Using Digital Book Metrics for Navigation and Browsing

Michael Huggett
University of British Columbia
m.huggett@ubc.ca

Edie Rasmussen
University of British Columbia
edie.rasmussen@ubc.ca

Abstract

As the scholar's work migrates from print to the digital realm, new ways of browsing, navigating and searching collections of digital books are needed. The Back-of Book Index is a carefully crafted source of information on a book's vocabulary and concepts, and if aggregated across multiple books, for a subject domain as a whole. Using a test collection of digital books in a variety of domains, we explore the use of index vocabulary to derive a series of metrics to indicate the relationships between index vocabulary and the digital collection, and the relationships between the books within a digital domain. We are investigating ways in which these metrics can be used to facilitate navigation and browsing at the domain level, to identify the most appropriate works within a digital collection for a particular subject or topic.

Keywords: digital books, back-of-book indexes, vocabulary, collections, browsing

Introduction

Millions of books are available in digital form as a result of mass digitization projects (such as the Gutenberg Project, the Million Books Project, the Open Content Alliance, and Google Books (Coyle, 2006)). As the scholar's work migrates from print to the digital realm, new ways of browsing, navigating and searching collections of digital books are needed. Currently, digital collections are searched at a macro level through metadata such as author, title, subject, etc., and at a micro level through keywords. In contrast, the back-of-book index (BoBI) operates at an intermediate level based on significant terms and concepts identified by a human indexer. Though relatively unstudied, the BoBI is a traditional knowledge structure to support search and browsing in the print domain (Jörgensen & Liddy, 1996; Liddy & Jörgensen, 1993); a few studies have examined its use within a single digital text, usually in comparison to a search engine (Abdullah & Gibb, 2008; Chi et al., 2006; Egan et al., 1989; Liesaputra, Witten & Bainbridge, 2009).

Because the BoBI is carefully crafted by a human domain expert, it is a rich source of information on a book's vocabulary and concepts, and if aggregated across multiple books, for a subject domain as a whole. The literature on BoBIs suggests that indexes serve a critical role in locating information in print and digital books, and that the indexing process generates a vocabulary that is richer, more structured, and more concentrated than that found in the book itself (Anderson & Pérez-Carballo, 2001; Gratch, Settel & Atherton, 1978). Such aggregate indexes could serve multiple purposes: to classify books based on summarized content (Enser, 1985), to generate a domain-descriptive vocabulary, to structure knowledge within the collection, and to navigate within the collection—applications which we are investigating in the Indexer's Legacy project (Huggett & Rasmussen, 2012).

In this poster we describe the use of index vocabulary to derive a series of metrics to indicate the relationships between index vocabulary and the digital collection, and the relationships between the books within a digital domain. We are investigating ways in which these metrics can be used to facilitate navigation and browsing at the domain level, to identify the most appropriate works within a digital collection for a particular subject or topic.

Acknowledgements: Funding from the UBC Hampton Fund and the Social Science and Humanities Research Council is gratefully acknowledged. We also thank our Graduate Research Assistants who have sourced and pre-processed the many books for our test collections.

Huggett, M., & Rasmussen, E. (2013). Using digital book metrics for navigation and browsing. *iConference 2013 Proceedings* (pp. 764-768). doi:10.9776/13370

Copyright is held by the authors.

Building a Test Collection

Our test collection consists of seven domain corpora in the Arts (Art History, Music), Humanities (Economics, Cooking), and Sciences (Geology, Anatomy, Darwin). Each domain is comprised of over 100 publicly-available digital books. The index is extracted from each book's PDF, and cleaned of OCR errors to leave only valid words. Each index entry is then 'expanded' to include on a single line its main heading, subheading(s) and page references. These expanded lines are then aggregated into a single file, sorted alphabetically, and compressed back into standard indented index format to create a single meta-index for an entire domain. The meta-index can be searched and browsed using the Meta-Dex User Interface[1] (Huggett & Rasmussen, 2012).

