Archiving Newspaper Websites: A Case Study of the Chicago Tribune

Kalev Leetaru – leetaru@illinois.edu

Archiving Newspaper Websites

- Little hard data on how much content goes up on a newspaper website daily and how hard it is to archive on a regular basis.
- Case study of the Chicago Tribune to explore this in detail.

What's New(s)?

- Hardest part is getting an inventory of what gets posted to the site each day.
- No easy master inventory list of the URLs of new articles each day.
- Tribune itself uses multiple content management systems and doesn't have a single point of overview to its site.

RSS / Site Maps

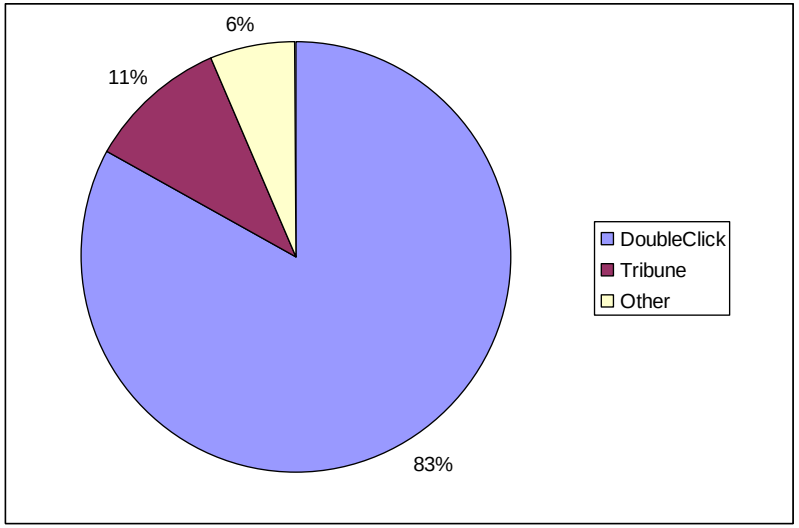
- Some sites like CNN offer strong date-sorted RSS feeds. Already in machine-friendly format. Just download every 30 minutes and you have a complete list of all new content on the site.
- Google News Sitemaps service allows news sites to provide a list of all new content to select users like Google News.

Gateway Pages

- Tribune has neither (has RSS feeds, but they are poor).
- Must manually identify the primary gateway pages for the site (main pages of each topic). There are 105 for the Tribune as of October 2010.
- Most recent X articles are listed on the gateway page. Must download them on a quick interval: every 30 minutes or you miss articles for some sections.

Exploring the Tribune

- Downloaded all 105 Tribune gateway pages every 30 minutes from 9/15/2010 to 10/19/2010: total of 136,605 snapshots.
- 83% of links are to the DoubleClick.net advertising network, with just 11% of links pointing to Tribune pages.



Tribune Findings

- Articles stay up, but the links to those articles last from 18 hours to 7 days, with an average lifespan of 56 hours.
- Roughly 39% of articles are linked for less than a day: if you miss downloading the gateway page in that timespan, you won't ever know that article every existed.
- Average of 735 new articles posted daily.
- Thursdays have highest adds, Sundays have the fewest.

