Forward Experiment

Investigation of content drift and link rot in Web-at-large references in a sample of two corpora of scholarly research papers

James Powell

High level outline

- First part explores available metadata for web-at-large references
  - Link rot (HTTP status, content length, mime type)
  - Content drift (content length, mime type)
- Second part focuses on content similarity from interval to interval
  - Content length and content drift
  - Various similarity measures – Hash, Vsim, Graph Edit?
  - Exploring content drift in aggregate for a given publication, e.g. visualizing drift, calculating a “drift score” based on aggregate drift among web at large references for a paper

High level outline … cont'd

- Third part looks at Machine Learning algorithms applied to this data
  - Can machine learning algorithms learn to identify/anticipate either of these problems?
  - Are there associations among the data that other methods didn't detect?
  - What are possible uses for ML in this context?

Question(s)?

- Is it the case that for this sample that both link rot and content drift increase with the passage of time, as has previously been noted?
- What do the link availability dynamics (changes in HTTP status and content availability) tell us about the potential for content drift for a given resource?
- Do status chains exhibit any predictable patterns over time?
- Is content change dramatic? Fast or slow? Is it ever prone to reversal?
- Does the information content of a cited resource increase, decrease or stay the same over time?
- Are some top level domains more volatile than others?

Other similarity measures

- Similarity hash comparisons
- Similarity hash results as compared to preliminary feature vector comparisons
- Analysis of citations when status code changes during review period
- Status and file size changes
- What else?

Describe experiment

- A sample of arXiv and PLOS papers were identified for this experiment
- For each collection for a period of time (whenever), all links that were not DOI/scholarly links were extracted
- These links were harvested on a regular basis
- At each harvest a copy was pushed to a Web archive, and saved locally.
- The initial HTTP status of a request along with size and mime (in fact, entire HTTP header) were stored

Descriptive stats for experimental data

- ArXiv : #source papers
  - 17,158 citations with 14, 966 unique URLs
  - 97.82% were HTML documents
  - 74.06% of cited items changed size
  - 74.50% of originals that were available were “always” available
- PLOS : #source papers
  - 8,284 citations with 6,549 unique URLs,
  - 99.12% were HTML documents
  - 80.84% of cited items changed size
  - 69.00% of originals that were available were “always” available

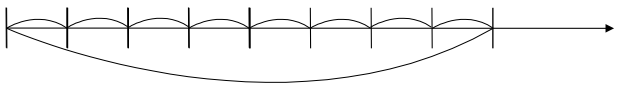
Characterize Data after Filtering

- What filtered and why? Network outage days.
- All papers and references affected by the outage were discarded. Stats.
- Stats of final corpus. Including % of URLs changed according to status code, content length, content type

	Metadata	Content
Compare to original *(examine all URIs)*		Similarity hash Vector-Cosine similarity
Compare with last item retrieved *(examine only URIs that exhibit change over lifespan)*	Byte count Status code	

Illustration of comparison strategies

Metadata: HTTP last status, previous vs current



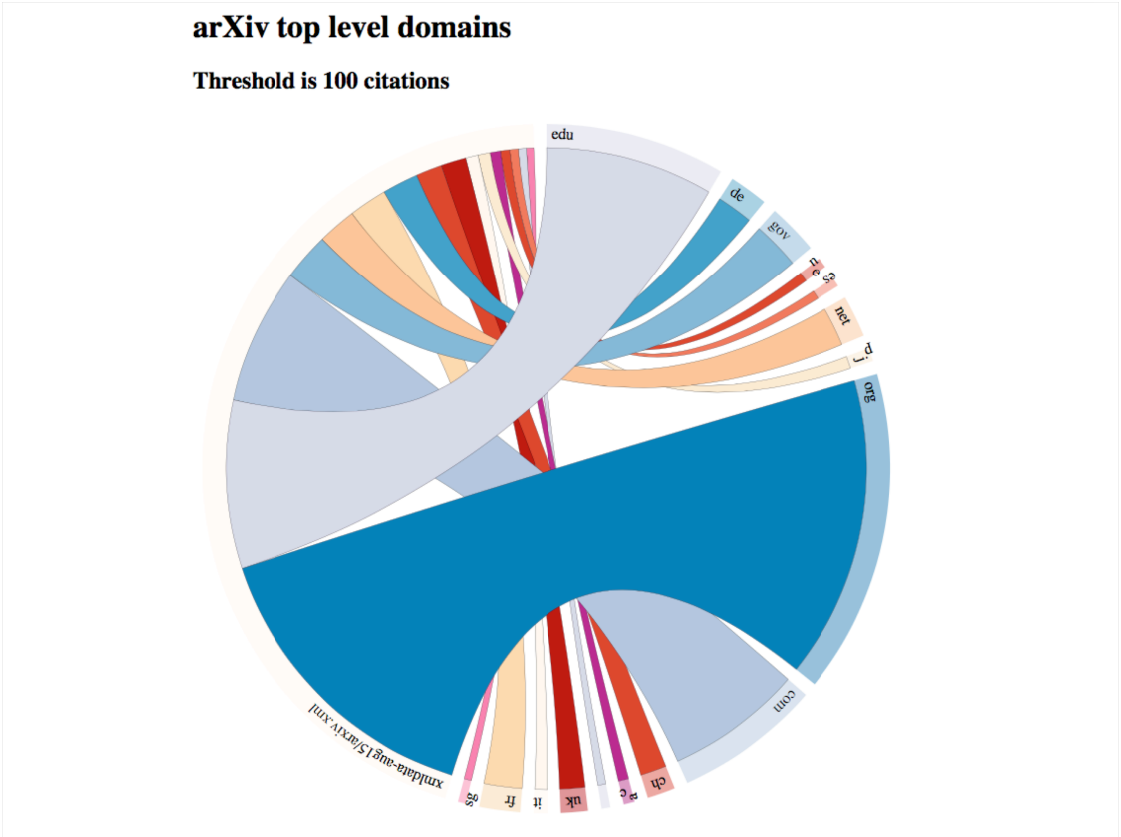
Metadata: Size, subsequent interval compared to initial

Content: Similarity, subsequent interval compared to initial

In sequence vs forward looking/predictive, e.g. previous vs current status, original size vs size of item retrieved at subsequent intervals, etc.

