[Proc. International Conference on Document Analysis and Recognition, ICDAR'2007, Curitiba, Brazil, Sep. 2007]

Google Book Search: Document Understanding on a Massive Scale

Luc Vincent
Google, Inc.
1600 Amphitheatre Parkway, Mountain View CA 94043, USA
luc@google.com

Abstract

Unveiled in late 2004, Google Book Search *is an ambitious program to make all the world's books discoverable online. The sheer scale of the problem brings a number of unique document analysis and understanding challenges that are outlined in this paper. We also go over some of the ways that Google is working with the document analysis research community to help push the state of the art.*

Google Book Search started as a dream: the dream by Google co-founders Larry Page and Sergey Brin that one day, all the world's books might be discoverable online, through a kind of electronic and distributed Library of Alexandria. Today, thanks to advances in scanning and storage technology, to significant investment by Google and to a lot of hard work by a lot of dedicated Googlers and partners, this dream is well on its way to becoming a reality. Google Book Search, originally named *Google Print,* was launched in late 2004. Over a million books have been scanned to date and Google Book Search is already the largest "Virtual Card Catalog" in existence.

Google Book Search is a huge endeavor: estimates for the number of books ever published range from 50 millions to 200 millions, and well over 100,000 new books are published annually. Making a significant dent at scanning or acquiring this content in electronic form has required a multifaceted aproach. First, partnerships with publishers were established, allowing Gogle to get access to mostly recent books directly from publishers. Second, our Library Program was put in place and Google entered into agreements with a number of libraries to scan their collections. Four major libraries joined the program when it was unveiled over two years ago and to date, over twenty libraries have signed up. More details are available at *http://books.google.com/googlebooks/partners.html.*

While daunting, the task of digitizing such massive numbers of books is only part of the challenge. In order for the wealth of information contained in these volumes to be made accessible and useful, it needs to be organized and indexed. One of the ways this is done is through optical character recognition (OCR), that is, the process of turning book pages into indexable text. While OCR is a fairly mature technology, the scale of the problem brings unique challenges that available OCR packages are relatively ill-equipped to handle. Chief among them is the fact that because of its global ambitions, Google Book Search requires handling books in *all* the world's scripts and languages! Furthermore, the language or even script is frequently unknown prior to OCR, and in fact, a substantial number of books contain multiple scripts or languages (Chinese and English, or English and French are common combinations). In contrast, commonly used OCR packages typically require that language information be provided beforehand. Additionally, these packages are tweaked to work optimally on documents not exceeding a few pages and to not spend more than a few seconds per page. As such, they fail to capitalize on shape and style redundancy typically found in books across hundreds of pages, redundancy that can take a fair amount of CPU cycles to exploit.

In addition to OCR, making these books easily accessible and useful on *http://books.google.com* has required developing a number of additional state-of-the-art systems. These include systems for automatically deskewing, cropping and cleaning-up scanned book pages, which is critical as pre-processing prior to OCR, but also to generate clean and small images for efficient web serving. While this may be a well understood problem for high-quality documents, doing this well on scanned century-old book pages is no small feat. Most of the advanced systems developed for

Google Book Search however involve some form of Document Understanding and as such, come after OCR in the book processing pipeline. Systems that have been developed, are being developed or are being considered as interesting research challenges include:

Page ordering: errors in scanning or limitations in scanning technology can occasionally result in the pages of a book being somewhat out of order, or some pages missing, or some being duplicated. The challenge is to put complete volumes back together fully automatically, even when page numbers are missing altogether, or in the presence of OCR errors.

Language identification: often useful to extract at the page level in order to improve search quality

Chapter detection: the cornerstone of our book-level understanding, provides information essential to book navigation.

Content linking: associate table of content entries to pages, paragraph or chapter boundaries, link index words to pages where they appear, etc.

Summarization: automatically generate summaries of various lengths (3 lines, a paragraph, a page)

Metadata extraction and cross-validation: extract book title, author, publisher, edition, publication year, etc. and cross-validate with bibliographic metadata that might otherwise be available

Topic identification: extract main topic(s) of a book from less than perfect OCR output

Book clustering and linking: create relationships between volumes by grouping books by author, topic, publisher, etc.

The wealth and diversity of volumes now available in Google Book Search make it a unique playground for document understanding scientists. In fact, we have probably only scratched the surface in terms of interesting research challenges that this program enables.

