HathiTrust Large Scale Search: Scalability meets Usability

www.hathitrust.org

Tom Burton-West
Information Retrieval Programmer
Digital Library Production Service
University of Michigan Library
www.hathitrust.org/blogs/large-scale-search

Code4lib
February 7, 2012

HathiTrust

- HathiTrust is a shared digital repository
- 50+ member libraries
- Large Scale Search is one of many services built on top of the repository
- Currently about 10 million books
- 450 Terabytes
  - Preservation page images;jpeg 2000, tiff (438TB)
  - OCR and Metadata about (12TB)

HATHI TRUST Digital Library

Help Feedback

Home About Collections My Collections

Catalog Search

Search information **about** the items.

Find

All Fields Full view only

Advanced Catalog Search | Search Tips

Full-text Search

Search words that occur **within** the items.

Find

Full view only

Search Tips

Collections

Browse, search, or make HathiTrust collections.

View Public Collections

Collections are a way to group items for private or public use. The full-text of items within a collection can be searched independently of the full digital library. Read more about creating collections.

Recent News and Publications

Try our new mobile website!

HathiTrust Constitutional Convention

Information about the Author's Guild Lawsuit

- Authors Groups From U.K., Canada, Norway and Sweden Join Authors Guild, Australian Society of Authors, and Quebec Writers Union in Suit Against HathiTrust - The Author's Guild
- HathiTrust and the Litigation Path - Columbia Copyright Advisory Office
- University of Michigan Information on the Authors Guild Lawsuit
- Collaborating Globally, Planning Locally: HathiTrust and New Opportunities in Collection Management
- The Laboratorium: The Procedural Swamp

view more news ››

Currently Digitized

9,710,967 total volumes
5,154,728 book titles
256,156 serial titles
3,398,838,450 pages
435 terabytes
115 miles
7,890 tons
2,641,982 volumes (~27% of total)
in the public domain
View visualizations of HathiTrust
call numbers, languages, and dates

statistics information ››

Featured Collection

UM Press

Univ. of Michigan Press's full-text titles available in HathiTrust

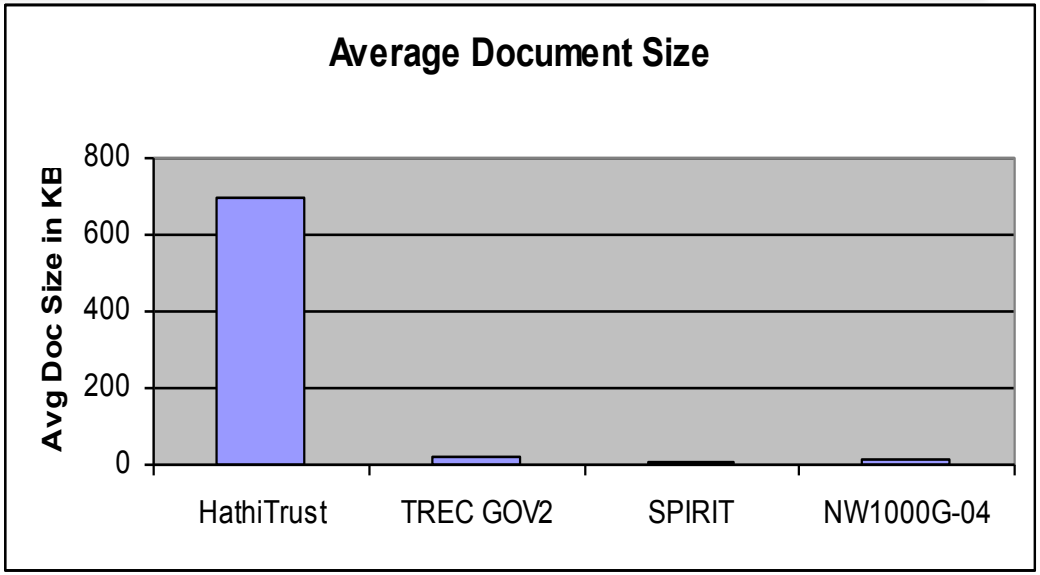
About Help Feedback Take-Down Policy Privacy Contact

Large Scale Search Challenges

- Goal: Design a system for full-text search that will scale to 10 million -20 million volumes (at a reasonable cost.)
- Challenges:
  - Must scale to 20 million full-text volumes
  - Very long documents compared to most large-scale search applications
  - Multilingual collection (400+ languages)
  - OCR quality varies

Long Documents

- Average HathiTrust document is 700KB containing over 100,000 words.
  - Estimated size of 10 million Document collection is 7 TB.
- Average HathiTrust document is about 38 times larger than the average document size of 18KB used in Large Research test collections
- 2 quadrillion words in corpus



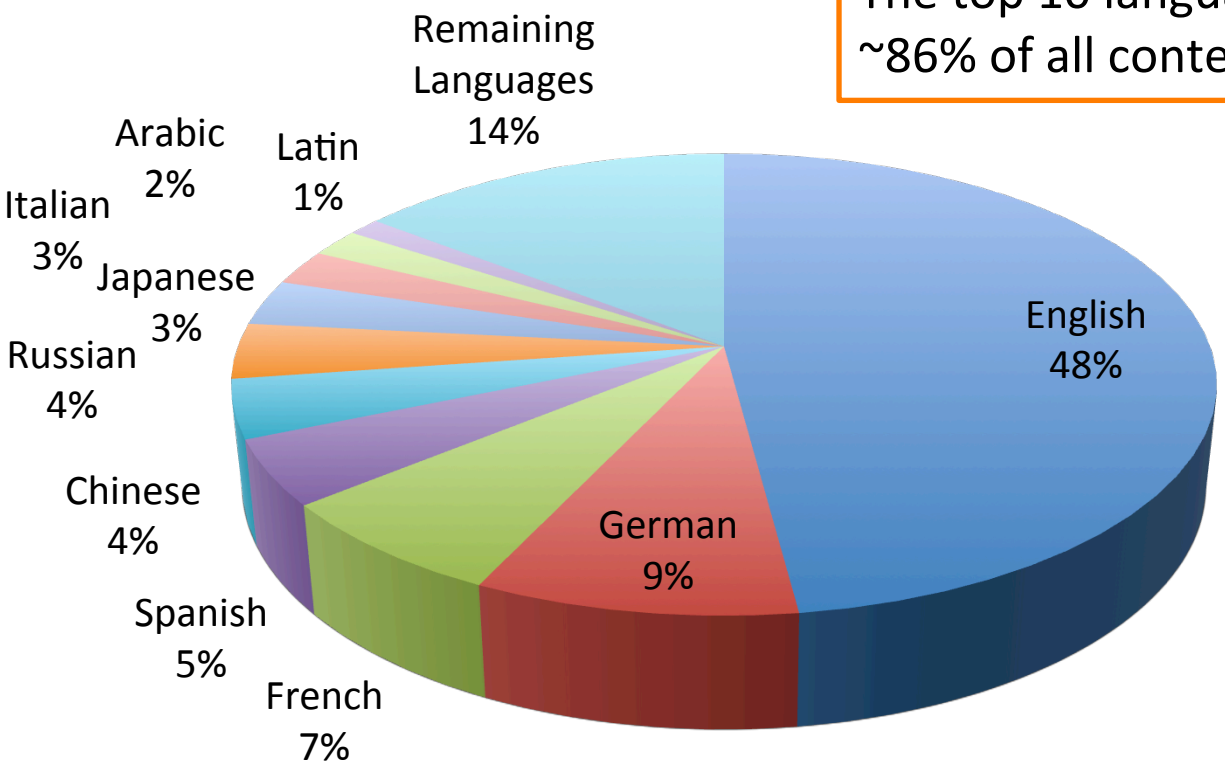
Collection	Size	Documents	Average Doc size
HathiTrust	7 TB	10 million	700 KB
TREC GOV2	0.456 TB	25 million	18 KB
SPIRIT	1 TB	94 million	10 KB
NW1000G-04	1.3 TB*	100 million	16 KB

Multilingual

- 400+ languages, 50 languages with over 1,000 volumes.
- About 500,000 (5%)of volumes in multiple languages
- Currently all languages in one index
- Lowest common denominator tokenizing
- Some languages are challenging for Information Retrieval
  - CJK
  - Arabic
  - Finnish, Turkish, German