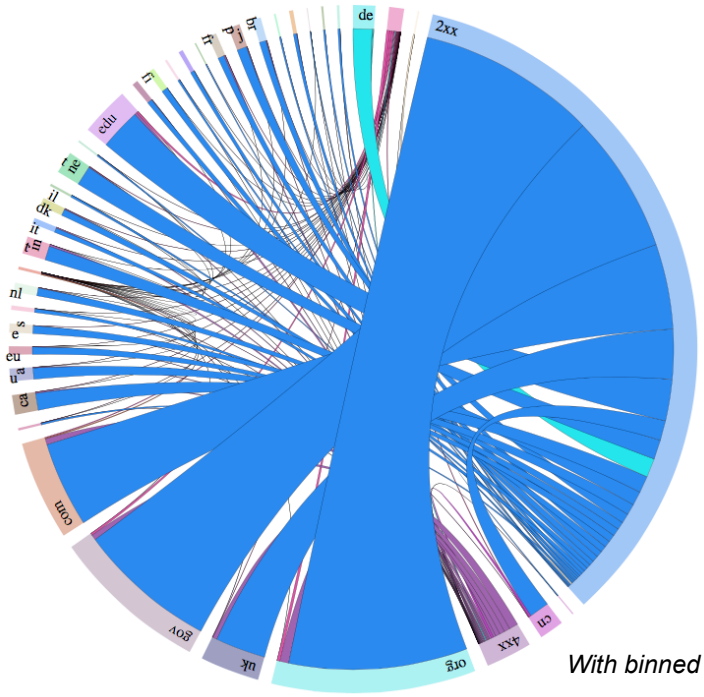
Metadata Extraction

- Procedure description
- HTTP headers: content-length, mime-type, status codes
- Status inspection can be used to characterize link rot
  - Most common top level domains for a corpus
  - Most frequently occurring tlds and their associated statuses
  - Most common tlds yielding a 404