Beyond document image processing, OCR, volume level understanding and indexing, another important topic has kept the Google Book Search team busy, namely ranking. Specifically, how should books that match a particular query be ranked? The web is notorious for its rich graph of hyperlinks, famously exploited by Google' PageRank algorithm [6]. This structure applies somewhat to technical publications, which typically contain numerous references to other technical publications. However the universe of books is different and most books (eg, novels) do not contain any references. Novel approaches therefore had to be developed, exploiting an array of new signals. Additionally, these techniques were recently extended to allow "blending" of book search results with web search resuts when appropriate.

Ranking is only part of the equation when returning book search results to users. Another thorny issue is copyright: the copyright status of each book determines how much of the book we can show and the kind of user experience we are able to provide. Not respecting copyright is oviously not an option. So how do we deal with this? As explained in [7], each book includes an "About this book" page with basic bibliographic data like title, author, publication date, length and subject. For some books additional information is available, such as key terms and phrases, references to the book from scholarly publications or other books, chapter titles and a list of related books. After that, books fall into four categories, summarized in Figs. 1 and 2:

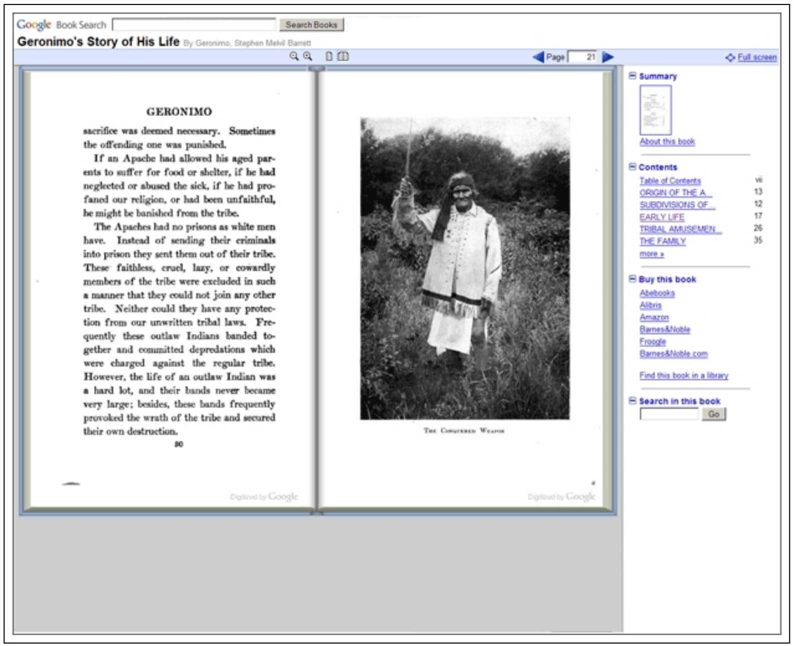
Full view: If we have determined that a book is out of copyright, or the publisher or rightsholder has given Google permission, users are able to page through the entire book from start to finish, as many times as they like. If the book is in the public domain, users are also able to download a PDF version of the book.

Limited preview: If a publisher or author has joined Google Books Partner Program, a few full pages from the book are available as a preview. Searches within the book can also be conducted.

Snippet view: For books which may still be in copyright, no full page is ever shown to users. However searching within the book is possible and for each search term, up to three snippets of text from the book are displayed, showing search term in context. While this is limited, as always, we provide links to places where the book can be purchased or borrowed.

No preview available: In some cases, we limit the display to metadata, that is, only the 'About the book' page is still available, along with pointers to bookstores or libraries that may carry this book.

Google Book Search is still in it infancy, but the program is already proving to be an indispensable tool to users, authors and publishers alike. Users love to be able to search over the universe of books as conveniently as they search the web; authors appreciate the exposure their works are given through the program; and many publishers have seen an uptick in sales, especially for some older, hard to find volumes, i.e., the famous "long tail" [1]. The Document Image Analysis and Understanding community has also arguably been somwewhat revitalized by Google Book Search, which has generated a lot of excitement among document analysis researchers and provided new research challenges and directions: document understanding and analysis on a truly



Google Book Search [Search Books]

Geronimo's Story of His Life By Geronimo, Stephen Melvil Barrett

Page 21 ◇ Full screen

GERONIMO

sacrifice was deemed necessary. Sometimes the offending one was punished.

If an Apache had allowed his aged parents to suffer for food or shelter, if he had neglected or abused the sick, if he had profaned our religion, or had been unfaithful, he might be banished from the tribe.

The Apaches had no prisons as white men have. Instead of sending their criminals into prison they sent them out of their tribe. These faithless, cruel, lazy, or cowardly members of the tribe were excluded in such a manner that they could not join any other tribe. Neither could they have any protection from our unwritten tribal laws. Frequently these outlaw Indians banded together and committed depredations which were charged against the regular tribe. However, the life of an outlaw Indian was a hard lot, and their bands never became very large; besides, these bands frequently provoked the wrath of the tribe and secured their own destruction.