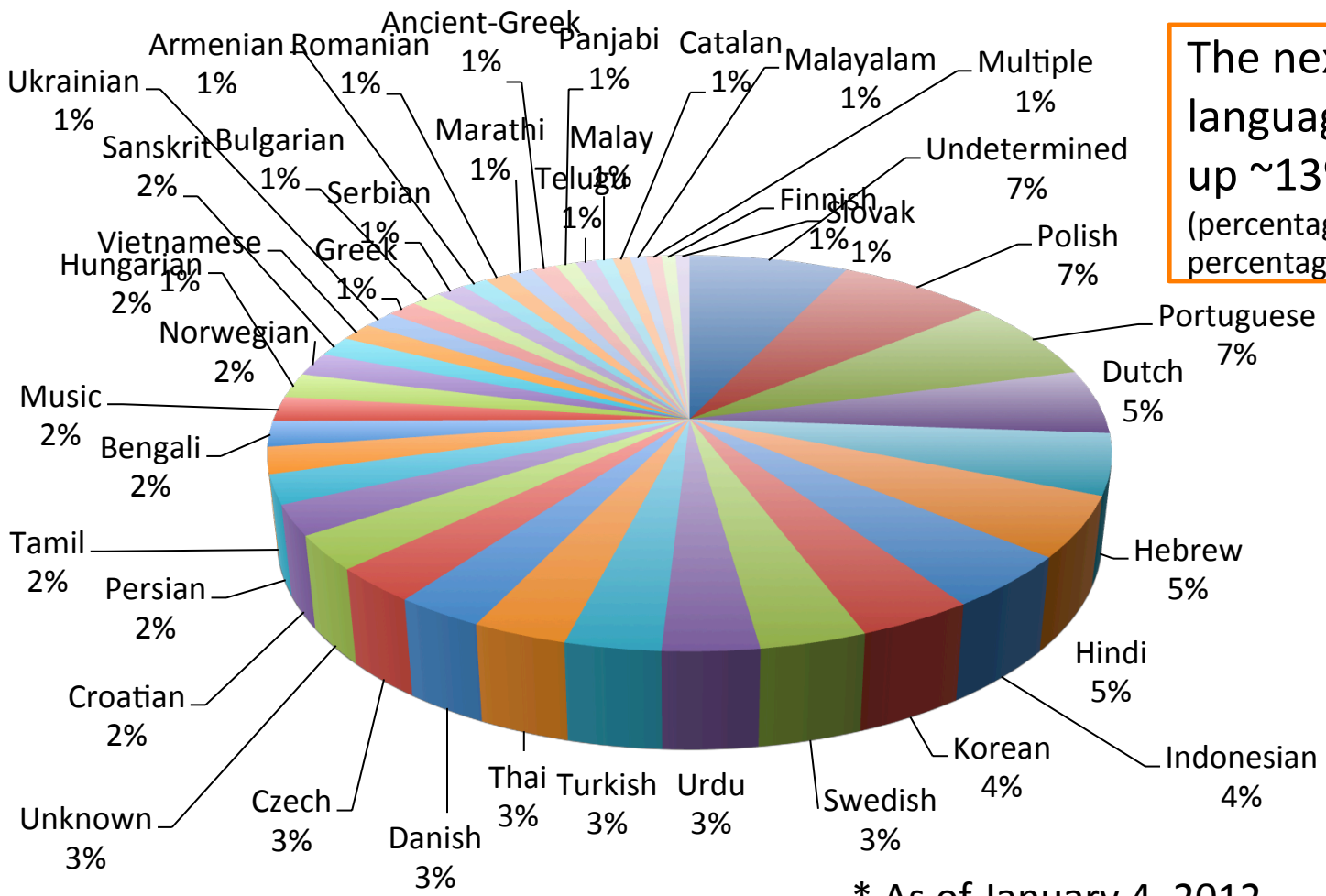
Language Distribution (1)

The top 10 languages make up ~86% of all content



* As of January 4, 2012

Language Distribution (2)

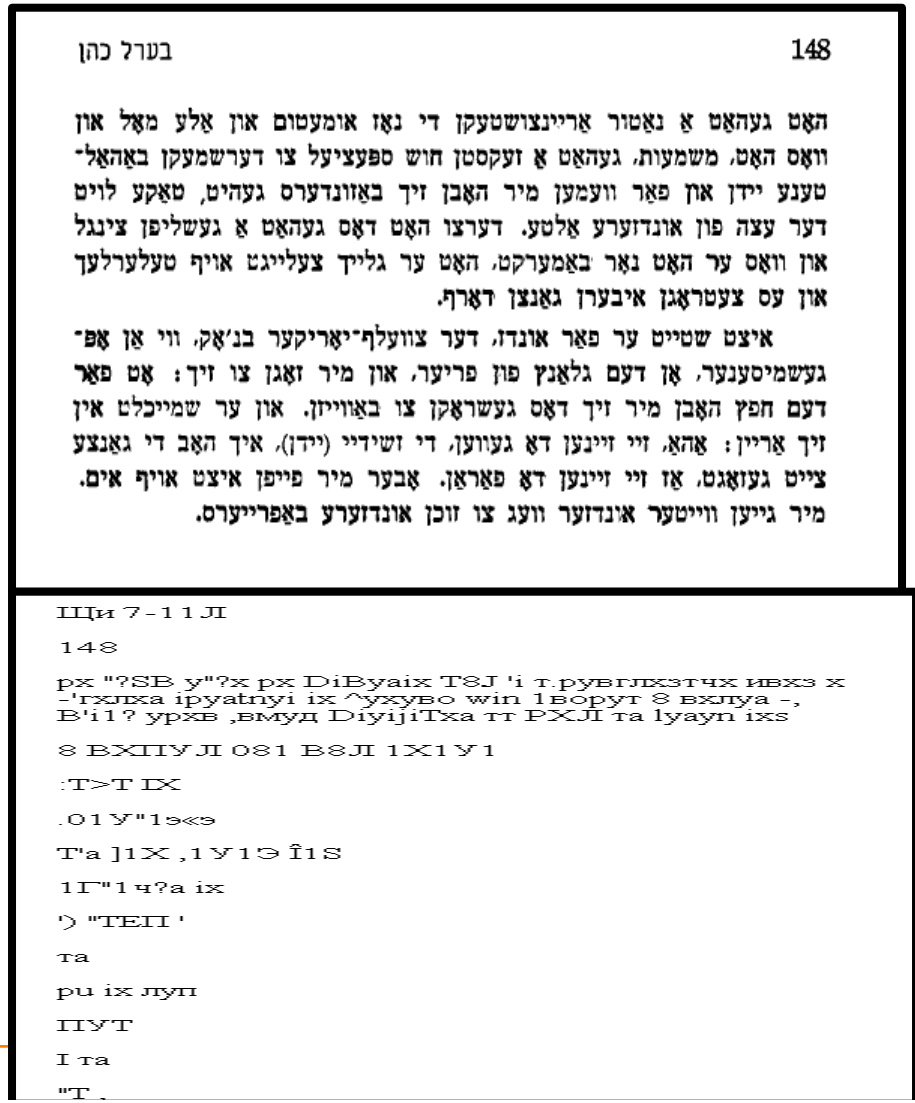


The next 40 languages make up ~13% of total
(percentages shown are percentage of the 13%)

* As of January 4, 2012

“Dirty” OCR

- The OCR varies in quality
- Dirty OCR results in large terms index, and high memory consumption
- Dirty OCR affects term statistics and makes some volumes difficult to find
- Example: Hebrew characters not recognized



בערל כהן 148

האָט געהאַט אַ נאַטור אַריינצושטעקן די נאָז אומעטום און אַלע מאָל און וואָס האָט, משמעות, געהאַט אַ זעקסטן חוש ספּעציעל צו דערשמעקן באַהאַל־טענע יידן און פאַר וועמען מיר האָבן זיך באַזונדערס געהיט, טאַקע לויט דער עצה פון אונדזערע אַלטע. דערצו האָט דאָס געהאַט אַ געשליפן ציינגל און וואָס ער האָט נאָר באַמערקט, האָט ער גלייך צעלייגט אויף טעלערלעך און עס צעטראָגן איבערן גאַנצן דאָרף.

איצט שטייט ער פאַר אונדז, דער צוועלף־יאָריקער בנ׳אָק, ווי אַן אָפּ־געשמיסענער, אָן דעם גלאַנץ פון פריער, און מיר זאָגן צו זיך: אָט פאַר דעם חפץ האָבן מיר זיך דאָס געשראָקן צו באַווייזן. און ער שמייכלט אין זיך אַריין: אַהאַ, זיי זיינען דאָ געווען, די זשידיי (יידן), איך האָב די גאַנצע צייט געזאָגט, אַז זיי זיינען דאָ פאַראַן. אָבער מיר פייפן איצט אויף אים. מיר גייען ווייטער אינדזער וועג צו זוכן אונדזערע באַפרייערס.