PLOS top level domains

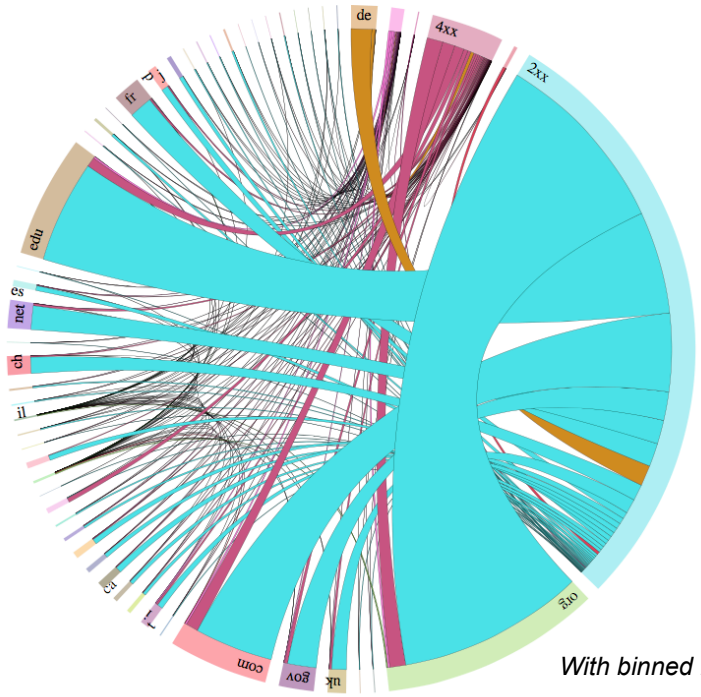
Threshold is 100 occurrences



With binned HTTP statuses

arXiv top level domains

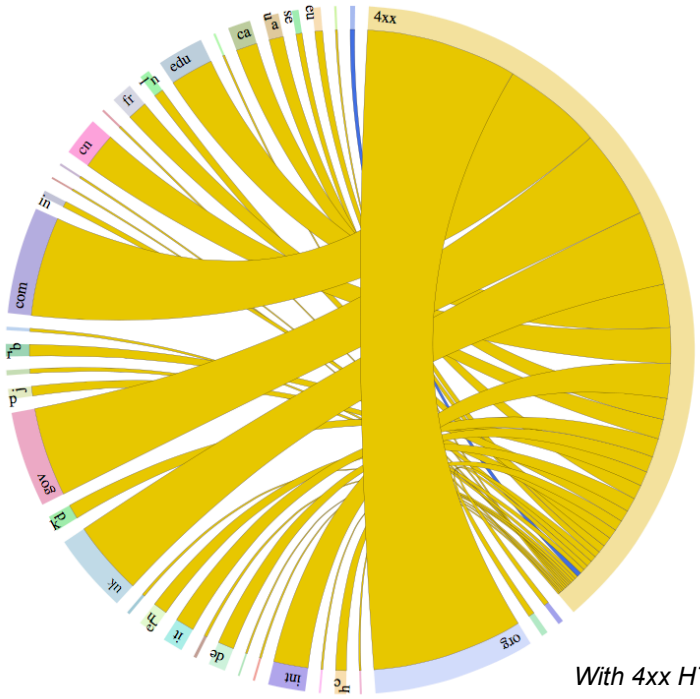
Threshold is 100 occurrences



With binned HTTP statuses

PLOS top level domains

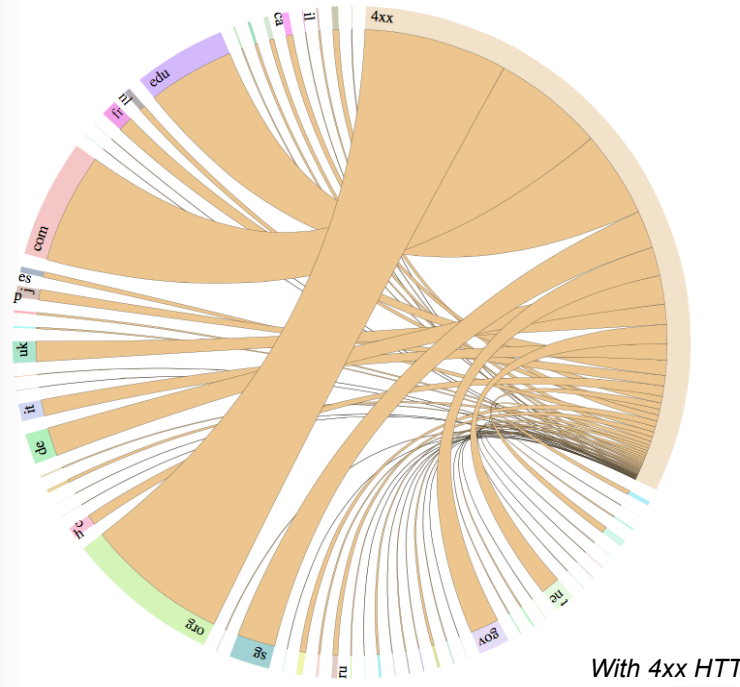
Threshold is 10 occurrences



With 4xx HTTP statuses

arXiv top level domains

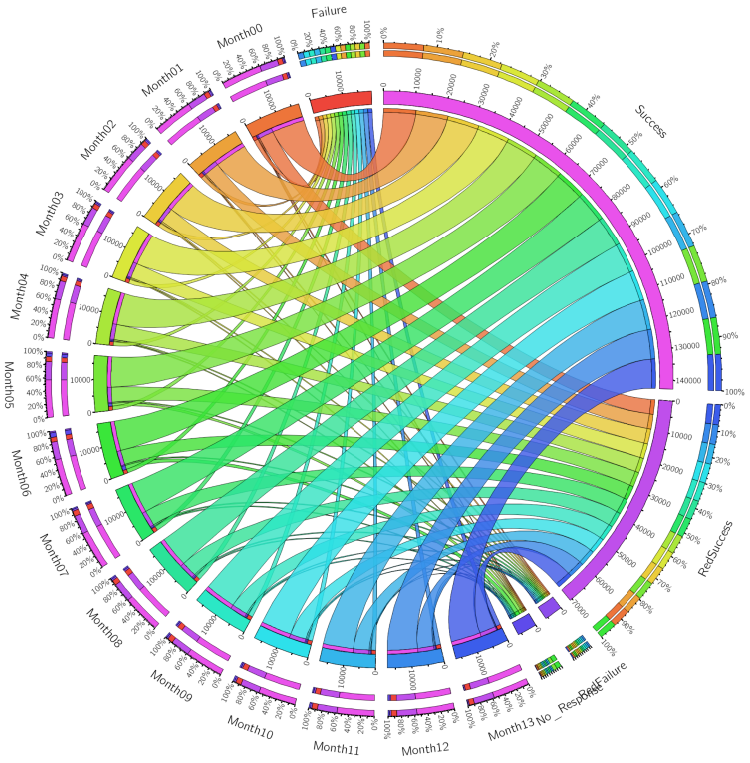
Threshold is 10 occurrences

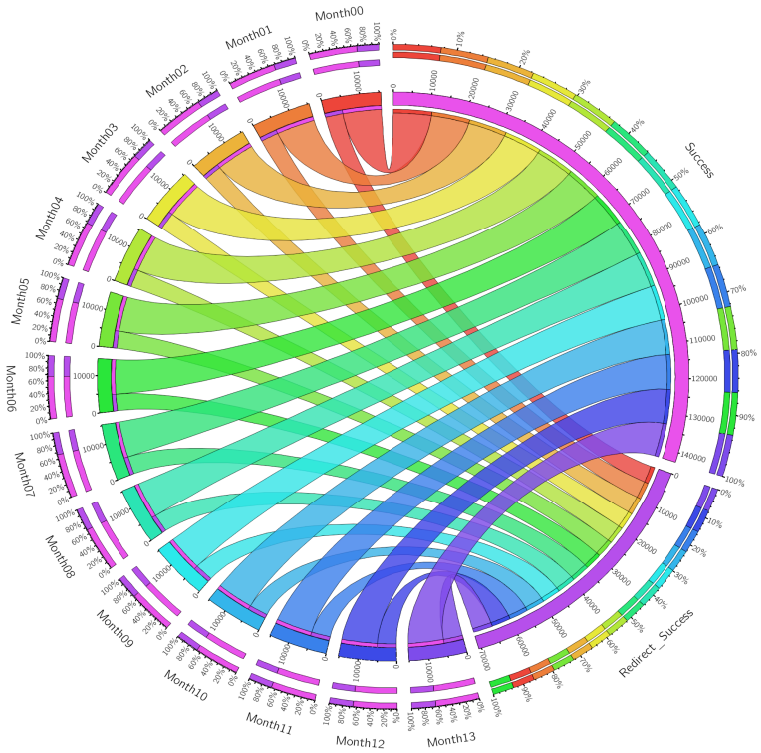


With 4xx HTTP statuses

Circos diagram

- Circos diagrams show relationship of referenced URIs to four categories of status:
  - ☒ Success
  - ☒ Redirect – Success
  - ☒ Failure
  - ☒ Redirect – Failure
- Circos with only successful statuses



