arabic OCR Degraded Text Collection. In International Conference on Machine Intelligence: Special Session on Arabic Document Image Analysis, (2005).
- [4] Darwish, K. and D. Oard. Term Selection for Searching Printed Arabic. In SIGIR-2002, pp 261-268 (2002).
- [5] Doerman, D. The Retrieval of Document Images: A Brief Survey. ICDAR, pp 945-949 (1997).
- [6] Doermann, D. The Indexing and Retrieval of Document Images: A Survey. Computer Vision and Image Understanding, 70(3): pp 287-298 (1998).
- [7] Harding, S., W. Croft, and C. Weir. Probabilistic Retrieval of OCR-degraded Text Using N-Grams. In European Conference on Digital Libraries, pp 345 - 359 (1997).
- [8] Harman, D. Overview of the First Text REtrieval Conference. Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, Pittsburgh, Pennsylvania, United States, pp 36 - 47 (1992).
- [9] Hawking, D. Document Retrieval in OCR-Scanned Text. Sixth Parallel Computing Workshop, paper P2-F (1996).
- [10] Kantor, P. and E. Voorhees. Report on the TREC-5 Confusion Track. TREC-5, pp 65, (1996).
- [11] Lam-Adesina, A. M. and G. J. Jones. Examining and improving the effectiveness of relevance feedback for retrieval of scanned text documents. Inf. Process. Manage. Vol. 42 (3) pp. 633-649, (2006).
- [12] Matveeva, I., C. Burges, T. Burkard, A. Laucius and L. Wong. High accuracy retrieval with multiple nested rankers. SIGIR 2006 (2006).
- [13] Metzler, D. and W. B. Croft. Combining the Language Model and Inference Network Approaches to Retrieval. Information Processing and Management Special Issue on Bayesian Networks and Information Retrieval, 40(5), 735-750, (2004).
- [14] Sanderson, M., J. Zobel, Information retrieval system evaluation: effort, sensitivity, and reliability. SIGIR 2005: 162-169, (2005).
- [15] Simske, S. and X. Lin. Creating Digital Libraries: Content Generation and Re-mastering. Proc. International Workshop on Document Image Analysis for Libraries, Palo Alto, January 2004, pp. 33-45, (2004).
- [16] Smith, S., An Analysis of the Effects of Data Corruption on Text Retrieval Performance. Technical Report DR90-1, Thinking Machines Corp: Cambridge, MA, (1990).
- [17] Smucker, M., J. Allan, and B. Carterette, A Comparison of Statistical Significance Tests for Information Retrieval Evaluation, CIKM '07, Lisboa, Portugal, (2007).
- [18] Song, R., J.R. Wen, S. Shi, G. Xin, T.Y. Liu, et al., Microsoft Research Asia at Web Track and Terabyte Track of TREC 2004. In 2004 Text REtrieval Conference (2004).
- [19] Taghva, K., J. Borsack, and A. Condit. An Expert System for Automatically Correcting OCR Output. Proc. IS&T/SPIE 1994 Intl. Symp. on Electronic Imaging Science and Technology, , San Jose, CA, pp 270-278 (1994a).
- [20] Taghva, K., J. Borasack, A. Condit, and J. Gilbreth. Results and Implications of the Noisy Data Projects. Technical Report 94-01, Information Science Research Institute, University of Nevada, Las Vegas, (1994b).
- [21] Taghva, K., J. Borasack, A. Condit, and P. Inaparthi. Querying Short OCR'd Documents. Technical Report 94-10, Information Science Research Institute, University of Nevada, Las Vegas, (1995).
- [22] Taghva, K., Borsack, J., & Condit A. Evaluation of Model-Based Retrieval Effectiveness OCR Text. ACM Transactions on Information Systems, 14(1):64-93, (1996a).
- [23] Taghva, K., Borsack, J., & Condit, A. Effects of OCR errors on Ranking and Feedback using the Vector Space Model. Information Processing and Management, 32(3):317-327, (1996b).
- [24] Thoma, G. and G. Ford. Automated Data Entry System: Performance Issues. Proc. SPIE Conference on Document Recognition and Retrieval IX, San Jose, 2002, pp. 181-190, (2002).
- [25] Tseng, Y. and D. Oard. Document Image Retrieval Techniques for Chinese. In Symposium on Document Image Understanding Technology, Columbia, MD, pp 151-158 (2001).
- [26] Voorhees, E. Evaluation by highly relevant documents. In Proceedings of SIGIR, pages 74–82, (2001).
- [27] Wu, H., G. Kazai, and M. Taylor, Book Search Experiments: Investigating IR Methods for the Indexing and Retrieval of Books. ECIR 2008: 234-245 (2008)