Щи 7-11 Л

148

px "?SB y"?x px DiByaix T8J 'i т.рувглхзтчх ивхз х
-'гхлха ipyatnyi ix ^ухуво win 1ворут 8 вхлуа -,
B'i1? урхв ,вмуд DiyijiTxa тт РХЛ та lyayn ixs

8 ВХПУЛ 081 В8Л 1Х1У1

:Т>Т IX

.01У"1э«э

Т'а]1Х ,1У1Э Î1S

1Г"1ч?а ix

Э "ТЕП '

та

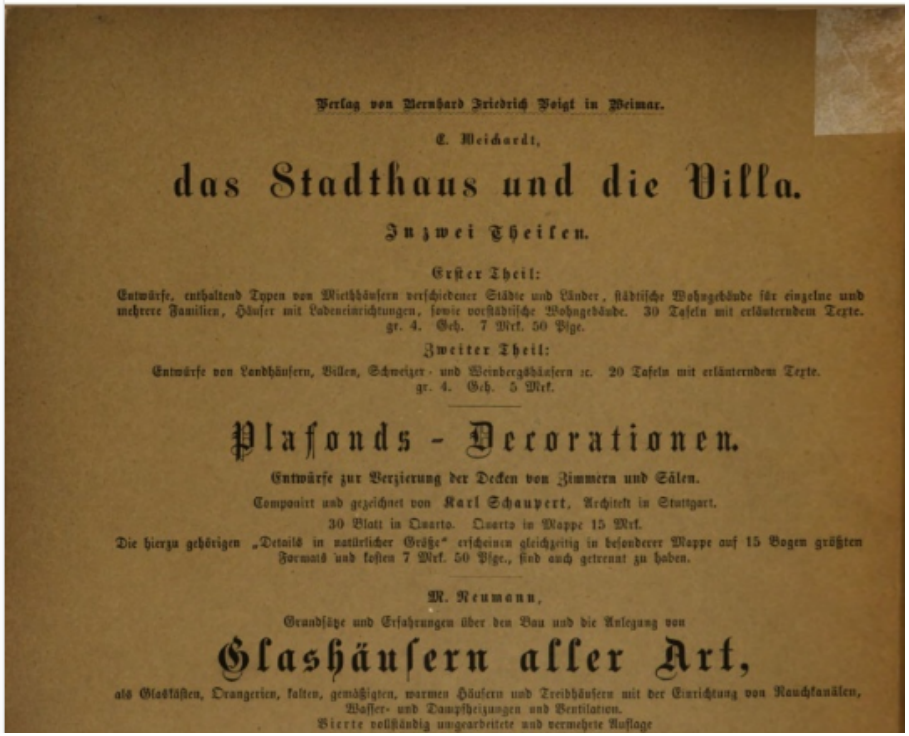
pu ix луп

ПУТ

I та

"Т ,

OCR from discolored text



Verlag von Bernhard Friedrich Voigt in Weimar.

E. Weichardt,

das Stadthaus und die Villa.

In zwei Theilen.

Erster Theil:

Entwürfe, enthaltend Typen von Miethhäusern verschiedener Städte und Länder, städtische Wohngebäude für einzelne und mehrere Familien, Häuser mit Ladeneinrichtungen, sowie vorstädtische Wohngebäude. 30 Tafeln mit erläuterndem Texte. gr. 4. Geh. 7 Mrk. 50 Pfge.

Zweiter Theil:

Entwürfe von Landhäusern, Villen, Schweizer- und Weinbergshäusern ꝛc. 20 Tafeln mit erläuterndem Texte. gr. 4. Geh. 5 Mrk.

Plafonds - Decorationen.

Entwürfe zur Verzierung der Decken von Zimmern und Sälen.

Componirt und gezeichnet von Karl Schaupert, Architekt in Stuttgart.

30 Blatt in Quarto. Quarto in Mappe 15 Mrk.

Die hierzu gehörigen „Details in natürlicher Größe" erscheinen gleichzeitig in besonderer Mappe auf 15 Bogen größten Formats und kosten 7 Mrk. 50 Pfge., sind auch getrennt zu haben.

M. Neumann,

Grundsätze und Erfahrungen über den Bau und die Anlegung von

Glashäusern aller Art,

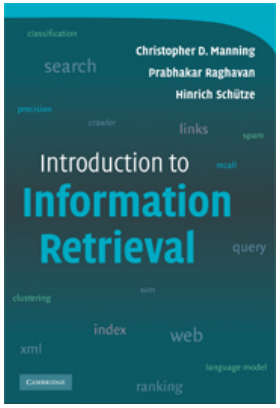
als Glaskästen, Orangerien, kalten, gemäßigten, warmen Häusern und Treibhäusern mit der Einrichtung von Rauchkanälen, Wasser- und Dampfheizungen und Ventilation.

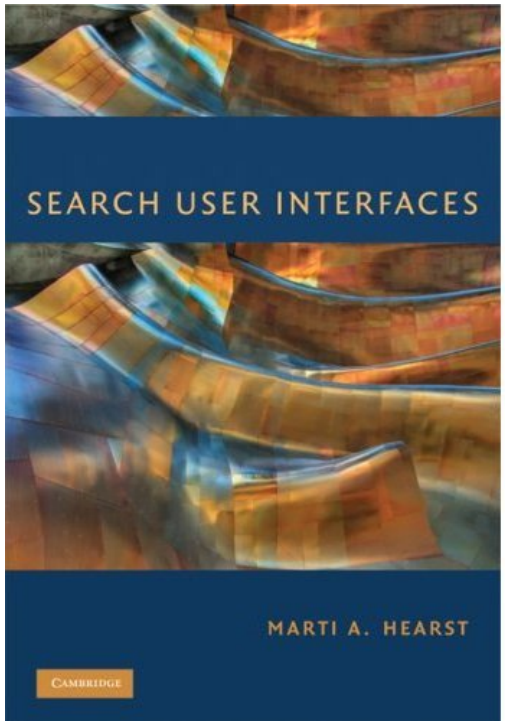
Vierte vollständig umgearbeitete und vermehrte Auflage

^018 ^1(1^11101118 1111(1^10 ^ ! ^^.
されき10^菊ゆ"【巛れ,
0;11110なホ, IIIゆ011^ 7^^6110011則"も【"I106^(^1^ ? 11 ? 1:(^"も16 ? 8"1)811 ^&'380^119665111)6 ^1 "II各""1
111(0
I"ゆ ひひ 80"111^11, お3"レ1:1111180^111(1ゆ("""が", 1。^^ 。! ^!^!^!々 ? ;11(0511361)8110し30 2.0^61x1 "1 ひ
(【ほ"ひ【111) ? "1 ^;;^.
01. 4, 0ゆ.7舰50新が
ほI110(11(6 0011 ? 0111)^11161:I!,おら由I13ひ ;Iは, 11I!)) 3861(1(1ひ 081)011【ひひな.20 3^0^11111111れは
力"!!)^!^ 3 ! ひな.
9^ん&^. 5 3^【1
I ^ 1011 (! ^ー^お
611^111^ 1(111お"がひ ""(^わはX^オ"Iわ011 3^111^11ひ 11& @な(は.
國ががゆ^ま國^なバ0ゆな11やはォ, I @1"十9。バ.
國もIIはさ"9ゆれ!が" , で^01ほI 1101な【1は"&V5^^"Iひ ^!"!^" I &[。"わ"は015お0が11915^611
^01111(1(8 11"も^0 7朋".50刺が.,國011も3"ひ 111^ I
311. 91(11111(1 II !!,
(^I:^!!^^ひ 11&ほ【|0&1:ひ &611お0は111I& &16 ^111^11114 0011
國@10さほ(!は,0|:0"0ひな", ? 0^", が"1&6;9^ん011(^"國|^^^00國國ほ^"!^""ョ國9 ? 011もひ 0,
3^0[^ひ 11"0 3)0"ゆめ6ナ""0む`11111^お凸111ぉ1011,
忍【ばでひ 000(1^11&は111110は1: ! ^"" ? 11I)& 0ば111ゆバ6 &11^0が
0011さ, ぶ0バ10モ0,
011161119110さ00" 25 1111& 244お&!?11わ""9"し9^ 8, @ゆ.10舰ル50 ? ? 9^

Scalability meets Usability : Challenges

- Must have reasonable response time as collection scales to 20 million volumes
- Provide good results on the first result page (relevance ranking)
- Provide good results for non-English searching
- Help users manage large result sets or drill down within them
- All the usability issues for search interfaces apply regardless of scale