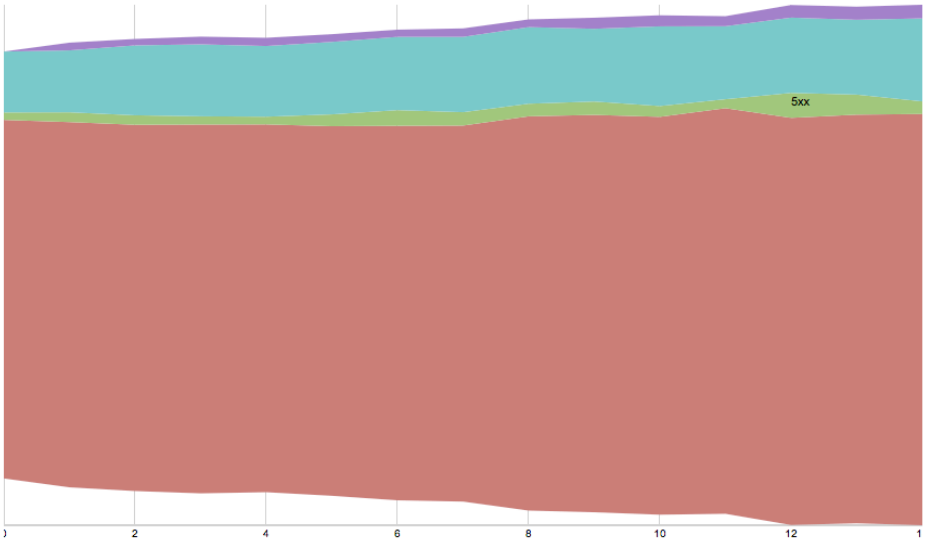


Stream diagrams

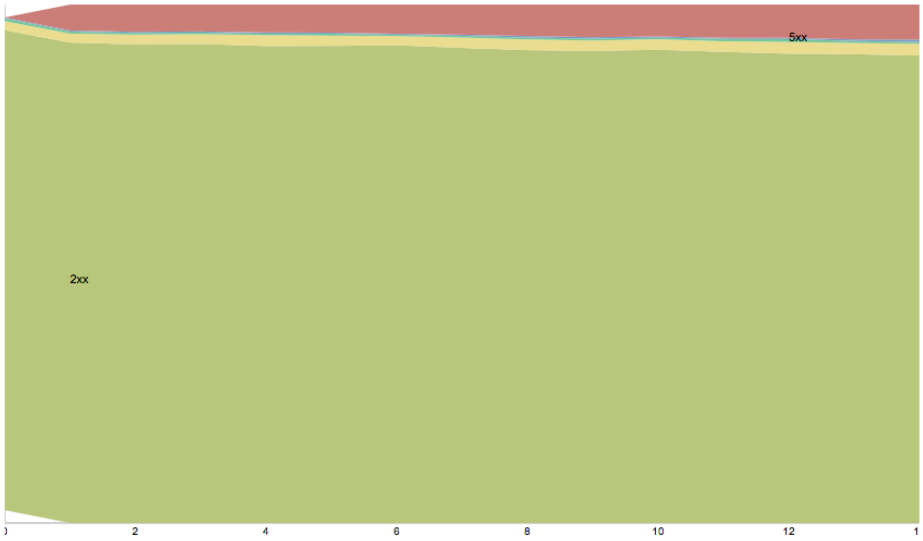
- Stream diagrams of statuses by interval reveal trends with respect to the occurrence of statuses
  - All statuses of a given type are grouped together to form a band for a given month
  - The aggregate represents a complete band of reported statuses for the retrieval period
  - Subsequent bands flow left to right to show change over time
  - Again, status changes indicate the stability or volatility of a set of web-at-large links over time

