

A Web Service for Long Tail Book Publishing

Prakash Reddy
HP Labs
1501 Pagemill Ave
Palo Alto, Ca 94304
(650) 859-2790

Prakash.reddy@hp.com

Jian Fan
HP Labs
1501 Pagemill Ave
Palo Alto, Ca 94304
(650) 857-2554

Jian.fan@hp.com

Jim Rowson
HP - IPG
3000 Hanover St
Palo Alto, Ca 94304
(650) 857-7267

Jim.Rowson@hp.com

Steven Rosenberg
HP Labs
1501 Pagemill Ave
Palo Alto, Ca 94304
(650) 857-5902

Steven.Rosenberg@hp.com

Andrew Bolwell
Corporate Ventures
3000 Hanover St
Palo Alto, Ca 94304
(831) 427-1291

Andrew.Bolwell@hp.com

ABSTRACT

More than 32M unique book titles are available in US libraries, but Amazon, the biggest retailer, had only 1.2M unique titles available for sale in 2004. Currently there is an effort underway by public libraries, universities, the Open Content Alliance, Google and others, to non-destructively scan these 32M unique books and make them available for on-line viewing and search. Twenty percent (6.4M) of the 32M titles are out of copyright and out of print. A publisher estimates that an average of 40 copies of each title can be sold per year if they could be made available for sale. This long tail opportunity represents a several billion dollar market with the right cost structure. To address this long-tail book market we need to take the cost out of several parts of the value chain: automatic book preparation to minimize publishing setup costs, print-on-demand to remove warehouse and waste costs, and web 2.0 techniques to minimize marketing costs. We have created this system with several partners based on HP technology, and available as an incubation business.

ACM Categories & Subject Descriptors:

I.7.4 Computing Methodologies, Document and Text Processing, Electronic Publishing

General Terms:

Algorithms

Keywords:

edge lines, page boundary, print on demand, scanning

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. PROBLEM STATEMENT

Chris Anderson, the author of "The Long tail", points out that the demand curve never drops to zero. The biggest money is often in the aggregate of small sales under the "long tail" portion of the demand curve. At Barnes & Noble, a brick and mortar store carries 130,000 unique titles. On the other hand, more than half of the books sold on Amazon are not from the top 130,000 titles. According to Nielsen Bookscan, Amazon itself only carries 1.2M unique titles, which represents just 4% of the 32M titles available from various US libraries. Approximately 20% of these available titles are out of copyright and out of print. Today, the cost to republish these books is prohibitive when each book only sells an average of 40 copies. The challenge is to make these books available for publishing at an acceptable cost.

2. OUR SOLUTION

The traditional value-chain for book publishing involves heavy investment in content preparation (buying reprint rights, editing, preparing for print), in production (speculative print runs that must be warehoused and potentially scrapped if not sold), and marketing (advertising, book tours, revenue sharing with retail channel). Most of these steps in the value chain must be rethought for long tail books. Through a combination of technology and business innovation, it is possible to radically reduce the costs in each of these areas, making it feasible to address this market.

3. AUTOMATIC BOOK PREPARATION

A vast majority of the out of print and out of copyright books are in libraries and primarily collecting dust. Fortunately, efforts are underway by the Open Content Alliance, Universal Digital Library, Google and others to digitally scan these books, bringing them online. These books are scanned using high resolution photography, to avoid destroying the originals. However, this photographic scanning produces page images that are not directly useful for print because of lighting, alignment, scan artifacts, as well as aging and wear and tear from use.

Preparing books for Publish-On-Demand (POD) from scanned images involves solving and automating a variety of technical steps, such as detecting page boundaries, adjusting for lighting,

removing background noise, deskewing, text sharpening and formatting into a form that is readily acceptable to a print service provider (PSP).

3.1 System overview

We have developed an hosted web service, which downloads the scanned book pages, and cost-efficiently processes them to produce print quality images. Furthermore, we've automated the production of a final book package for delivery to a print-on-demand supplier. Currently we support scanned books from the Internet Archive (www.archive.org), Universal Digital Library (www.ulib.org), and scans produced from several commercial scanning providers.