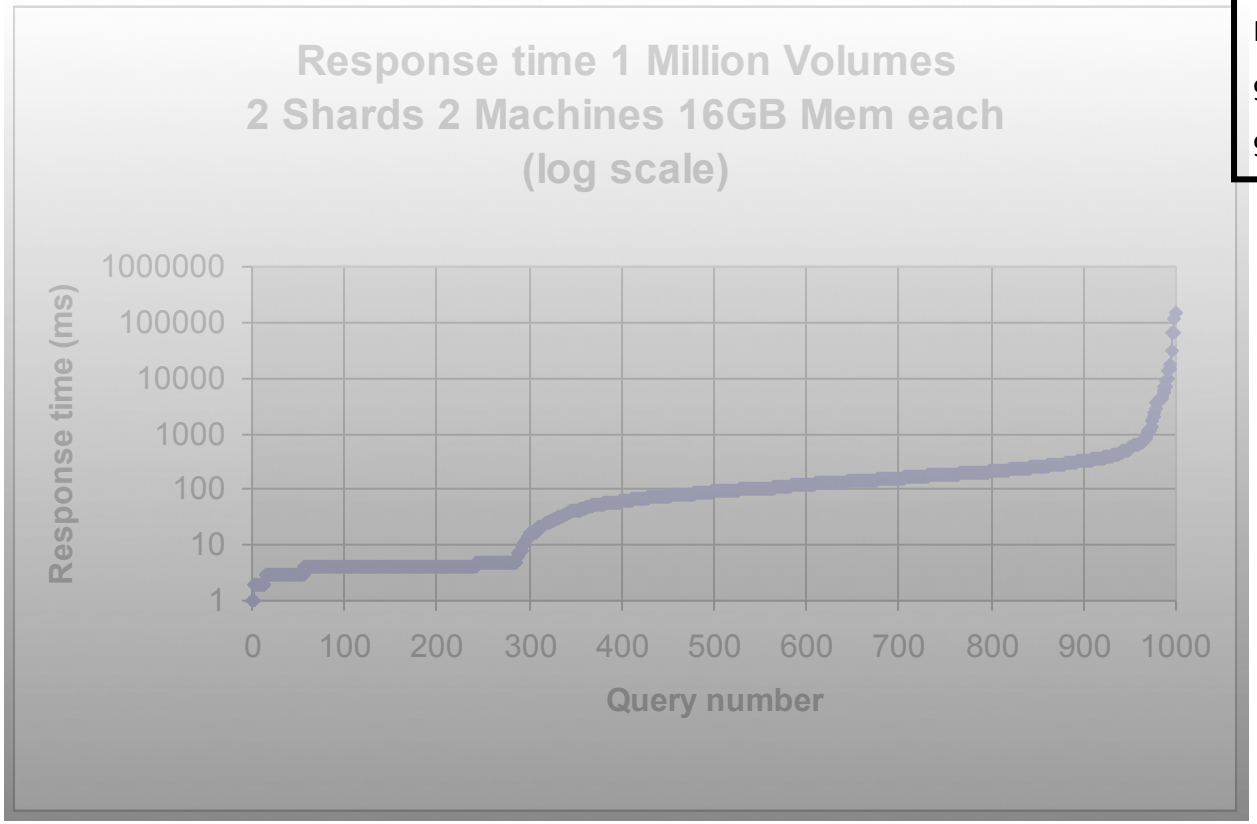


Response time and scalability

- Ideally response time should be under 1 second*
- Response time of over 10 seconds distracts users from their tasks*
- Early scalability tests with 1 million volumes showed the slowest 1% of queries took from 10 seconds to two minutes

*See: http://www.useit.com/papers/responsetime.html

Response Time Varies with Query



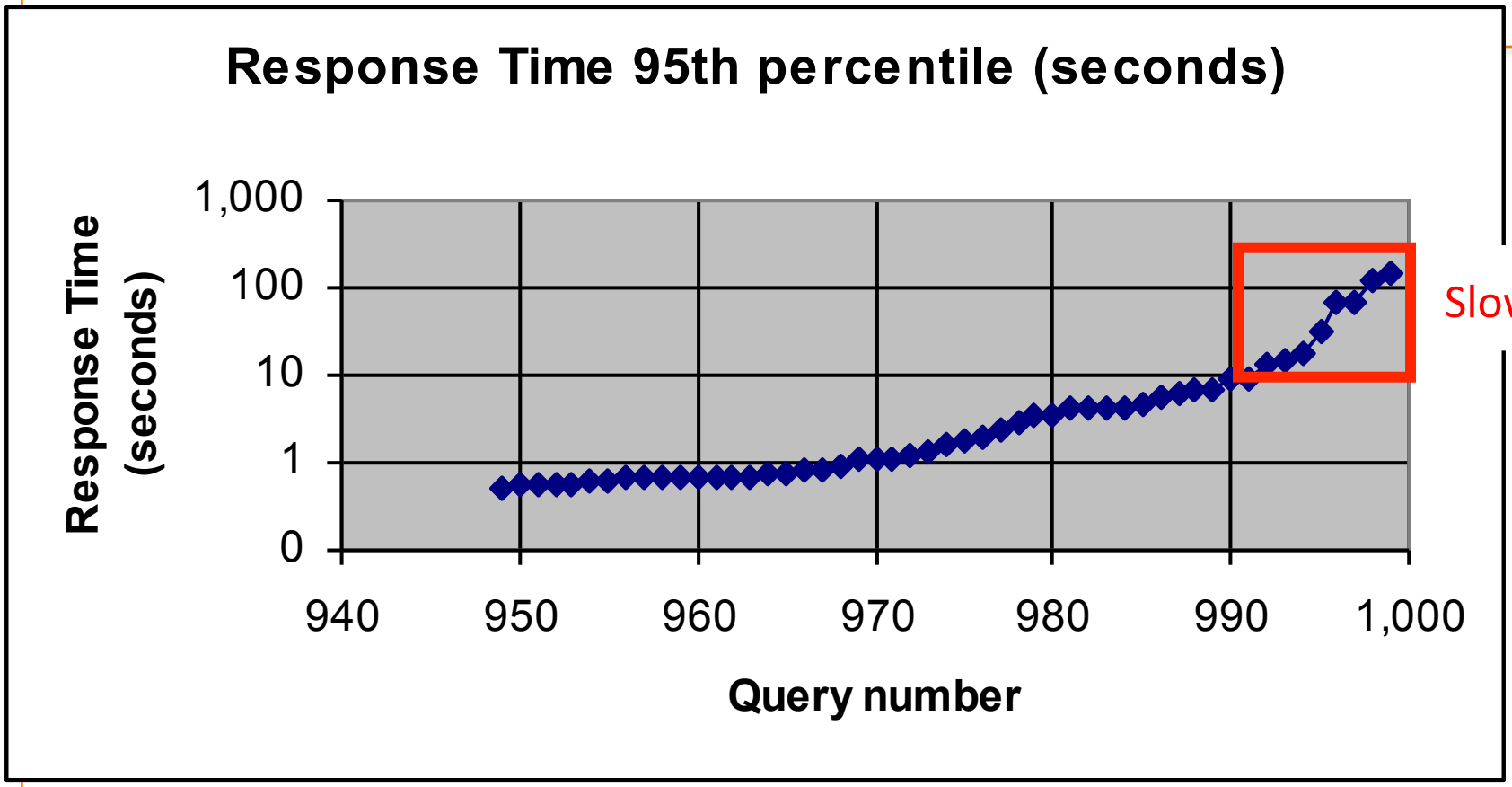
Average: 673

Median: 91

90th: 328

99th: 7,504

Slowest 5% of queries



Performance problem: phrase queries too slow

- Slowest query: “The lives and literature of the beat generation” took 2 minutes
- Cause of slow queries is high disk I/O
- Phrase queries require reading the “positions index” from disk
- Common words such as “the” occur so frequently that reading their positions list requires many GB of disk I/O

Stop Words

- The word “the” occurs 40 billion times in the corpus or about 4 billion times per million documents.
- Removing “stop” words (“the”, “of” etc.) not desirable
- Couldn’t search for many phrases
  - “to be or not to be”
  - “the who”
  - “man in the moon” vs. “man on the moon”

Stop Words

- Stop words in one language are content words in another language
- German stopwords “war” and “die” are content words in English
- English stopwords “is” and “by” are content words (“ice” and “village”) in Swedish

CommonGrams

- Instead of stop words, create bi-grams for common words
- Example: Slowest query: “The lives and literature of the beat generation”
  - “the-lives” “lives-and”
  - “and-literature” “literature-of”
  - “of-the” “the-beat” “generation”

Standard index vs. CommonGrams

(500,000 document index)

- Standard Index

WORD	TOTAL OCCURRENCES IN CORPUS (MILLIONS)	NUMBER OF DOCS (THOUSANDS)
the	2,013	386
of	1,299	440
and	855	376
literature	4	210
lives	2	194
generation	2	199
beat	0.6	130
TOTAL	4,176	

- Common Grams

TERM	TOTAL OCCURRENCES IN CORPUS (MILLIONS)	NUMBER OF DOCS (THOUSANDS)
of-the	446	396
generation	2.42	262
the-lives	0.36	128
literature-of	0.35	103
lives-and	0.25	115
and-literature	0.24	77
the-beat	0.06	26
TOTAL	450	