arXiv, all statuses

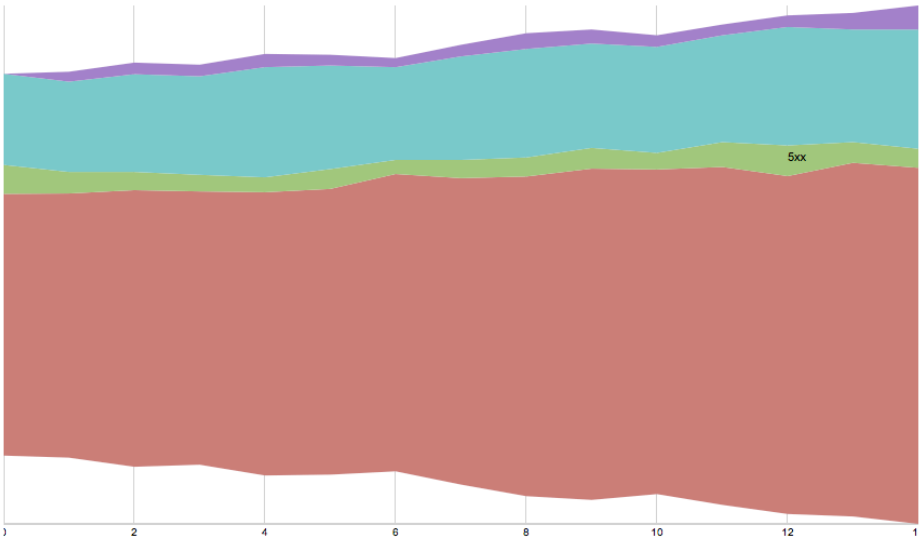
arXiv, “bad” statuses



PLOS, all statuses

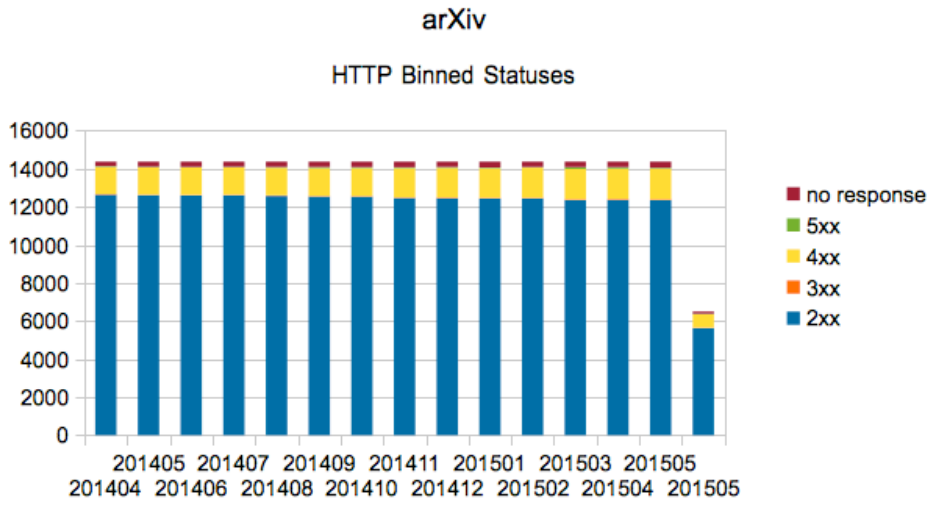


PLOS, “bad” statuses

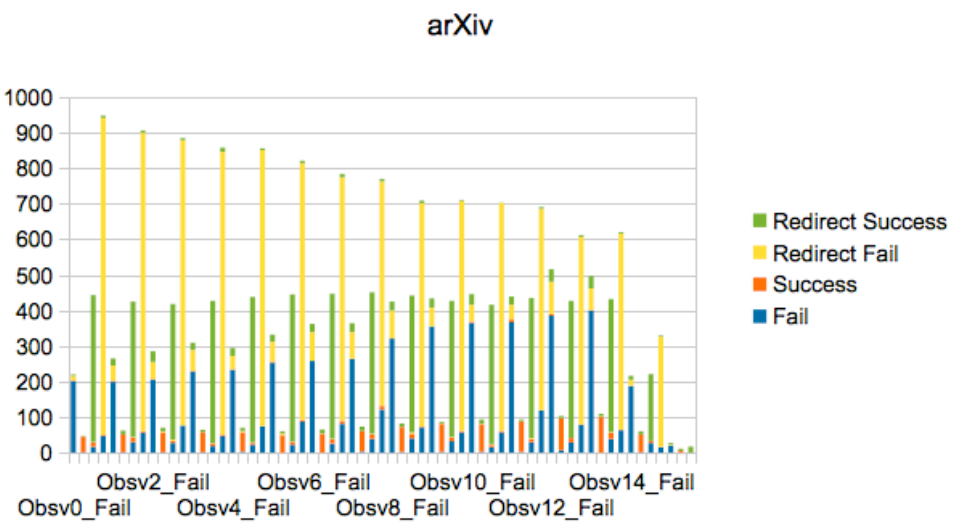


Still more about HTTP statuses

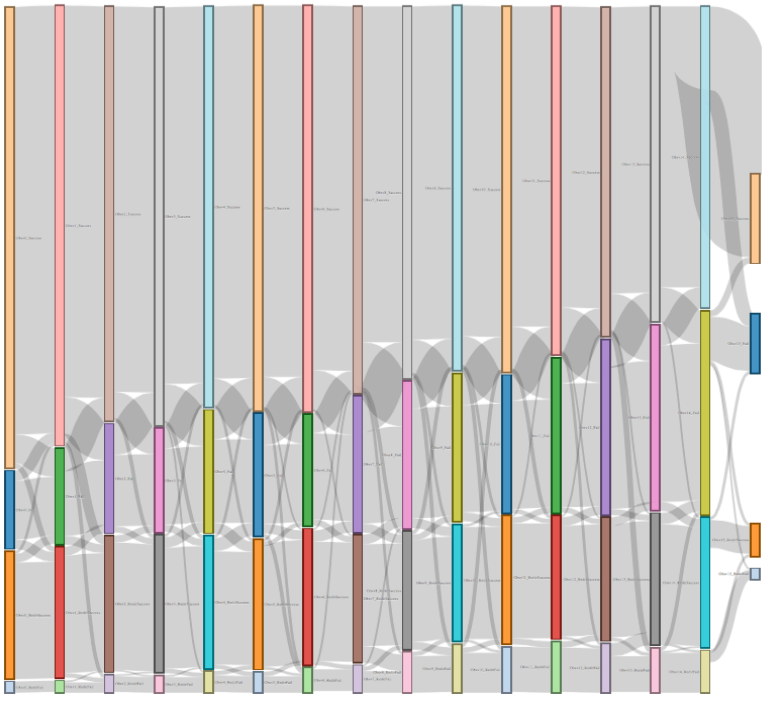
- Sometimes a success or failed attempt to retrieve an object occurs only after one or more redirects
- It may be useful to
  - Examine status chain length and content
  - Longer status chains may hint that link rot will occur in the future even if the object was successfully retrieved this interval
  - Look at the first and last status in the chain to characterize whether this was a success or failure, indicative of link rot
  - Bar charts and sankey diagrams can help understand these possible trends



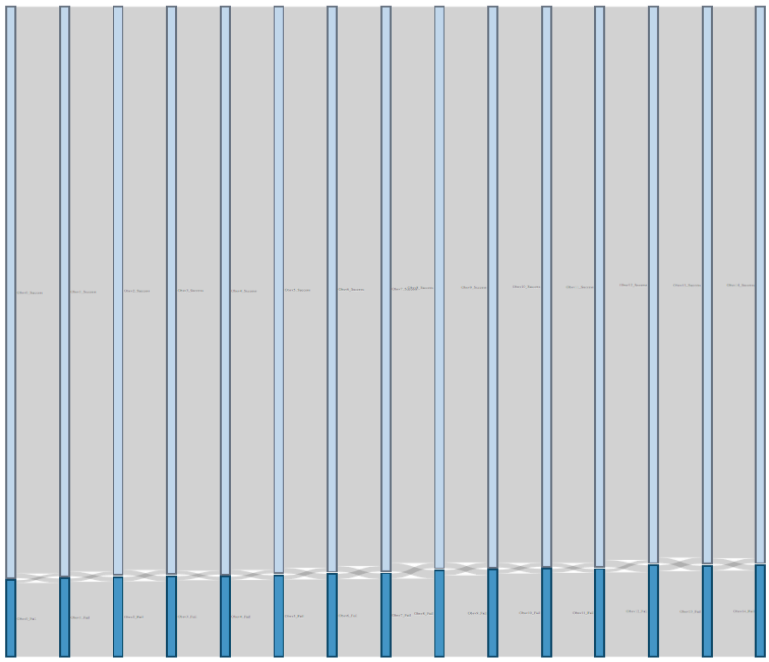
Fail, Redirect-Fail, Success, Redirect-Success stacked by intervals



Successes, failures, with redirects



Only success/fail statuses for arXiv



Refs that change status at some interval

0	http://endo.endojournals.org/cgi/content/abstract/132/6/2614		200	301 302 302 200			text/html	45811	org
1	http://endo.endojournals.org/cgi/content/abstract/132/6/2614	200	200	301 302 302 200	90.3796658767	77	text/html	46846	org
2	http://endo.endojournals.org/cgi/content/abstract/132/6/2614	200	200	301 302 302 200	90.2841777831	71	text/html	46927	org
3	http://endo.endojournals.org/cgi/content/abstract/132/6/2614	200	200	301 302 302 200	90.1297140115	71	text/html	47024	org
4	http://endo.endojournals.org/cgi/content/abstract/132/6/2614	200	503	302 503	2.8024073079	incomparable	text/html	1397	org
5	http://endo.endojournals.org/cgi/content/abstract/132/6/2614	503	200	301 302 302 200	89.790107923	68	text/html	47395	org
6	http://endo.endojournals.org/cgi/content/abstract/132/6/2614	200	200	301 302 302 200	81.1282894744	66	text/html	71890	org
7	http://endo.endojournals.org/cgi/content/abstract/132/6/2614	200	200	301 302 302 200	80.8311160244	66	text/html	71636	org
8	http://endo.endojournals.org/cgi/content/abstract/132/6/2614	200	200	301 302 302 200	80.5321845623	63	text/html	73495	org
9	http://endo.endojournals.org/cgi/content/abstract/132/6/2614	200	200	301 302 302 200	80.5321845623	63	text/html	73952	org
10	http://endo.endojournals.org/cgi/content/abstract/132/6/2614	200	200	301 302 302 200	80.4100148788	63	text/html	75110	org
11	http://endo.endojournals.org/cgi/content/abstract/132/6/2614	200	200	301 302 302 200	37.1635988601	61	text/html	72302	org
12	http://endo.endojournals.org/cgi/content/abstract/132/6/2614	200	none		nan	failed		0	org
13	http://endo.endojournals.org/cgi/content/abstract/132/6/2614		200	301 302 302 200	37.7072221217	65	text/html	71052	org
14	http://endo.endojournals.org/cgi/content/abstract/132/6/2614	200	200	301 302 302 200	78.9085847846	65	text/html	76417	org