As the indexes are processed, we collect and calculate a variety of metrics that characterize the vocabulary, the books, and the domain. The metrics can be used to supplement the meta-index in browsing and navigating the digital collection.

Domain- and Vocabulary-Level Properties of the Test Collection

The basic properties of each domain are shown in Table 1. Each domain dominates in some measure. The domains show a high degree of variability in the number of tokens (i.e. instances of terms) and unique terms, within both content (i.e. the book's chapters) and index. The table also shows high variability in the number of main entries (*anchors*), sub-entries (*subs*), and page references (*refs*).

Table 1. Domains Compared by Basic Properties

domains	anatomy	arthistory	cookbooks	darwin	economics	geology	music	avg
num files	102	102	102	102	102	102	102	102
content terms	187366	165400	71519	152372	129897	194903	161315	151824.5714
content tokens	8655337	3815980	3269603	4509749	5511824	6378458	3927793	5152677.7143
content tokens / term	46.1948	23.0712	45.7166	29.5970	42.4323	32.7263	24.3486	34.8695
index terms	27276	37962	9430	33290	15595	38980	31200	27676.1429
index tokens	544537	233320	218963	210818	168526	276315	207960	265777
index tokens / term	19.9640	6.1461	23.2198	6.3328	10.8064	7.0886	6.6654	11.4604
num anchors	16142	24222	4943	20241	8372	25723	18378	16860.1429
num subs	130096	56671	58401	54057	44993	72105	49835	66594
subs / anchor	8.06	2.34	11.81	2.67	5.37	2.80	2.71	5.1086
num refs	226456	99152	78161	96944	46111	138335	98331	111927.1429
refs / book	2220.16	972.08	766.28	950.43	452.07	1356.23	964.03	1097.3257
refs / anchor	14.03	4.09	15.81	4.79	5.51	5.38	5.35	7.8514
refs / sub	1.74	1.75	1.34	1.79	1.02	1.92	1.97	1.6471
shared anchors	6879	7328	1966	6906	2840	8775	6713	5915.2857
shared_%	42.62	30.25	39.77	34.12	33.92	34.11	36.53	35.9029
coherence_%	3.7020	1.5836	3.5531	1.8788	2.4332	2.0634	2.2559	2.4957

The table's measures indicate some interesting properties. A low number of anchors suggests a 'compactness' of the domain's set of concepts. A higher number of subs per anchor suggest that a domain's individual concepts are more thoroughly elaborated, as does a higher ratio of references (per book, anchor, or sub).

Three calculated measures show subtler relationships:

- ***Shared anchors*** indicates the number of main entries in the domain that are shared between at least two books, providing a base measure of agreement of important concepts within the domain.
- ***Shared_%*** reflects the proportion of main entries in the domain that are shared between books. A higher value indicates better agreement on core concepts between the books of that domain. The

[1] http://meta-dex.no-ip.org

determination of whether low scores indicate a broad vocabulary or specialization into well-defined sub-areas may be solved by clustering books by dominant terms—a subject of future research.

- ***Coherence_%*** shows the degree to which all of the domain's main entries appear in all of its books. If all of a domain's entries appear in all of its books, the coherence is 100%. If none of books share any of their entries with another book, the coherence is 0%. The remarkably low degree of coherence of entries in each domain reveals how little texts agree on what is important, and how to express it. In other words, while a third or more of all main entries are shared within a domain, the vast majority are shared with very few other books. The degree of dispersal could possibly be lessened through better use of stemming algorithms that can aggregate strongly-related entries by variants such as such as singular and plural forms.

How much can we say about the domains based on these numbers, and to what degree of confidence? At this stage much is conjectural, and informed by our working familiarity with the domains. The numbers seem to support statements to the effect that Anatomy focuses on listing and organizing facts without much conjecture or argumentation, whereas Economics exposes ideas more by argumentation than by relying on a fixed set of discrete, specific facts. Clearly there is much more that we could do to compare domain properties—this forms a part of our future work.