CommonGrams
Comparison of Response time (ms)

	average	median	90th	99th	slowest query
Standard Index	459	32	146	6,784	120,595
Common Grams	68	3	71	2,226	7,800

http://www.hathitrust.org/blogs/large-scale-search/slow-queries-and-common-words-part-2

http://www.hathitrust.org/blogs/large-scale-search/slow-queries-and-common-words-part-1

Good Response Time with Large Indexes: Tuning Caching and Memory

- Our 10 million document index is about 7 terabytes.
  - About 700 GB per million documents
- Large index means disk I/O is bottleneck
- Tradeoff JVM vs OS memory
  - Solr uses OS memory (disk I/O caching) for caching of postings
  - Tests showed memory available for disk I/O caching has most impact on response time (assuming adequate cache warming)
- Performance tuning
  - Reduce memory needed by Solr (16GB for 3 shards)
    - termInfosIndexDivisor and termIndexInterval
  - Leave a large amount of free OS memory for caching (about 50GB per server)
  - Run cache-warming queries every morning

Full-text search and "search within a book": Tuning for response time

- Full-text search of 10 million books uses the whole book as a Solr document.
  - Size about 7 TB, so we need to optimize the index for minimal query latency.
  - We index off-line
  - We optimize off-line before putting the index into production for searching
- "search within a book" indexes a book on-the fly and indexes each page of a book as a Solr document.
  - Each page has a "book id" field so a search within the book includes a filter query: q1=bacon&fq= book_id=12345
  - It is optimized for minimal latency between the time the book is sent for indexing and the time the book is available for searching
  - We keep the index small by deleting it nightly. Average size is about 4GB
  - We index and serve queries in the same Solr instance
  - We configure the index to minimize merging
  - We store the pages so that we can display and highlight snippets

Good Results and Relevance

- System magic
  - query processing
    - Field weighting 1 (add boosts for matches in author/title/ subject metadata)
    - Query normalization (lower casing, diacritic normalization)
    - Query expansion (synonyms, stemming, added phrase queries)
    - Field weighting 2 , weighted query expansion/normalization
    - query intent guessing
- Help users refine their search
  - user query reformulation (UX issues)
  - Advanced search including MARC metadata
  - faceting

Relevance: Recall and Precision

- Recall = how many of the relevant results in the collection are returned
- Precision = the ratio of relevant results to the total number of results.
- Normally for relevance ranking these are measured at some number of results for example p@10 is the precision at 10 results

Precision, Recall and Performance

- Generally there is an inverse relationship between precision and recall
- Different types of queries favor recall or precision
- Compare these queries:
  - dog OR food (4,431,164) hits
  - dog AND food (2,006,995) hits
  - “dog food” (52,784) hits

For large-scale search, favor precision over recall

- With the full text of over 10 million very long documents large result sets can be a problem
  - We changed the default Solr Boolean operator from “OR” to “AND”
  - We allow phrase queries so users can enter very precise queries
  - We don’t use stop words
  - We don’t do stemming or use synonyms*

Weighted query expansion to boost recall without hurting precision. (Scalable?)

- User query: [dog food]
  - Boost query: (“dog food”)^100 OR (“dog food”~200) ^50 OR(dog AND food)^10 OR (dog OR food) ^1
  - Dismax query: _query_:"{!edismax

    qf='ocr^500+marcfields^10+

    pf =ocr^5000+marcfields^40+

    pf2=ocr^2500+marcfields^20

Usability v. Scalability:

Relevance v.s. performance

- Stopwords
  - (solved with CommonGrams and cache warming)
- Adding weighted phrases (dismax pf,pf2,pf3)
  - (High disk I/O for OCR phrase searches)
- Proximity (words on same page/within 200 words)
  - (High disk I/O for OCR phrase searches)

Usability v. Scalability:

Relevance v.s. performance complicated by multiple languages

- truncation operator
  - 3+ billion unique words make truncation computationally intractable
- Fuzzy search (helps with dirty OCR)
  - 3+ billion unique words=intractable
- Character ngrams (language independent stemming/helps with dirty OCR)
  - Index size increase of 5-20x not scalable

Stemming and synonyms

- Stemming conflates different forms of the same word to increase recall.
  - cat, cats
  - organized, organizer, organizing, organization
- However over-stemming leads to loss of precision by conflating words with different meanings
  - Organize, Organic, Organ =>Organ

Stemming and synonyms

- Solution: create separate field and combine unstemmed and stemmed field with lower weight for stemmed field
  - <copyField source="unstemmed" dest="stemmed">
  - Boost: q=unstemmed^1000+stemmed^50
- Problems
  - Stemming is language-specific
  - Copying the OCR field would double the index size from 6TB to 12TB
- Current hack
  - Use copyfield stategy with English stemmer for MARC fields only.
- Synonyms have similar issues

Normalization, diacritics and thé

- We normalize both text and queries. Words are lower-cased and diacritics are removed
- Removing diacritics allows searching for non-English words without special keyboard mappings
- The French word for tea is thé
  - This gets normalized to "the" resulting in millions of hits.
  - The term statistics (idf) and relevance ranking are also affected as thé would normally occur much less frequently than "the"
  - Copyfield solution would double size of index

Normalization, diacritics and thé

- Copyfield solution would double size of index
  - <copyField source="diacritics" dest="noDiacritics">
  - Boost: q=diacritics^1000+noDiacritics^50
- Possible alternative: create filter that outputs token both with and without diacritics as if they were synonyms (only when there is a diacritic in input)
    - Search for thé would only match docs containing thé
    - Search for the would match both the and thé
    - Index size increase proportional to number of tokens with diacritics

Tokenization and other language-specific issues

- With over 400 languages, using language detection and putting each language in a separate field is not feasible
  - What to do with 500,000 volumes in multiple languages?
- We put all languages in one field
- Need to use analyzer that works ok for all languages: ICUFilters

Tokenization and other language-specific issues

- Some languages such as Chinese, Japanese, Thai, and Vietnamese don't use spaces to separate words*
  - Language-specific solutions (SmartChineseAnalyzer)
  - Script based solutions (CJK tokenizer, ICUTokenizer)
- For Chinese there are two character sets: Traditional and Simplified. Users expect to enter terms in one character set and get results in both.
- Japanese uses 4 different scripts, the same word may be written in numerous ways

* http://www.hathitrust.org/blogs/large-scale-search/multilingual-issues-part-1-word-segmentation

Tokenization and other language-specific issues

- Arabic has very large number of word forms and needs stemming for effective retrieval.
  - Solr has a very good light stemmer for Arabic
  - Problem is that other languages such as Urdu use the Arabic script
- Other languages (German for example)need decompounding
- We can implement script-specific processing but not yet language specific processing

New Features

- Help users with large result sets
  - Facets based on MARC metadata
  - Advanced search based on MARC metadata
- Improve relevance ranking
  - Use added weight for matches in MARC metadata
  - “Search within a book” back-end moved to Solr: Relevance ranked results when searching within a book

Facets based on MARC Metadata

Full-text Search | bacon | Full view only | Find | Advanced Full-text Search

Refine Results

Viewability
- Limited (search only) (1,027,590)
- Full View (643,305)

Subject
- United States (52,817)
- Great Britain (34,098)
- Agriculture (23,932)
- Science (23,316)
- Education (19,455)
- more...

Author
- Shakespeare, William, 1564-1616. (3,112)
- Scott, Walter, Sir, 1771-1832. (1,510)
- United States. (1,457)
- Voltaire, 1694-1778. (1,334)
- United States. Patent Office. (1,298)
- more...

Language
- English (1,332,659)
- French (138,005)
- German (111,596)
- Spanish (42,447)
- Undetermined (39,285)
- more...

Place of Publication

Search Results: 1,670,895 items found for : **bacon** in *Everything*

25 per page | Prev **1** 2 3 4 5 6 7 8 9 10 ... 66836 >> Next

Select all on page | Select Collection | Add Selected

- Bacon Families Association quarterly. v4-8 1984-1989
  Published 1984
  Catalog Record | Limited (search-only)
- The Bacon family from England to America : from Washington County, New County, California / Edw. J. Young.
  by Young, Edward J.
  Published 1965
  Catalog Record | Limited (search-only)
- Bacon Families Association quarterly. v.9-12 1989-1994
  Published 1984
  Catalog Record | Limited (search-only)
- The flower of night : for cello and piano / by Ernst Bacon.
  by Bacon, Ernst, 1898-1990.
  Published 2002
  Catalog Record | Limited (search-only)
- Bacon's Adventure / adapted by Elizabeth Bacon Rodewald.
  by Bacon, Herbert Marion, 1867-
  Published 1947
  Catalog Record | Full view
- Bacon's Adventure / adapted by Elizabeth Bacon Rodewald.
  by Bacon, Herbert Marion, 1867-
  Published 1947

1.6 million hits for Bacon!

