Buying bestsellers online: A case study in Search and Searchability

Trystan Upstill
Australian National University

Nick Craswell and David Hawking
CSIRO Mathematical and Information Sciences
The Plan

• Why study book buying on the Web?

• The Experiment

• Results

• Bookstore Properties

• Recommendations and Conclusion
Why study book buying on the WWW?

- People buy books
- Many bookstores use their online presence to boost their sales
  - Some depend on online sales
- Search engines should help people buy books
- (We hope to be able to generalise our results across all similar products sold on the WWW)
Question I - Search

- How well can search engines process search requests that result in transactions?
  - **Transactional search** (Travis and Broder, 2001)
Question II - Searchability

• How compatible are the different service/product providers (bookstores) with existing search engines?
  – Can the pages be crawled/indexed?
  – Will pages be thrown out as duplicates?
  – How well will the pages be matched and ranked?
    • Is the best page returned?

• Providers that cannot be searched and/or return irrelevant pages may cut-off potential customers
The Experiment

- Convert books to proper case, Remove Duplicates
- Check bookstores for category best sellers
- Identify correct URLs (using ISBNs or bookstore codes)
- Google Directory (DMOZ)
- Yahoo Directory
- The New York Times
- 09/2002 bestselling book titles
- Microsoft Network (MSN)
- altavista
- alltheweb
- Evaluation

bookstores
bestselling books
book titles
unique book identifiers
bookstores with > 3 bestsellers
Manually Collected Data

Book information

<table>
<thead>
<tr>
<th>Book</th>
<th>ISBN</th>
<th>AllDirect ID</th>
<th>Cat</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Painted House</td>
<td>044023722X, 038550120X</td>
<td>2993</td>
<td>8</td>
</tr>
<tr>
<td>Artemis Fowl</td>
<td>0786808012, 0786817070</td>
<td>8883</td>
<td>1</td>
</tr>
<tr>
<td>Olivia Saves the Circus</td>
<td>068982954X</td>
<td>9928</td>
<td>1</td>
</tr>
<tr>
<td>Red Rabbit</td>
<td>0399148701</td>
<td>3774</td>
<td>4</td>
</tr>
</tbody>
</table>

Bookstore information

<table>
<thead>
<tr>
<th>Bookstores</th>
<th>Uses ISBN</th>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>1BookStreet.com</td>
<td>Yes</td>
<td>1,2,3,4,6,7,8,9</td>
</tr>
<tr>
<td>A1Books.com</td>
<td>Yes</td>
<td>3,5,6,7,8,9</td>
</tr>
<tr>
<td>AllDirect.com</td>
<td>No</td>
<td>1,2,3,4,5,6,7,8,9</td>
</tr>
<tr>
<td>Amazon.com</td>
<td>Yes</td>
<td>1,2,3,4,5,6,7,8,9</td>
</tr>
</tbody>
</table>

ADCS 2002 - Buying bestsellers online: A case study in search and searchability
Evaluation Measures

- Success Rate:
  - Represents how often search engines are returning relevant transactional pages at a certain cutoff
  - Represented by S@n
  - In the example n=1 (and 5, 10 ...) is a success

- Precision
  - Represents the amount of choice a search engine is offering (high precision indicates more book buying alternatives). What proportion of the results at a cutoff are relevant transactional pages.
  - In the example P@5 is 2/5
  - P@1 = S@1
Search Engine Success Rates
Search Engine Precision

![Graph showing precision vs. number of documents for different search engines: AltaVista, Fast, Google, and MSN Search. The graph indicates a decrease in precision as the number of documents increases.](image-url)
Bookstore Success Rates

![Graph showing success rates at different numbers of documents for Amazon, Barnes and Noble, Walmart, and BookSite.](image)

**Success rate at n (S@n):**

- Amazon
- Barnes and Noble
- Walmart
- BookSite

**Number of documents (n):**

- 0
- 1
- 10
- 100
- 1000

ADCS 2002 - Buying bestsellers online: A case study in search and searchability
The Winners

- The Winners
  - Bookstore:
    - Amazon
      - Success Rate @ 1 was 0.125 (or 1/8 queries)
      - Barnes and Noble second best: S@1 was 0.028 (1/35 queries)
  - Search engine
    - MSN Search and Google
      - MSN Search returns more books as their first result, Google gives you more book buying choices.
        » MSN S@1 was 0.36 (~ 1/3 queries) P@5 was 0.13 (~ 1/8 correct)
        » Google’s S@1 was 0.15 (~ 1/7 queries) P@5 was 0.20 (~ 1/5 correct)
      - Google’s success rate betters MSN Search after 10 results

- However...
  - Results still relatively poor.
    - Navigational Search S@1: 0.74 (Google & Fast, Airline HP Search)
Bookstore characteristics

• We measured:
  – Bookstore URL Coverage
    • How many URLs from a bookstore are indexed by a search engine
      – If few URLs are indexed there is little chance that deep transactional pages would have been crawled/indexed.
  – Bookstore Link Coverage
    • How many links to a bookstore has a search engine observed
      – If few links are observed it is less likely that a bookstore will be crawled. Link ranking algorithms also favour sites with a large number of in-links.

• Coverage is only one aspect of searchability
Link/URL Coverage

- Amazon
- B&N
- Walmart
- NetstoreUSA
- Powells

Average Link Coverage (2 search engines)

URL Coverage (4 search engines)
URL Coverage Distribution

Amazon

Altavista

4,000,000

2,000,000

MSN Search

Fast

Google

Walmart

Altavista

1,000,000

500,000

MSN Search

Fast

Powells

Altavista

150,000

100,000

50,000

MSN Search

Fast

Google
Transactional search suggestions

- Perform deep crawls of at least one bookstore
- Be prepared to crawl dynamic sites
- Consider service provider partnerships
- Make sure you get your referral credits
Observed Good vs. Bad Site Structure

GOOD

NetstoreUSA

category.html

000008477574
000008477545
000003777545

Amazon, Barnes and Noble

encourage incoming links at this level

BAD

Powells

/ 

encourage incoming links at this level

BookWorks

frameset

Powells, Walmart, Barnes and Noble


urls continually change, duplicate information

ADCS 2002 - Buying bestsellers online: A case study in search and searchability

14 December 2002
Conclusions

• Most bookstores have low searchability
  – Amazon > Barnes and Noble > Walmart > Other

• Transactional search is not performed very well by current search engines
  – MSN Search ~ Google > AltaVista > Fast

• People should be aware of these issues and consider them when designing Web sites
Future Work

• Find out what the other retrieved pages are
  – Are the informational, navigational, different media type?

• Attempt to classify transactional pages to improve search performance

• Expansion/Generalisation of experiments using CDs, Movies, Electronics search
Questions?

- Presentation available at:

- Email:
  Trystan Upstill: trystan@cs.anu.edu.au
  Nick Craswell: Nick.Craswell@csiro.au
  David Hawking: David.Hawking@csiro.au