80

Digitized by Google

THE CONQUERED WEAPON

Digitized by Google

⊟ Summary

About this book

⊟ Contents

Table of Contents	vii
ORIGIN OF THE A...	13
SUBDIVISIONS OF...	12
EARLY LIFE	17
TRIBAL AMUSEMEN...	26
THE FAMILY	35
more »	

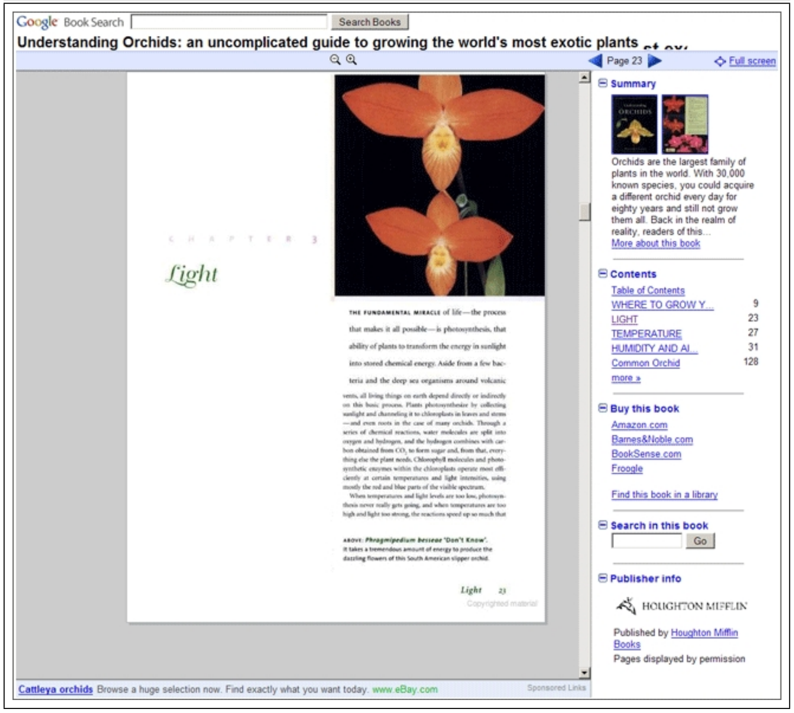
⊟ Buy this book

Abebooks
Alibris
Amazon
Barnes&Noble
Froogle
Barnes&Noble.com

Find this book in a library

⊟ Search in this book

[Go]



Google Book Search [Search Books]

Understanding Orchids: an uncomplicated guide to growing the world's most exotic plants

Page 23 ◇ Full screen

CHAPTER 3

Light

THE FUNDAMENTAL MIRACLE of life—the process that makes it all possible—is photosynthesis, that ability of plants to transform the energy in sunlight into stored chemical energy. Aside from a few bacteria and the deep sea organisms around volcanic vents, all living things on earth depend directly or indirectly on this basic process. Plants photosynthesize by collecting sunlight and channeling it to chloroplasts in leaves and stems—and even roots in the case of many orchids. Through a series of chemical reactions, water molecules are split into oxygen and hydrogen, and the hydrogen combines with carbon obtained from CO_2 to form sugar and, from that, everything else the plant needs. Chlorophyll molecules and photosynthetic enzymes within the chloroplasts operate most efficiently at certain temperatures and light intensities, using mostly the red and blue parts of the visible spectrum.

When temperatures and light levels are too low, photosynthesis never really gets going, and when temperatures are too high and light too strong, the reactions speed up so much that

ABOVE: *Phragmipedium besseae* 'Don't Know'. It takes a tremendous amount of energy to produce the dazzling flowers of this South American slipper orchid.

Light 23

Copyrighted material

⊟ Summary

Orchids are the largest family of plants in the world. With 30,000 known species, you could acquire a different orchid every day for eighty years and still not grow them all. Back in the realm of reality, readers of this...
More about this book

⊟ Contents

Table of Contents	
WHERE TO GROW Y...	9
LIGHT	23
TEMPERATURE	27
HUMIDITY AND AI...	31
Common Orchid	128
more »	

