



IWAW'06

Gordon Mohr, Internet Archive

September 22, 2006

INTERNET ARCHIVE

WayBackMachine

- > 60 billion captures
- > 600TB compressed
- 10 years (1996-)
- ...but only lookup-by-URL

Prior IA search tries

- ‘AndyVista’ ~ 2002 ~ one Alexa engineer
 - Word presence only: no ranking, phrases
 - Hits by year, in chronological order (!)
 - Never promoted; not open source; not maintained or scaled
- ‘Recall’ ~ 2004 ~ one Stanford CS postdoc
 - Concept extraction, Lisp
 - Tiny index, cool time-frequency views
 - Only subset of archive
 - Not open source; never scaled; fell apart when creator moved on

Time is Ripe

- What’s stopped us in the past:
 - fragile code (closed source, departing authors)
 - underprovisioned (little slack/funding/staff for indexes, servers)
- Now finally:
 - open-source search scaling up: Nutch
 - stable staff w/ experience: collection search
 - cheaper hardware
 - outside interest as keen as ever

Still: 60 billion/ 600TB is a lot!

- Fortunately:
 - 1/2 of archive *by date*
is only 1/12 *by captures*
 - even smaller *by size*
- Ergo: “20th Century Find”
 - 1996-2000: the 20th Century Web
 - IA + Library of Congress + Yahoo (Thanks!)

High risk, reward

- Very much a research project:
 - new, changing software
 - new, unproven hardware
 - unknown parameters until we're deeper
 - unknown traffic
- But could be:
 - largest Nutch/open-source search ever?
 - largest searchable web archive?
 - first time since 2000 this content has been searchable
 - researcher bonanza
 - driver for improving component open source projects

Software / Web Content Needed

- Hadoop 0.6 (++)?
 - MapReduce cluster processing
- Nutch 0.8 (++)?
 - Open source search
- NutchWAX 0.6.0 (++)?
 - UI, exclusion work needed for broad deployment
- NG Wayback Machine 0.8.0 (++)?
 - Open source ‘wayback’ browsing
- 20thCF Collection customization
 - entry & support pages, bells & whistles

The 20th Century Web Corpus

- Internet Archive ARC files
 - 100MB concatenated HTTP response transcripts
 - 526,469 known by name, likely from 1996-2000
 - ~22TB
 - URI Captures:

		HTML	ALL
1996		20,475,783	32,168,637
1997		214,802,568	318,294,397
1998		195,422,208	245,919,083
1999		757,339,164	847,540,628
2000	(2.5B (est))	3,328,040,041	
		=====	=====
	(est) 3,700,000,000		4,771,962,786
 - Stages:
 - 1996(easy!), 1997 (x10), 1998(=), 1999(x3), 2000(x4)

Deployment data size estimates

- Deduplication conjecture:
 - Exact duplicates are 40% or more of Archive, so effective deduplication yields 13TB corpus
 - ARC successor draft format, WARC, offers duplicate backpointers
- Text index conjecture:
 - Nutch index runs up to 10% of ARC size, will be < 2TB
- Location index conjecture:
 - Wayback browse index runs 3% of ARC size, will be < 500GB

Traffic

- Wayback Machine
 - 100,000 lookups-by-URL each day
 - > 50 URL retrieval requests every second
 - ~15% of the dated URL retrieval requests are for 1996-2000
- Traffic conjecture:
 - search will mean several times that traffic to this collection (even after novelty wears off)
 - WAG: 3x
 - > 45,000 queries per day
 - > 23 retrieval requests per second.

Hardware

- New IA “power rack”
 - 35x (1U Athlon64 X2 3800+ / 4GB RAM / 4x500GB HD / gigabit NIC)
- Hardware conjecture:
 - Power rack enough for all phases & services
 - usable storage > 60TB
 - 2 copies ARC corpus (2x13TB=~26TB)
 - 10x text index replication (10x2TB=20TB)
 - 10x URI-location index (5x500GB = 2.5TB)
 - Serving: over-provisioned?

The Phases

- Preparation ~70% done
 - Assessing, provisioning, researching, designing, collecting
- Implementation ~30% done
 - Improving components, indexing, integrating, customizing
- Testing
 - Testing, deploying
- Launch
 - Launching, maintaining

Open technical questions

- Index organization
 - Optimization?
 - Distribution?
- Practical serving capacity?
- Link-juice across time?
- Nutch or WM-BDB location index?
- Non-text indexing by URL, inlink text?

Open usability questions

- Time results comprehended?
- Visualization/timeplots needed?
- New WM features comprehended?

Open policy questions

- Who'll be angry?
- Will our existing policies hold up?
 - Retroactive robots compliance
 - Prompt manual excludes

Risks

- Hardware flaky, insufficient
- Software doesn't stabilize & scale
- Capacity & traffic surprises
- Avalanche of complaints

Future

- Creation & service should generate several lessons learned papers
- Annual updates -- always 5 years behind?