The Present and Future of Google Books

[BooksOnline’10 Keynote Talk - Extended Abstract]

James Crawford
Google Books
Mountain View, CA
jimic@google.com

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms
Algorithms, Design, Human Factors, Performance

Keywords
Google Books

The Google Books project has the modest goal of scanning all of the world’s books, converting them to digital form, and making them searchable and accessible. To date over twelve million books, containing over four billion pages, have been scanned and digitized. This is an impressive number but it turns out that scanning is only the beginning of the challenge.

One part of the challenge in making books searchable and accessible is that a scan produces an image of a page, and often a blurred or partially obscured one at that, but searching requires a digital representation of the text on the page. Converting the image to text is also critical to creating a good reading experience since the text can then be reformatted to match the display size and the user can control the font size and layout. This is especially important for tablet devices and smart phones.

Another part of the challenge is that a search query will often match thousands or even tens or hundreds of thousands of books. Consider for example the query ‘to be or not to be’. This line has been quoted in an untold number of books (plus one more if the notes from this workshop are published). How should we choose the top ten matches to return on the first page of search results? Much work has been done on search rankings for web pages. Unfortunately these techniques do not apply well to books. The Google books teams has had to invent largely new techniques to rank book results.

A third challenge is that copyright law was not written with the digital world in mind. In the days of print books the cost of a print run was sufficiently high that when books went out of print they relatively rarely came back into print. However, in the digital world we have the technical ability to make millions of out of print books available as ebooks. For books that are out of print, but in copyright, the largest expense in creating ebooks is the cost of accurately identifying the owner of the digital rights. This cost has emerged as an important non-technical challenge to opening up many millions of out of print books.

This talk will discuss these challenges and a variety of others (how do you scan a one hundred year old book without breaking the binding?). We will also look at some of the new opportunities that arise from the emerging digital books corpus. These range from social collaboration to linguistic analysis to other new areas that are only beginning to be discovered.