Book-Level Properties of the Test Collection

The book-level properties across different domains (Table 2) mirror the domain-level properties, but the variability of size and formatting of each domain is more clearly exposed in the standard deviations (SD). For the number of anchors per book, the SD comes close to the average, whereas for subs and tokens the median is often lower and the SD often noticeably higher than the average, which suggests the influence that individual large books have upon the metrics of a domain.

Table 2. Book Properties across Domains.

books	anatomy	arthistory	cookbooks	darwin	economics	geology	music	avg
num books	102	102	102	102	102	102	102	102
anchors avg	685.09	533.42	260.69	476.25	289.67	717.87	528.38	498.7671
anchors med	509.00	368.00	179.00	337.00	231.50	509.00	466.50	371.4286
anchors SD	537.03	463.42	276.95	404.13	214.77	641.72	337.56	410.7971
subs avg	1272.34	254.16	496.57	279.27	239.72	367.74	213.50	446.1857
subs med	696.00	101.50	234.50	160.50	138.50	184.50	142.00	236.7857
subs SD	1725.52	447.27	795.29	392.85	295.10	553.74	288.22	642.57
tokens avg	6609.31	2859.50	2461.57	2847.88	2255.41	3234.90	2508.73	3253.9
tokens med	3999.00	1319.50	1287.00	1827.50	1490.50	2274.00	1732.50	1990
tokens SD	8357.62	3841.09	3174.88	3895.45	2396.85	3058.93	2252.00	3853.8314

To get an idea of the properties of individual books in the test domains, we compared them by basic properties (Table 3). First we ranked the books by a simple count of their index entries, and by the proportion (p) of that count to the total number of unique entries in the domain as a whole. Across domains, we find that a handful of books hold a disproportionate number of the entries available in the domain. The proportion of usage in the top book ranges from 8% (Art History) up to 16% (Anatomy) of available entries.

Table 3. Book Rankings within Domain: the Top Books in the Anatomy Domain.

(Left-to-right are tables for number of index entries, *representativeness*, and *information gain* per book. Book ID numbers are listed in the books column. Note the close agreement of book rank across tables.)

Most entries [of 16148]	
avg	642.5922330097087
med	147.0
min	1
max	2569
std	514.7325148865394

Rank	#ent	p	#@	books
1.	2569	0.16	1	493
2.	2427	0.15	1	373
3.	2353	0.15	1	546
4.	2131	0.13	1	504
5.	1969	0.12	1	372
6.	1764	0.11	1	396
7.	1712	0.11	1	297
8.	1673	0.10	1	490
9.	1669	0.10	1	516
10.	1443	0.09	1	507
11.	1335	0.08	1	506
12.	1305	0.08	1	388
13.	1125	0.07	1	312
14.	961	0.06	1	494
15.	941	0.06	1	538

Representative [6894 >1 ref]	
avg	552.747572815534
med	144.0
min	1
max	2195
std	420.2271858612045

Rank	#ent	p	#@	books
1.	2195	0.32	1	373
2.	2018	0.29	1	546
3.	1967	0.29	1	493
4.	1835	0.27	1	372
5.	1501	0.22	1	297
6.	1495	0.22	1	490
7.	1256	0.18	1	396
8.	1209	0.18	1	506
9.	1162	0.17	1	504
10.	1128	0.16	1	507
11.	1107	0.16	1	388
12.	1035	0.15	1	516
13.	1014	0.15	1	312
14.	900	0.13	1	375
15.	888	0.13	1	494

Book info gain		
Rank	effect	ID
1.	13.0327	493
2.	12.9493	373
3.	11.9996	546
4.	9.5926	372
5.	8.0737	504
6.	7.4140	297
7.	7.1929	490
8.	7.0660	396
9.	5.9894	516
10.	5.2209	507
11.	5.0179	506
12.	4.6046	388
13.	3.3794	312
14.	2.2448	494
15.	2.1505	375

Second, we ranked the books by what we call their *representativeness*: the proportion of entries that they share with other books of the domain. Again, a small number of books share a disproportionate number of their entries: the top book in a domain ranges from 14% (Art History) up to 32% (Anatomy).