- Bacon the food?
- Francis Bacon?
- Roger Bacon?

Facets: Scalability and Usability

- Facets work best with small value sets of 10-30 values. LCSH provides millions of values!
- Solr will show the top N facet values by facet count
- The facet counts are based on all the results
- In HT full-text search large result sets (over 10,000 records) are common, the facet counts are based on the entire result set, not the most relevant of those 10,000+ records

Irrelevant Facets

Jaguar car, Jaguar animal, Jaguar OS?

Full-text Search | jaguar | Full view only | Find | Advanced Full-text Search

Refine Results

Viewability
- Limited (search only) (144,871)
- Full View (33,247)

Subject
- Science (3,184)
- United States (2,857)
- English language (2,370)
- Encyclopedias and dictionaries. (2,157)
- Great Britain (2,128)
- Mexico (2,056)
- Geography (2,029)
- Art (2,004)
- Automobiles (1,940)
- Ethnology (1,939)
- Natural history (1,913)
- Latin America (1,750)
- France (1,619)
- Science Periodicals. (1,607)
- Geography Periodicals. (1,527)

Search Results: 178,118 items found for : **jaguar** in *Everything*

25 per page | Prev **1** 2 3 4 5 6 7 8 9 10 ... 7125 >> Next

Select all on page | Select Collection | Add Selected

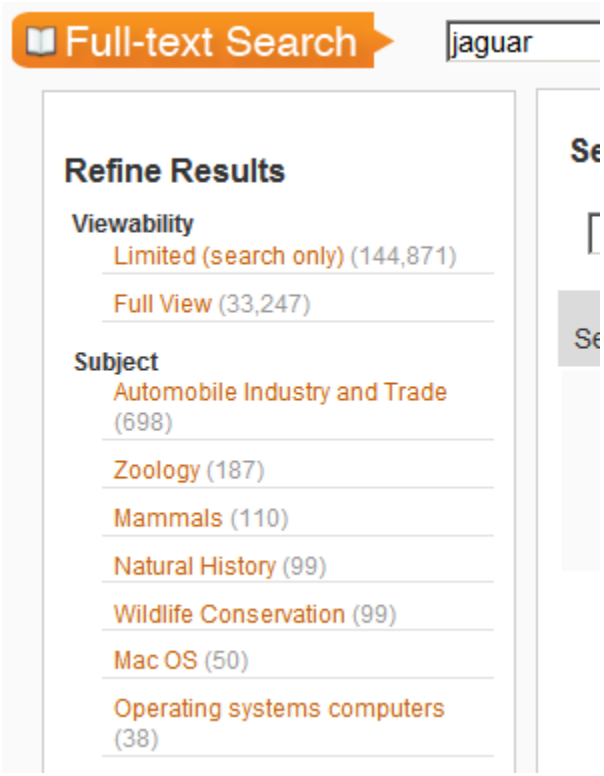
The will to win : John Egan and Jaguar / John Underwood.
by Underwood, John.
Published 1989
Catalog Record | Limited (search-only)

Jaguar [by] Lord Montagu of Beaulieu.
by Montagu of Beaulieu, Edward John Barrington Douglas-Scott-Montagu, Baron, 1926-
Published 1975
Catalog Record | Limited (search-only)

Jaguar, a biogrpahy
Published
Catalog Record | Limited (search-only)

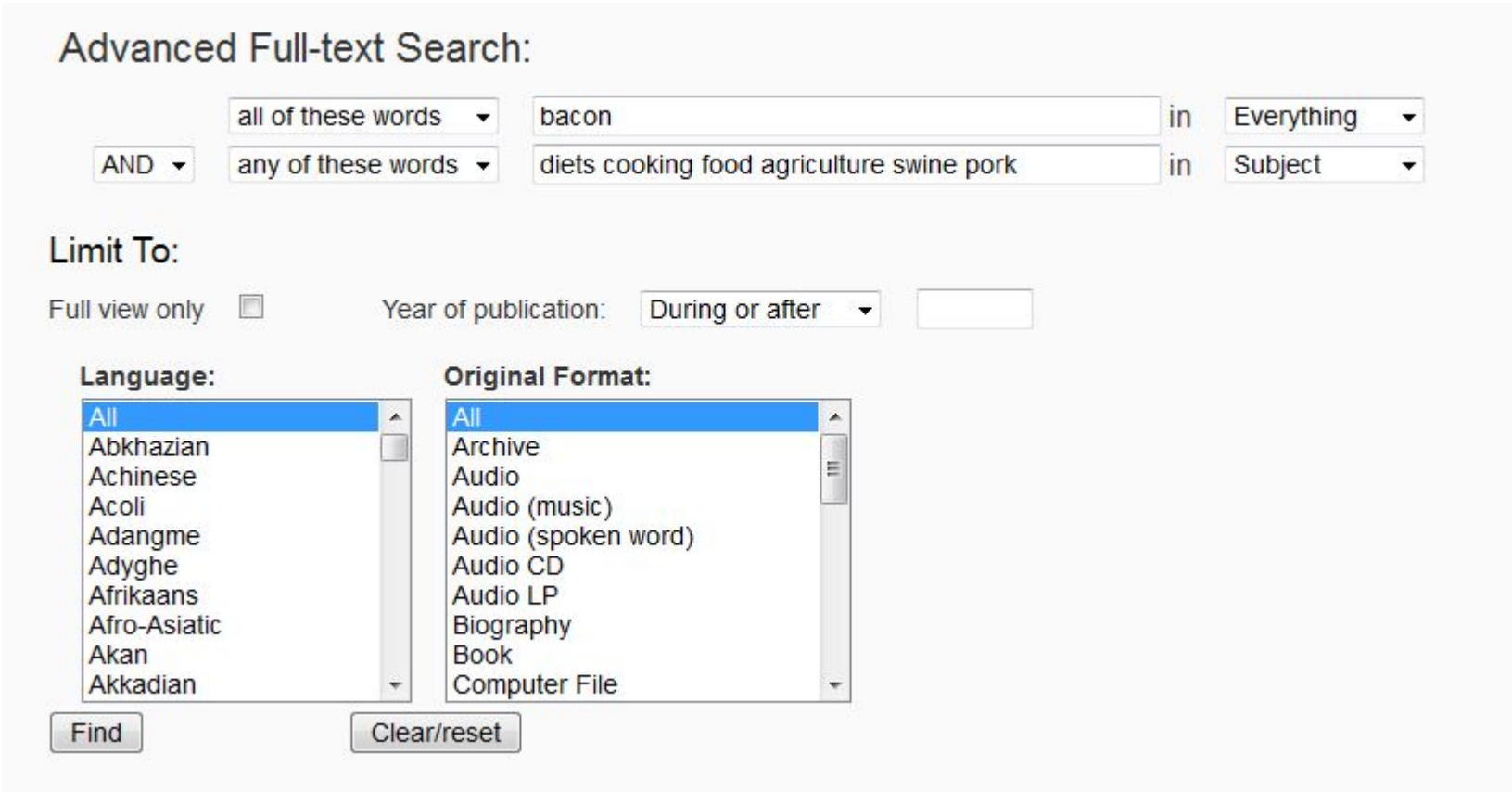
Jaguar, the history of a great British car / Andrew Whyte ; foreword by Sir William Lyons.
by Whyte, Andrew
Published 1985

Relevant facets



http://lucene.472066.n3.nabble.com/Getting-facet-counts-for-10-000-most-relevant-hits-td3363459.html

Advanced search based on MARC metadata and OCR



Advanced search based on MARC metadata and OCR

Full-text Search | Full view only | Find | Advanced Full-text Search

Refine Results

Viewability
- Limited (search only) (27,808)
- Full View (17,095)

Subject
- Agriculture (23,932)
- Agriculture Periodicals. (3,253)
- Agriculture Periodicals Periodicals. (2,590)
- Food (2,340)
- Agriculture. (1,418)
- Food industry and trade (1,359)
- Cooking, American. (1,292)
- Nutrition (1,275)
- Agriculture United States Periodicals. (1,159)
- Cooking (1,085)
- Cookbooks (1,028)
- Swine (978)
- United States. (900)
- Agriculture United States. (884)
- Cooking. (860)
- Diet (726)
- Food. (725)
- Agriculture, Cooperative (713)
- Food supply (600)
- Agriculture Economic aspects Periodicals. (597)
- Agricultural experiment stations (583)

Search Results: 44,903 items found for

all of these words: **bacon** in *Everything*

AND *any of these words:* **diets cooking food agriculture swine pork** in *Subject*

Revise this advanced search

25 per page | Prev **1** 2 3 4 5 6 7 8 9 10 ... 1797 >> Next

Select all on page | Select Collection | Add Selected

Report of the Reorganisation Commission for Pigs and Bacon.
by Great Britain. Agricultural Marketing Reorganisation Commission.
Published 1956
Catalog Record | Limited (search-only)