⊟ Buy this book

Amazon.com
Barnes&Noble.com
BookSense.com
Froogle

Find this book in a library

⊟ Search in this book

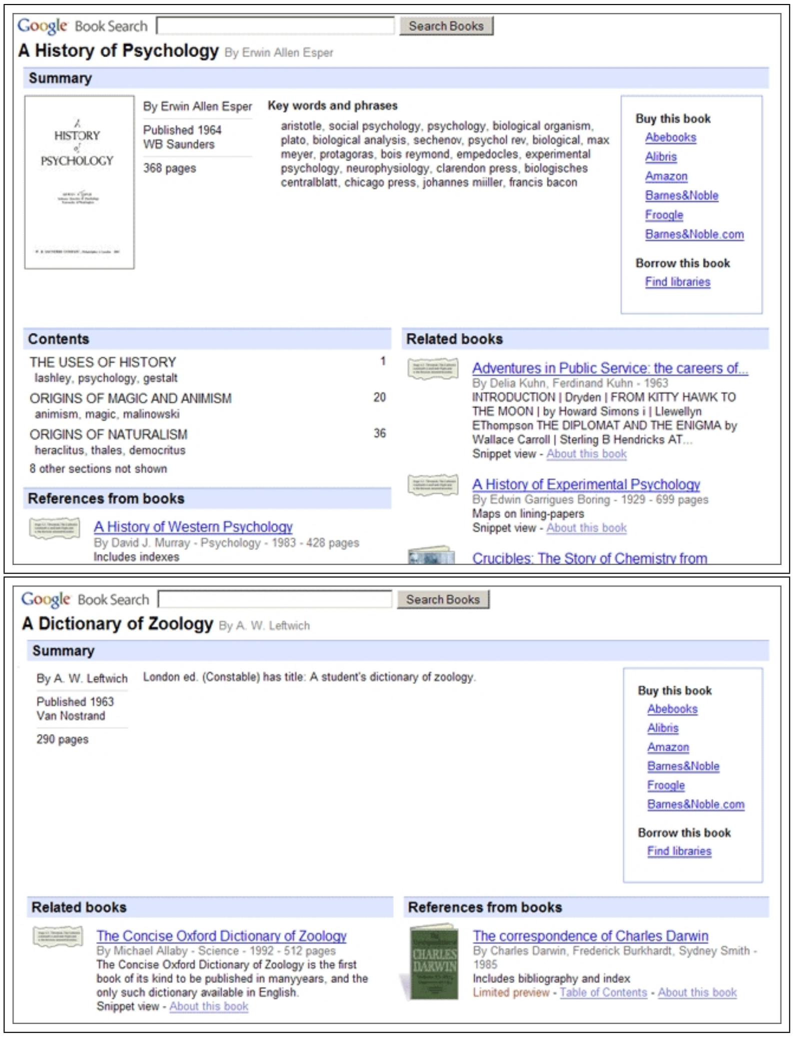
[Go]

⊟ Publisher info

HOUGHTON MIFFLIN

Published by Houghton Mifflin Books
Pages displayed by permission

Figure 1. Four different user experiences with Google Book Search, 1 of 2: Full View (top), Limited Preview (bottom).



A History of Psychology By Erwin Allen Esper

Summary

By Erwin Allen Esper

Published 1964
WB Saunders

368 pages

Key words and phrases

aristotle, social psychology, psychology, biological organism, plato, biological analysis, sechenov, psychol rev, biological, max meyer, protagoras, bois reymond, empedocles, experimental psychology, neurophysiology, clarendon press, biologisches centralblatt, chicago press, johannes miiller, francis bacon

Buy this book
Abebooks
Alibris
Amazon
Barnes&Noble
Froogle
Barnes&Noble.com

Borrow this book
Find libraries

Contents

THE USES OF HISTORY lashley, psychology, gestalt	1
ORIGINS OF MAGIC AND ANIMISM animism, magic, malinowski	20
ORIGINS OF NATURALISM heraclitus, thales, democritus	36

8 other sections not shown

References from books

A History of Western Psychology
By David J. Murray - Psychology - 1983 - 428 pages
Includes indexes

Related books

Adventures in Public Service: the careers of...
By Delia Kuhn, Ferdinand Kuhn - 1963
INTRODUCTION | Dryden | FROM KITTY HAWK TO THE MOON | by Howard Simons i | Llewellyn EThompson THE DIPLOMAT AND THE ENIGMA by Wallace Carroll | Sterling B Hendricks AT...
Snippet view - About this book

A History of Experimental Psychology
By Edwin Garrigues Boring - 1929 - 699 pages
Maps on lining-papers
Snippet view - About this book

Crucibles: The Story of Chemistry from

A Dictionary of Zoology By A. W. Leftwich

Summary

By A. W. Leftwich

Published 1963
Van Nostrand

290 pages

London ed. (Constable) has title: A student's dictionary of zoology.