Third, we ranked books by their *information gain*, measured with respect to a book's effect on domain *coherence***.** We define coherence as the degree to which all of the books of a corpus share all of the available index entries in the corpus (see above). A book's information-gain value is calculated as the effect on domain coherence if the book were removed from the domain. If a book shares entries with many other books, the book is effectively a 'hub node' in a shared-entry network model: removing a hub can drastically reduce a network's connectedness, in the worst case potentially splitting the domain into discrete unconnected sub-networks.

We find that only a few books in each domain score highly as 'hubs', and that the level of gain tends to drop off rapidly. The gain of the top book between domains also varies from 4.45 (Art History) up to 13.03 (Anatomy), indicating that the top books in Anatomy capture more of the gist of their field than do the top books of Art History. In comparing these three measures, the chief observations are that books that have many entries also tend to share more of their entries with other books (*representative*), and tend to act as central hubs that bind the domain together (*information gain*).

It may seem self-evident that big books are important, but the results say more than that: a book with a lot of entries and a high information-gain value is broad and general, rather than highly specific and exhaustively detailed, since even a big book would have a low information-gain score if it shared few of its entries with other books.

Discussion

The main goal of the Indexer's Legacy project is to explore ways in which the back-of-book index can contribute to the effective use of digital books within the subject domains of a collection. The meta-index provides a topic-based approach for searching a domain, but we are also interested in exploring domains based on their vocabulary, the unique properties of individual books, and relationships between books. We are currently building tools that employ visualizations and network models to support domain navigation. These tools address questions such as "Which books are most central to the domain?";

"Which books have the greatest coverage?"; "What book is most closely related to another?" We are also interested in developing new metrics that will better support exploration at the domain and collection level.

References

Abdullah, N. & Gibb, F. (2008). Using a task-based approach in evaluating the usability of BoBIs in an E-book environment. *ECDL 2008*, *LNCS* 4956, 246-257.

Anderson, J. D. & Pérez-Carballo, J. (2001). The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part I: Research, and the nature of human indexing. *Information Processing & Management* 37: 231-254.

Chi, E. H., Hong, L., Heiser, J., & Card, S. K. (2006). ScentIndex: Conceptually reorganizing subject indexes for reading. *VAST 2006*, 159-166.

Coyle, K. (2006). Mass digitization of books. *Journal of Academic Librarianship* 32(6), 641-645.

Egan, D.E., Remde, J.R., Gomez, L.M., Landauer, T.K., Eberhardt, J. & Lochbaum, C.C. (1989). Formative design evaluation of Superbook. *ACM Transactions on Information Systems* 7(1), 30-57.

Enser, P.G.B. (1985). Automatic classification of book material represented by back-of-the-book index. *Journal of Documentation* 41(3), 135-155.

Gratch, B., Settel, B., & Atherton, P. (1978). Characteristics of book indexes for subject retrieval in the humanities and social sciences. *The Indexer* 11(1): 14-23.

Huggett, M. & Rasmussen, E. (2012). Dynamic online views of meta-indexes. *JCDL 2012*: 233-236.

Jörgensen, C. & Liddy, E.D. (1996). Information access or information anxiety?—an exploratory evaluation of book index features. *The Indexer* 20, 64-68.

Liddy, E. D & Jörgensen, C. (1993). Reality check! Book index characteristics that facilitate information access. *Proceedings of the 25th Annual Meeting of the American Society of Indexers*, 125-138.

Liesaputra, V., Witten, I.H., and Bainbridge, D. (2009). Searching in a book. *ECDL 2009*, *LNCS* 5714, 442-446.