Tribune Findings

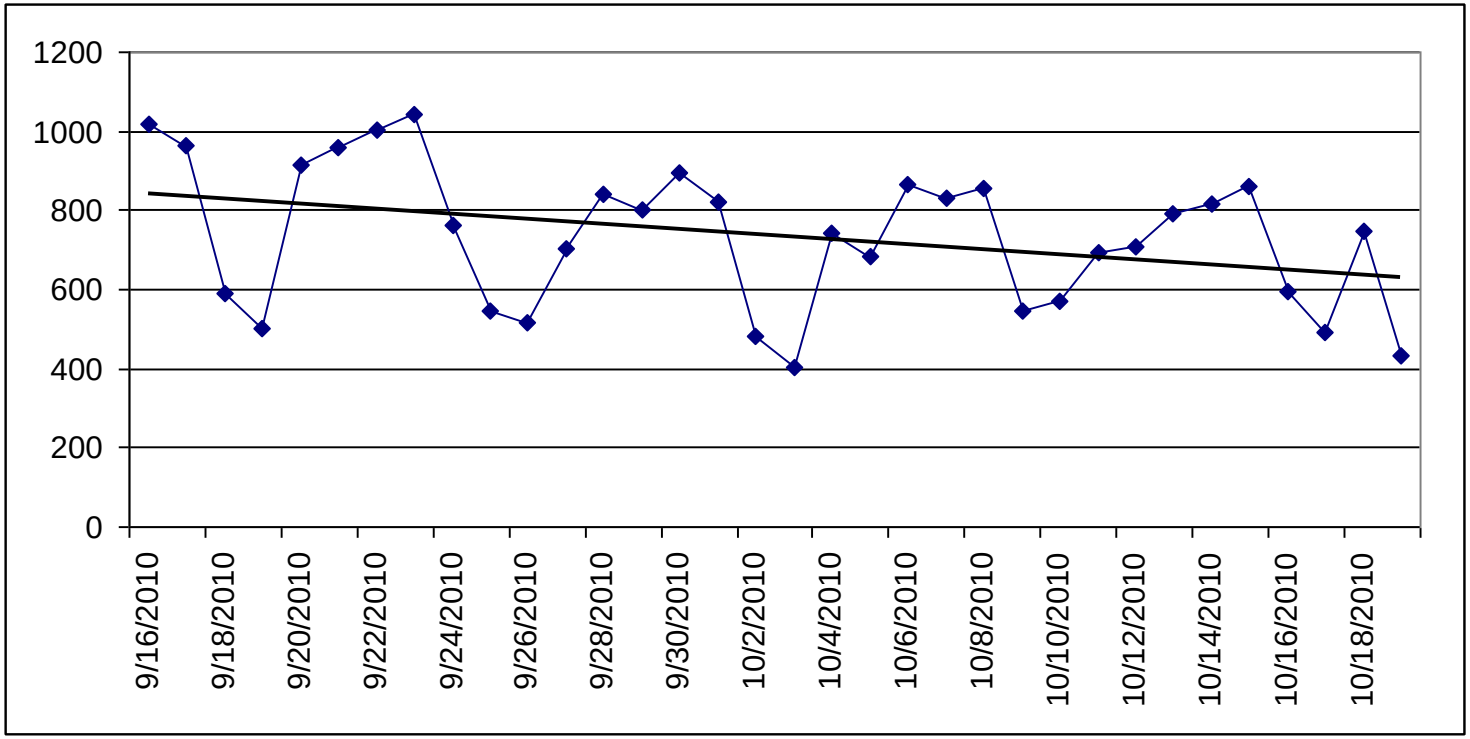


Figure 3 - Number new Tribune links seen by day

Tribune Findings

- Content sections have high stratification: some have just a few articles added a day, others have very high add rates, adding new content 24/7 at a very high velocity.

Tribune Findings

Table 1 - Gateway pages ordered by average link lifespan

Section	Total	Tribune	%Tribune	DoubleClick	Lifespan
/business/	9003	2180	24.21	5134	18.97
/sports/	9940	3384	34.04	5192	20.21
/entertainment/celebrity/	5403	1090	20.17	3900	22.08
/features/horoscopes/	5503	244	4.43	5200	25.08
/	13030	3855	29.59	5200	25.48
/technology/deals/	3345	689	20.60	2602	25.61
/news/nationworld/	5704	1628	28.54	3900	29.38
/news/local/chicago/	4927	426	8.65	3900	29.77
/sports/football/bears/	5310	1083	20.40	3903	35.15
/sports/college/	5362	1118	20.85	3897	35.19
/health/	5394	789	14.63	4096	36.19
/news/education	4477	464	10.36	3894	36.62
/news/opinion/blogs/	8041	735	9.14	3892	37.48
/news/opinion/share/	5573	345	6.19	5168	38.01
/entertainment/	5762	1458	25.30	3848	38.10
/news/politics/	4710	619	13.14	3903	38.84
/news/local/	6309	1057	16.75	3879	43.35
/news/opinion/	5037	831	16.50	3900	45.79
/news/columnists/all/	4867	908	18.66	3898	47.37
/news/columnists/all/	4867	908	18.66	3898	47.37
/news/corrections/	4301	337	7.84	3903	50.16
/sports/baseball/whitesox/	5028	850	16.91	3900	50.73
/sports/highschool/	4847	838	17.29	3798	52.01

Conclusions

- News sites can offer high-quality RSS and Sitemaps: benefits libraries and benefits them from increased consumer awareness of content.
- Otherwise really need newspapers to engage with libraries to provide them content lists, too hard to do externally via monitoring, but newspapers themselves don't even know their content due to fragmented content management infrastructures.

Thank You

- Kalev Leetaru – leetaru@illinois.edu
  - http://contentanalysis.ichass.illinois.edu
  - I-CHASS (Institute for Computing in the Humanities, Arts, and Social Sciences)
  - NCSA (National Center for Supercomputing Applications)
  - University of Illinois