The bacon curing industry; report of the committee of enquiry into the bacon curing industry set up by the Economic Development Committee for Food Processing (now part of the EDC for Food Manufacturing)
by Great Britain. Economic Development Committee for the Food Manufacturing Industry.
Published 1968
Catalog Record | Limited (search-only)

No-nitrite meats : handle carefully.
by United States. Food Safety and Quality Service.
Published 1979
Catalog Record | Full view

Pigs and bacon curing; a practical manual on the various methods of feeding pigs, curing hams and bacon, and utilising the by-products of swine.
by Davies, Robert Evan, 1877-
Published 1948
Catalog Record | Limited (search-only)

Economic series. no.37 1932

Relevance ranking with MARC Metadata

- Example (simplified)

```
q= _query_:"{!edismax
qf='ocr^5000
+AnyMARCField^2
+TitleField^50
+AuthorField^80
+SubjectField^50
+mm='100%'
tie='0.9' } bacon“
```

- mm = minimum must match
- tie = weight of all fields other than the field with maximum score

* see: http://wiki.apache.org/solr/DisMaxQParserPlugin

Relevance ranking with MARC Metadata

- Example (actually used)
  - q= _query_:"{!edismax qf='ocr^50000+allfieldsProper^2+allfields^1+titleProper^50+title_topProper^30+title_restProper^15+title^10+title_top^5+title_rest^2+series^5+series2^5+author^80+author2^50+issn^1+isbn^1+oclc^1+sdrnum^1+ctrlnum^1+id^1+rptnum^1+topicProper^2+topic^1+hlb3^1+fullgeographic^1+fullgenre^1+era^1+'
  - pf='title_ab^10000+titleProper^1500+title_topProper^1000+title_restProper^800+series^100+series2^100+author^1600+author2^800+topicProper^200+fullgenre^200+hlb3^200+allfieldsProper^100+' mm='100%25'
  - tie='0.9' } bacon"
- qf = query
- pf = query words as a phrase query
- pf2 and pf3 = disjunction of 2 and 3 word combinations as phrases
- mm = minimum must match
- tie = weight of all fields other than the field with maximum score

Tuning Relevance Ranking

- How should we adjust the relative weight between OCR (full-text) and MARC fields?
- What are the Solr settings affecting ranking?
- How do you tell when one group of settings provides more relevant results than another
  - Need to test against some set of queries to insure that improving results for one query doesn't make results for other queries worse

Solr relevance knobs

- Query expansion
- Field weighting
- Boost queries or dismax queries
- Expert
  - length normalization
    - SweetSpotSimilarity
  - Solr relevance formula tf *idf
    - write custom Similarity class or use Solr 4.0/Trunk

Testing Relevance

- Use test collection and relevance judgements
  - TREC/XML book track*
  - Are queries similar to your user's queries?
  - Is collection similar to your collection?
  - What is the appropriate metric (P@10) other real-world issues?
- Relevance is complicated and different users will consider different sets of documents relevant
  - Ambiguous queries

*https://inex.mmci.uni-saarland.de/tracks/books/

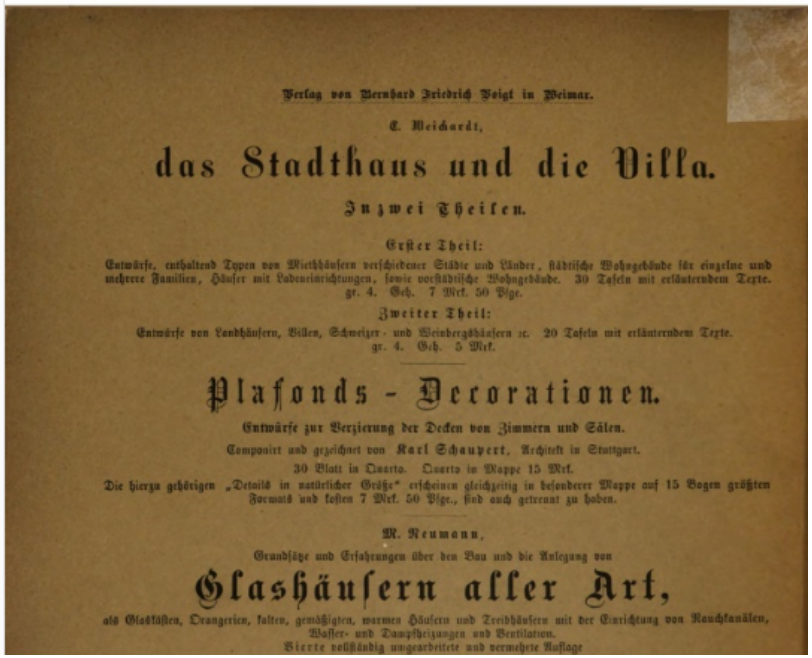
Testing Relevance

- Use clickthrough logs and A/B testing*
  - clickthrough not always measure of relevance
  - A/B testing controversial in library settings
  - Use interleaving to overcome order effects
- Usability testing?
- Beta testers ?

*Radlinski, F., Kurup, M., & Joachims, T. (2008). How does clickthrough data reflect retrieval quality? In Proceeding of the 17th ACM conference on Information and knowledge management, CIKM '08 (pp. 43–52). New York, NY, USA: ACM. doi:10.1145/1458082.1458092

“Dirty” OCR: Causes

- noise, varible ink, bleedthrough, markings by users
- tight bindings affect scanning, words near binding or edge can be distorted
- background color, fading, water or mold damage/stains
- bad scanning, blurred, operator scanning
- 400+ languages (x volumes in more than one language)
  * Feng and Manmatha (2010?)



Verlag von Bernhard Friedrich Voigt in Weimar.

E. Weichardt,

das Stadthaus und die Villa.

In zwei Theilen.

Erster Theil:

Entwürfe, enthaltend Typen von Miethhäusern verschiedener Städte und Länder, städtische Wohngebäude für einzelne und mehrere Familien, Häuser mit Ladeneinrichtungen, sowie vorstädtische Wohngebäude. 30 Tafeln mit erläuterndem Texte. gr. 4. Geh. 7 Mrk. 50 Pfge.

Zweiter Theil:

Entwürfe von Landhäusern, Villen, Schweizer- und Weinbergshäusern ꝛc. 20 Tafeln mit erläuterndem Texte. gr. 4. Geh. 5 Mrk.

Plafonds - Decorationen.

Entwürfe zur Verzierung der Decken von Zimmern und Sälen.

Componirt und gezeichnet von Karl Schaupert, Architekt in Stuttgart.

30 Blatt in Quarto. Quarto in Mappe 15 Mrk.

Die hierzu gehörigen „Details in natürlicher Größe“ erscheinen gleichzeitig in besonderer Mappe auf 15 Bogen größten Formats und kosten 7 Mrk. 50 Pfge., sind auch getrennt zu haben.

M. Neumann,

Grundsätze und Erfahrungen über den Bau und die Anlegung von

Glashäusern aller Art,

als Glaskästen, Orangerien, kalten, gemäßigten, warmen Häusern und Treibhäusern mit der Einrichtung von Rauchkanälen, Wasser- und Dampfheizungen und Ventilation.

Vierte vollständig umgearbeitete und vermehrte Auflage

“Dirty” OCR: Effects on OCR

- OCR engine produces no OCR for page or volume
- OCR engine misrecognizes language/script and produces garbage
- OCR engine interprets images as text and produces garbage

“Dirty” OCR: Effects on Search

- Large number of garbage terms result in over 3 billion unique terms per index.
  - blew up lucene *
  - increased memory requirement to 32GB until we changed *termIndexInterval*
  - MultiTerm queries (wildcard, prefix/truncation) performance unusable
  - Problem for spelling suggestion
  - Documents with lots of garbage from images have incorrect document length normalization

* http://www.hathitrust.org/blogs/large-scale-search/too-many-words-again

“Dirty” OCR: Effects on Search

- Missing OCR text not retrievable, some volumes not searchable
- Incorrectly recognized text not retrievable
- Term statistics and relevance ranking skewed
- If students underline important parts and underlines cause misrecognition, important parts less retrievable

“Dirty” OCR:Solutions?

- IR techniques for retrieving dirty OCR such as ngrams, fuzzy matching, error correction and query garbling have disappointing performance or are not scalable (TREC 4&5 and Savoy 20xx)
- Removing hapax (terms that occur only once) will likely remove huge amount of real words including proper names
- Correction algorithms are designed for single language
- heuristics will remove some fraction of bad words but must work with 400 languages
- Google improves OCR engine and re-OCR’s in a 2 year cycle

Collection Builder

HATHI TRUST Digital Library

Login Help Feedback

Home About Public Collections My Collections

Full-text Search

Full-text Search "library of universal knowledge" Full view only Find

Search Results: 950 items found for **"library of universal knowledge"** in the full text of all items (in 0.024 sec.)