The system has four primary components a) Content acquisition b) Processing pipeline c) Content packager and d) the User Interface, which gives publishers the ability to make minor edits to the processed content and control the flow.

3.2 Content Acquisition

This module is responsible for acquiring content from a variety of sources and normalizing the representation for downstream processing. This has been designed to support a variety of scanned sources. Plug-ins can be developed and integrated to support additional scan sources. The most common practice is to download the content (raw scanned images) and the associated meta data using a File Transfer Protocol (FTP). This would be simpler if there existed a standard on how to package the raw content.

3.3 Processing pipeline

The stage is where we turn raw, photographically scanned page images into print-ready content through conversion, rotation, margin detection/cropping, and image/text enhancement.

One of the key problems in preparing scanned images for printing is detecting page boundaries on the scanned images and finding a crop region that can be applied to all the pages in a book. To detect page boundary we obtain the gradient, detect lines, determine the page edges, determine skew and finally crop at boundary and skew. See Fig 1.

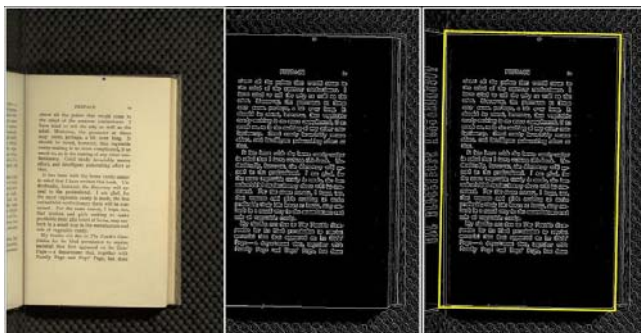


Figure 1: Shows raw scanned image (left), image with edge lines detected (middle) and detected page boundary yellow rectangle (right)

Finding the page boundary accurately is critical to producing acceptable quality output. We combine the individually detected page data and compute a uniform page size to account for scanning idiosyncracies.



Figure 2: Shows page margin detection and cropping

Once the pages are cropped, we run the images through a set of enhancement algorithms, where each image is deskewed if necessary, text is sharpened, and images are adjusted. Color conversion also happens at this stage. Fig 2 shows examples of original images and processed images after color/illumination correction and text sharpening.

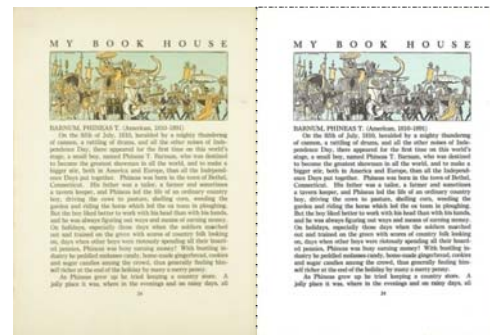


Figure 3: Shows color/illumination correction and text sharpening

3.4 Content Packaging

Once the book block has been processed it needs to be packaged into a form that is print ready. This involves formatting the processed content to the appropriate trim size, generating covers, managing ISBNs and generating pricing/barcode information. Our system has flexibility to support a variety of trim sizes and binding options. The cover generation is based on a flexible and extensible template system.

3.5 Publisher tools

Our system provides support to publishers to manage their ISBN numbers, define cover templates, and generate copyright/title pages. In addition we have tools which allow publishers to preview, edit and approve processed books.

3.6 Fulfillment service

For POD to succeed, we need service providers who can fulfill the generated demand. It has been estimated that 50% of all books printed are scrapped due to poor estimation of demand. All these books must be warehoused and shipped to bookstores. The wasted books must also be shipped back from bookstores and destroyed. All these costs are prohibitive for long-tail books.

We have designed our Publish On Demand book service around print-on-demand suppliers, to eliminate warehouse and waste costs of traditional book production processes.