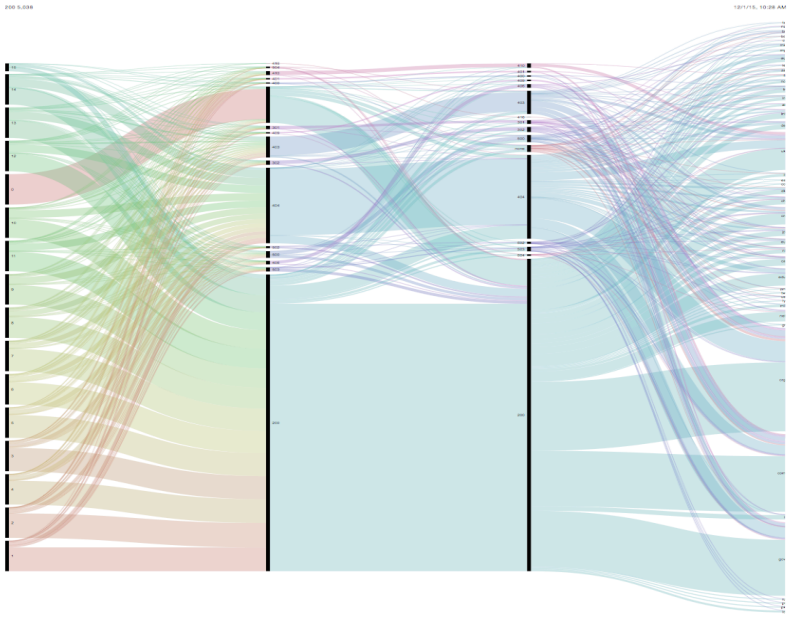
0	http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=17030180&retmode=ref&cmd=prlinks		200	302 301 301 301 301 200			text/html	41010	gov
1	http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=17030180&retmode=ref&cmd=prlinks	200	200	302 301 301 301 301 200	95.6512004697	96	text/html	41285	gov
2	http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=17030180&retmode=ref&cmd=prlinks	200	200	302 301 301 301 301 200	95.4551942199	96	text/html	41216	gov
3	http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=17030180&retmode=ref&cmd=prlinks	200	200	302 301 301 301 301 200	95.1081448961	96	text/html	41222	gov
4	http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=17030180&retmode=ref&cmd=prlinks	200	200	302 200	5.8739140255	incomparable	text/html	916	gov
5	http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=17030180&retmode=ref&cmd=prlinks	200	200	302 301 301 301 301 200	80.8437653193	80	text/html	40605	gov
6	http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=17030180&retmode=ref&cmd=prlinks	200	200	302 301 301 301 301 200	80.9121366538	80	text/html	40605	gov
7	http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=17030180&retmode=ref&cmd=prlinks	200	200	302 301 301 301 301 200	77.5871293726	69	text/html	50567	gov
8	http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=17030180&retmode=ref&cmd=prlinks	200	200	302 301 301 301 301 200	77.5815311343	69	text/html	50568	gov
9	http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=17030180&retmode=ref&cmd=prlinks	200	200	302 301 301 301 301 200	77.5523402744	66	text/html	51760	gov
10	http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=17030180&retmode=ref&cmd=prlinks	200	200	302 301 301 301 301 200	76.2263407524	61	text/html	50662	gov
11	http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=17030180&retmode=ref&cmd=prlinks	200	200	302 301 301 301 301 200	72.160162617	58	text/html	54598	gov
12	http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=17030180&retmode=ref&cmd=prlinks	200	200	302 301 301 301 301 200	72.1198027825	54	text/html	54172	gov
13	http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=17030180&retmode=ref&cmd=prlinks	200	404	302 301 404	7.5940354026	incomparable	text/html	6469	gov
14	http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=17030180&retmode=ref&cmd=prlinks	404	200	302 301 301 301 301 200	71.8744148764	0	text/html	62191	gov

Refs that change status at some interval

0	http://dbpartners.stanford.edu/RegaSubtyping/		200	200			text/html	7826 edu
1	http://dbpartners.stanford.edu/RegaSubtyping/	200	200	200	100	100	text/html	7826 edu
2	http://dbpartners.stanford.edu/RegaSubtyping/	200	200	200	100	100	text/html	7826 edu
3	http://dbpartners.stanford.edu/RegaSubtyping/	200	200	200	100	100	text/html	7826 edu
4	http://dbpartners.stanford.edu/RegaSubtyping/	200	403	403	0	incomparable	text/html	315 edu
5	http://dbpartners.stanford.edu/RegaSubtyping/	403	403	403	0	incomparable	text/html	315 edu
6	http://dbpartners.stanford.edu/RegaSubtyping/	403	403	403	0	incomparable	text/html	315 edu
7	http://dbpartners.stanford.edu/RegaSubtyping/	403	403	403	0	incomparable	text/html	315 edu
8	http://dbpartners.stanford.edu/RegaSubtyping/	403	200	200	7.6037510794	incomparable	text/html	1690 edu
9	http://dbpartners.stanford.edu/RegaSubtyping/	200	200	200	7.6037510794	incomparable	text/html	1690 edu
10	http://dbpartners.stanford.edu/RegaSubtyping/	200	200	200	7.6037510794	incomparable	text/html	1690 edu
11	http://dbpartners.stanford.edu/RegaSubtyping/	200	200	200	7.6037510794	incomparable	text/html	1690 edu
12	http://dbpartners.stanford.edu/RegaSubtyping/	200	200	200	6.1813916982	incomparable	text/html	1309 edu
13	http://dbpartners.stanford.edu/RegaSubtyping/	200	200	200	6.1813916982	incomparable	text/html	1309 edu
14	http://dbpartners.stanford.edu/RegaSubtyping/	200	200	200	6.1813916982	incomparable	text/html	1309 edu