All Items (950) Only Full view (642)

25 per page Previous 1 2 3 4 5 6 ... 37 38 Next

Select all on page Select Collection Add Selected

General encyclopedias in print., 1963
Published 1963
Catalog Record Limited (search-only)

History by the topical method.
by Woody, John W. [from old catalog]
Published 1889
Catalog Record Full view

Erasmus & Henry VIII.
by Merle d'Aubigné, J. H. 1794-1872.
Published 1884
Catalog Record Full view

General encyclopedias in print., 1964
Published 1963
Catalog Record Limited (search-only)

Anglo-American general encyclopedias : a historical bibliography, 1703-1967 /
by Walsh, S. Padraig, 1922-

Collection Builder

HATHI TRUST Digital Library

Login Help Feedback

Home About Collections My Collections

HathiTrust Digital Library - Collections Create New Collection

Login to create and save permanent collections or see your private collections.

All | Recently Updated | Featured | My Collections

Find a collection

Collections with at least (all items)

Sort by: Collection Title

Showing 627 collections

'Generations'
Owner: quoddy
2 items
last updated: 10/11/10

1850s-1870s Japan
Travel narratives, photography and treaty ports
Owner: theisenc (University of Michigan)
20 items
last updated: 10/18/10

19-20th C. Psychology Texts-Gen
A collection of selected psychology books in various areas and from the 19th century to late 20th century (mostly full-view).
Owner: Michael Palij (University of Michigan)
127 items
last updated: 09/05/11

19th c cookbooks
Owner: nonesuch (University of Michigan)
184 items
last updated: 07/28/11

19th c periodicals
Owner: nonesuch (University of Michigan)
23 items
last updated: 03/16/10

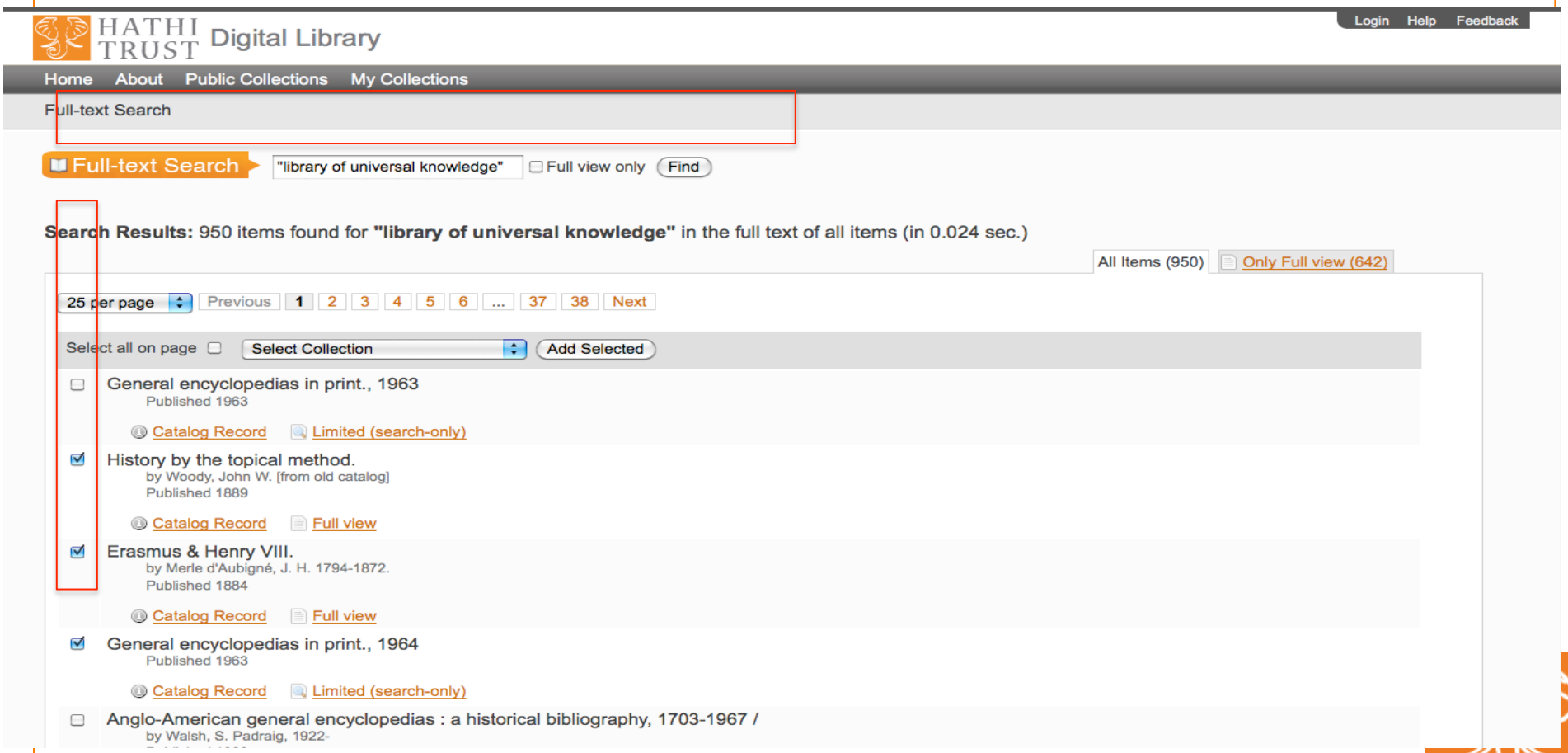
19th c. Ottoman binding
Bindings made for Sultan Abdul Hamid II at the palace workshop at Yıldız Sarayı
Owner: iemilian
1 item
last updated: 05/27/10

Featured Collection

Adventure Novels: G.A. Henty

Novels from the Henty Collection and books about George A. Henty.

Collection Builder



HATHI TRUST Digital Library

Login Help Feedback

Home About Public Collections My Collections

UMPress (*)'s UM Press collection

Search in this collection Find

All Items (1404) Only Full view (782)

Collection Name: UM Press

Collection Description: The Univ. of Michigan Press titles available in HathiTrust

Visit UM Press website

Sort by: Title A-Z 25 per page Previous 1 2 3 4 5 6 ... 56 57 Next

Select all on page Select Collection Copy Selected

[The relationships of the eastern North American crayfishes, with a revision of the genus Faxonius]
by Creaser, Edwin P.
Published 1933
Catalog Record Full view
In My Collections: ---

"Eastern magnificence & European ingenuity" : clocks of late imperial China / Catherine Pagani.
by Pagani, Catherine.
Published 2001
Catalog Record Full view
In My Collections: ---

The "Goldhagen effect" : history, memory, Nazism--facing the German past / edited by Geoff Eley.
Published 2000
Catalog Record Limited (search-only)
In My Collections: ---

1986. Crashaw; a study in baroque sensibility.
by Warren, Austin, 1899-1986.
Published 1957
Catalog Record Full view
In My Collections: ---

9226 Kercheval; the storefront that did not burn.
by Milio, Nancy.
Published 1970
In My Collections:

Collection Builder

- http://babel.hathitrust.org/cgi/mb
- Need to update one field, i.e. add user's collection number to the collection field
- Solr needs to re-index whole record including 800MB+ of OCR
- Solution #1 Keep collection-unique id table in database send query with ids in collection to Solr
  - q1=ocr:bacon AND id:(2 OR 5 …or 1023)
  - not scalable
- This is one of many use cases for the ability to merge a list of ids from the result of a database or other external system search into a Solr query.
  - See: http://lucene.472066.n3.nabble.com/filter-query-from-external-list-of-Solr-unique-IDs-tc1709060.html#a1709716
  - https://issues.apache.org/jira/browse/SOLR-139
  - https://issues.apache.org/jira/browse/LUCENE-1879

Collection Builder

- Solution 2
- Bite the bullet and update the entire record every time someone add a record to their collection. i.e. update the coll_ids field
- To minimize the latency between updating a record and having it be available for searching we needed a separate small Solr index configured/optimized for minimal update latency

Collection Builder

- New requirement: medium to large collections i.e. 10,000 – 100,000 volumes
- Solution 3/Current hack
  - Eliminate small Solr and use same Solr as full-text search
  - If collection is under 1000 (1024) items in size store list of unique ids in database and send id list to Solr. Collection update just changes a row in the database and requires no Solr re-indexing so immediate collection update
  - If collection is over 1000 items, add the items to the nightly indexing queue with an update to the coll_ids field. Message to user saying they will have to wait overnight to be able to search new items

Thank You !

Tom Burton-West
tburtonw@umich.edu

www.hathitrust.org/blogs/large-scale-search