3.8 Marketing costs

Traditional bookstores and even online stores demand a 50% markup. However, new technologies on the web have shown some potentially different and cheaper ways to market these niche long tail books. The web has splintered into millions of specialized websites and communities, each with its own highly involved audience. User created content, social recommendations, and other "web 2.0" techniques can be used to potentially market niche, long-tail books more effectively than other outlets.

One of the approaches we have taken is "social retail" a branded site that can be used to build highly focused communities where the books are available for sale comment, recommendation and to be mixed with user created content.

4. PROCESSING PERFORMANCE

Publishers today use a manual approach for preparing books for the POD market. Skilled designers take as input scanned images and make heavy use of editing tools like InDesign from Adobe to prepare a book. Each page is manually cropped, rotated as needed, color and background is corrected, content is deskewed and a final PDF for the book is generated. It typically takes an experienced person on the order of 1-2 days to prepare a book. This does not include the effort required to design the cover and package and send the content off to a print service provider. Our system has automated all these steps. We have provided some interactive tools to publishers to edit processed books as and when needed. It takes an average of 4-5 hours to prepare a 300 page book. Processing destructively scanned books takes less time than non-destructively scanned books. The system has been architected to scale up/down based on demand.

5. STATUS AND FUTURE WORK

The goal of making every book ever published available to anyone at anytime in a form that is convenient poses several challenges. We have built a system that demonstrates an end to end solution that takes raw book scans and make them available for online viewing as well as on demand printing. We have been able to resolve conflicting requirements for online and print media formatting. In the future it is possible to retarget the processed books for ebook readers. We have successfully deployed the system and are actively working with a publisher and a few

libraries. Several of the books processed by our system are offered for sale thru various channels.

As a result of building this system we have certainly uncovered several challenges. These challenges provide an opportunity to innovate and collaborate.

5.1 Unified discovery

One of the biggest challenges is to allow users to discover content that resides in different digital repositories. Currently there is no single repository of all the books that have been scanned and available for consumption. It would make it easy if we had a unified meta-data repository of all books that have been processed. This would allow people to easily find books of interest. This does not require that the actual book content be made available but simply the meta-data about each book. This also requires defining a standard for representing meta-data. Libraries do have some standards and we would be better served if we were adopt such a standard.

5.2 Processing quality

The quality of processed books depends heavily on the quality of the scan and the algorithms applied in the cleaning process. Some specific issues we have encountered while cleaning up books include a) identifying noise b) dealing with page curl effects c) identifying missing pages, d) accurately detecting page boundary all the time.

If the scans are not cleaned up properly it could adversely affect the quality of the OCR which in turn would affect readability and ability to search. It certainly affects the printed quality of books.

Crowd sourcing is one of the techniques that can be employed to identify and improve the quality.

We are working on a system which allows adding confidence factors at every processing step. The lower the confidence factor the higher the probability that the book needs special attention. The hope is to minimize the number of pages that need to be inspected.

Currently we have an accept rate of 85% of the books we process.

5.3 Book composition

The large indexed searchable digital book repositories open up possibilities for composing a book by selecting logical units from multiple books. Combining personal content with professional content becomes technically feasible. However this poses several challenges in terms of copyrights, proper compensation, format consistencies etc.

5.4 Book recommendation

A book recommendation system based on understanding the content and combining it with the library classification data and the Book Industry Standards and communications (BISAC) will help achieve the goal of reaching the target audience for these books.

REFERENCES

Web Sources

www.archive.org

www.ulib.org

www.foodsville.com

- [1] Chris Anderson, "The Long Tail" Wired magazine, Issue 12.10 Oct 2004
<http://www.wired.com/wired/archive/12.10/tail.html>
- [2] Jian Fan, "A local orientation coherency weighted color gradient for edge detection", ICIP05, p.1132-1135, Sept. 2005
- [3] Jian Fan, Xiaofan Lin, Steven Simske, "A comprehensive image processing suite for book re-mastering", ICDAR05, p.447-451, Aug. 2005
- [4] Jian Fan, "Enhancement of Camera-captured Document Images with Watershed Segmentation", CBDAR07, p.87-93, Sept.