0	http://scott.sherrillmix.com/R-GMT_HexPlot.php		200	200			text/html	3774 com
1	http://scott.sherrillmix.com/R-GMT_HexPlot.php	200	406	406	8.3440653028	incomparable	text/html	387 com
2	http://scott.sherrillmix.com/R-GMT_HexPlot.php	406	406	406	8.3440653028	incomparable	text/html	387 com
3	http://scott.sherrillmix.com/R-GMT_HexPlot.php	406	200	200	99.9999986584	100	text/html	3774 com
4	http://scott.sherrillmix.com/R-GMT_HexPlot.php	200	200	200	99.9999986584	100	text/html	3774 com
5	http://scott.sherrillmix.com/R-GMT_HexPlot.php	200	200	200	99.9999986584	100	text/html	3774 com
6	http://scott.sherrillmix.com/R-GMT_HexPlot.php	200	200	200	99.9999986584	100	text/html	3774 com
7	http://scott.sherrillmix.com/R-GMT_HexPlot.php	200	200	200	99.9999986584	100	text/html	3774 com
8	http://scott.sherrillmix.com/R-GMT_HexPlot.php	200	200	200	99.9999986584	100	text/html	3774 com
9	http://scott.sherrillmix.com/R-GMT_HexPlot.php	200	200	200	99.9999986584	100	text/html	3774 com
10	http://scott.sherrillmix.com/R-GMT_HexPlot.php	200	200	200	99.9999986584	100	text/html	3774 com
11	http://scott.sherrillmix.com/R-GMT_HexPlot.php	200	200	200	99.9999986584	100	text/html	3774 com
12	http://scott.sherrillmix.com/R-GMT_HexPlot.php	200	200	200	99.9999986584	100	text/html	3774 com
13	http://scott.sherrillmix.com/R-GMT_HexPlot.php	200	200	200	99.9999986584	100	text/html	3774 com
14	http://scott.sherrillmix.com/R-GMT_HexPlot.php	200	200	200	99.9999986584	100	text/html	3774 com
15	http://scott.sherrillmix.com/R-GMT_HexPlot.php	200	200	200	99.9999986584	100	text/html	3774 com

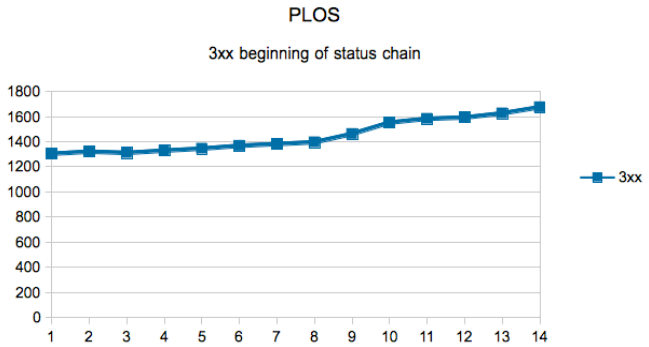
Intervals, previous status, current status, and top level domain (PLOS)

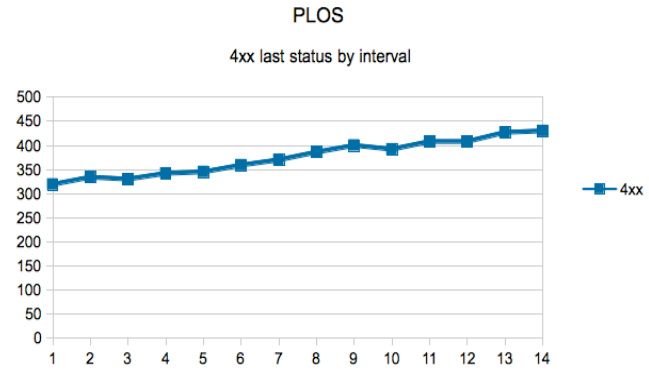


HTTP statuses and link rot

- Is it the case that for this sample that link rot increases with the passage of time?
  - During the course of this experiment, the plos sample experienced a 34% increase in 4xx final statuses and a 28% increase in status chains that began with 3xx (indicating redirection).
  - arXiv exhibited a 12% increase in 4xx final statuses, and an 11% increase in status chains that began with 3xx.
- So the answer would appear to be **yes**.

PLOS

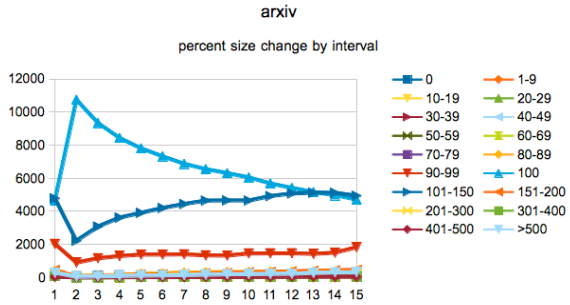




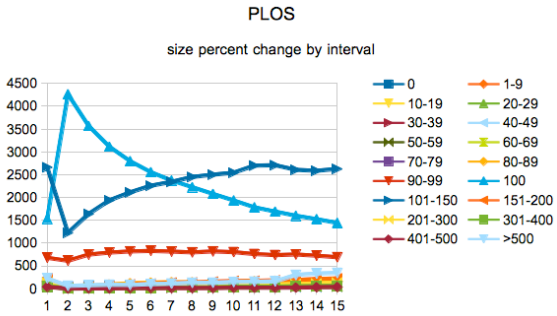
Size (content length) changes

- Size changes hint that content drift may have occurred
- Do document sizes vary over time?
- If so, by how much?
- To what extent do they vary?
- What might this mean?