Buy this book
Abebooks
Alibris
Amazon
Barnes&Noble
Froogle
Barnes&Noble.com

Borrow this book
Find libraries

Related books

The Concise Oxford Dictionary of Zoology
By Michael Allaby - Science - 1992 - 512 pages
The Concise Oxford Dictionary of Zoology is the first book of its kind to be published in manyyears, and the only such dictionary available in English.
Snippet view - About this book

References from books

The correspondence of Charles Darwin
By Charles Darwin, Frederick Burkhardt, Sydney Smith - 1985
Includes bibliography and index
Limited preview - Table of Contents - About this book

Figure 2. Four different user experiences with Google Book Search, 2 of 2: Snippet View (top), No Preview Available (bottom).

massive book collections was not considered feasible or even interesting until recently.

Google is committed to helping the (IC)DAR community further and in novel ways. Some of what we can do includes:

- **Open Source:** as a company, Google is firmly committed to open source and has already open sourced a substantial number of packages [8]. Among them is the Tesseract OCR package [10], originally developed by Hewlett-Packard, and which we recently helped resurrect and open source [12, 11]. With the help of the open source community at large, we are actively improving this engine and hope that some day, it will match and even surpass the accuracy of leading commercial engines.

- **Research and Open Source Grants:** Google has a very active research grant program as well as several ways to fund open source initiatives, including the *Google Summer of Code*. Recipients of such grants in the Document Analysis community include: Thomas Breuel of IUPR at DFKI received Google grants for the development of OCROpus, an advanced open source OCR and document analysis framework [5, 9]. OCROpus is able to leverage OCR engines like Tesseract, and its superior document image analysis and language modeling components enable it to squeeze more OCR accuracy out of these systems. As part of his work on OCRopus, Breuel also developed the very interesting *hOCR microformat,* designed to describe OCR workflow and results in a flexible and open manner [4]. Another document image analysis researcher, Apostolos Antonacopoulos, recently received a Google grant to assist in the creation of new groundtruth dataset for page segmentation and layout analysis [3, 2].

- **Datasets for Research:** last but not least, we are in the process of preparing datasets that we will be sharing with the research community. In the field of Document Analyis, and more generally Computer Vision, collecting good datasets for research is often an overwhelming task. As a result, the same old datasets get used over and over, which is undesirable. Also, some authors compile a new dataset just for the purpose of their experiments, which can make the quality of their work difficult to evaluate. We believe we can help by making some large chunks of our (out of copyright) data available to the Document Analysis research community.

The Google Book Search team is looking forward to working with the document analysis and understanding community on research programs that will forever change the way we use books and interact with massive book collections.

References

[1] C. Anderson. *The Long Tail.* Random House, 2006.

[2] A. Antonacopoulos, B. Gatos, and D. Bridson. ICDAR2007 page segmentation competition. In *ICDAR'2007, International Conference on Document Analysis and Recognition*, Curitiba, Brazil, Sept. 2007.

[3] A. Antonacopoulos, D. Karatzas, and D. Bridson. Ground truth for layout analysis performance evaluation. In H. Bunke and A. Spitz, editors, *Document Analysis Systems VII, Proceedings of the International Association for Pattern Recognition (IAPR) Workshop on Document Analysis Systems (DAS2006)*, pages 302–311, Nelson, New Zealand, Feb. 2006. Springer Lecture Notes in Computer Science, LNCS 3872.

[4] T. M. Breuel. The hOCR microformat for OCR workflow and results. In *ICDAR'2007, International Conference on Document Analysis and Recognition*, Curitiba, Brazil, Sept. 2007.

[5] T. M. Breuel. OCRopus home page, *http://code.google.com/p/ocropus/*, 2007.

[6] Google. Google search technology, *http://www.google.com/technology*, 1998.

[7] Google. About Google Book Search, http://books.google.com/googlebooks/about.html, 2004.

[8] Google. Google open source projects, *http://code.google.com/projects.html*, 2006.

[9] T. Raman. Google and open source OCR, *http://googleblog.blogspot.com/2007/06/google-and-open-source-ocr.html*, 2007.

[10] R. Smith. An overview of the Tesseract OCR engine. In *ICDAR'2007, International Conference on Document Analysis and Recognition*, Curitiba, Brazil, Sept. 2007.

[11] R. Smith and L. Vincent. Tesseract OCR home page, *http://code.google.com/p/tesseract-ocr/*, 2006.

[12] L. Vincent. Announcing Tesseract OCR, *http://google-code-updates.blogspot.com/2006/08/announcing-tesseract-ocr.html*, 2006.