A bit confusing... 100% actually indicates that a copy is exactly the same size as the original

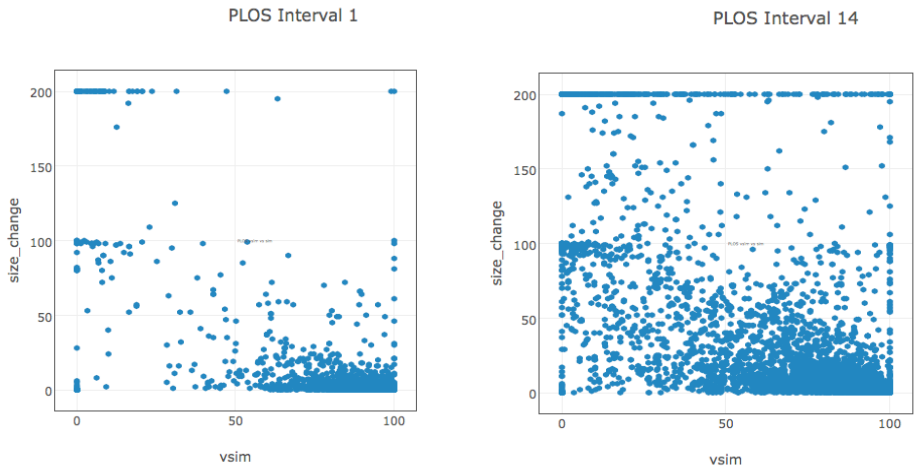


arXiv	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Percent size change in relation to the original document, by interval															
0	279	72	112	131	124	138	150	175	186	189	222	205	240	235	275
1-9	253	52	62	78	103	112	137	158	175	198	195	206	224	253	238
10-19	145	19	28	37	44	65	71	89	92	94	98	130	131	139	149
20-29	114	9	15	27	35	45	47	57	98	114	110	113	117	119	120
30-39	78	7	15	24	38	43	42	45	54	58	60	67	70	77	74
40-49	107	7	11	16	20	29	34	44	90	88	94	92	94	98	106
50-59	110	16	50	57	60	63	72	72	83	95	95	104	103	99	103
60-69	148	25	42	56	71	68	87	83	104	103	121	121	124	129	144
70-79	164	28	65	97	104	124	135	145	149	145	145	149	155	170	159
80-89	387	78	124	143	222	252	281	310	327	332	339	333	341	342	379
90-99	2025	894	1159	1292	1383	1383	1396	1335	1319	1448	1459	1460	1424	1495	1824
100	4656	10744	9333	8434	7804	7320	6867	6547	6307	6045	5679	5411	5162	4971	4712
101-150	4782	2213	3074	3616	3906	4198	4437	4644	4656	4649	4917	5072	5121	5118	4937
151-200	447	62	90	129	172	191	214	241	268	318	334	372	399	431	455
201-300	257	71	83	97	114	138	169	166	172	191	189	213	268	267	263
301-400	76	5	10	17	24	38	49	48	55	57	55	50	68	64	70
401-500	41	12	15	19	23	20	27	34	30	34	35	37	39	40	40
>500	287	42	69	88	110	130	142	164	192	199	210	222	277	310	311



PLOS	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Percent size change in relation to the original document, by interval															
0	139	38	60	63	78	66	63	84	98	90	87	87	116	110	129
1-9	161	26	44	60	69	72	89	100	113	114	107	142	153	153	159
10-19	36	13	8	10	13	18	21	24	34	46	40	39	37	41	34
20-29	45	10	19	19	23	21	27	27	26	32	40	44	49	52	49
30-39	49	5	12	11	16	17	19	22	18	30	29	34	40	55	56
40-49	48	12	13	15	16	21	30	37	38	32	35	28	31	31	44
50-59	42	2	6	14	18	23	33	28	32	36	34	41	45	45	43
60-69	62	7	15	20	21	26	37	27	40	45	47	72	54	59	58
70-79	80	27	37	41	46	52	55	50	58	70	70	77	76	98	86
80-89	153	28	56	92	118	133	141	140	123	149	163	174	147	153	155
90-99	675	610	746	791	816	828	812	795	816	800	758	735	749	722	689
100	1533	4263	3567	3113	2796	2555	2372	2225	2078	1934	1783	1690	1599	1524	1441
101-150	2661	1223	1638	1936	2111	2257	2342	2449	2500	2542	2699	2707	2605	2585	2625
151-200	231	25	42	55	93	110	132	136	154	167	172	182	190	212	224
201-300	136	9	18	24	26	37	48	57	71	78	86	92	100	112	134
301-400	59	5	8	14	15	21	24	29	26	39	46	45	53	56	62
401-500	36	2	5	5	10	11	12	10	13	18	19	20	23	30	35
>500	221	62	73	84	82	99	110	127	129	145	152	158	300	329	344

Another view of vsim vs size change



Correlation coefficient was also nominal -.2 to -.35 average, vs >.8 for sim vs vsim
Linear feature at 200 is artificial, due to a threshold for size change values

Content size discussion

Size changes mean something...

- Quantitatively, we want to be able to say that a document changed by some meaningful fraction (half as large as it originally was, doubled in size, etc)
- We might also investigate whether size change correlates with a change in similarity (meaning) as compared to the original document

Size changes and content drift

- Is it the case that for this sample that content drift increases with the passage of time?
  - Based only on size change, it appears that it does
  - The most common change was for content size to increase, with content growth most commonly falling into the 1 – 50% increase range
  - Also the number of documents that were exactly the same size as the original plummeted over time
    - PLOS by 66%
    - arXiv by 56%

Content Similarity

- One approach: Binary hash of content, compare subsequent interval binary hash to original harvested web-at-large resource using python library ssdeep
- Byte level comparison
- Works for any mime type
- Sometimes small changes to a file can cause hashes to appear more different than they really are – especially true for textual content

Feature vectors

- Between 97 and 99% of the objects harvested and monitored for this experiment were of mime type text/html
- Textual content can be represented as feature vectors and compared using various measures (e.g. Cosine, Euclidean distance)
- Feature vectors are content based so they are less vulnerable to changes in content formatting or minor rearrangement of content

Feature Vector construction

- First, all HTML, CSS and Javascript content is stripped from document using the python library *BeautifulSoup* with the lxml parser
- Next a stop word list is consulted, so frequently occurring words are omitted
- Next, Porter stemmer is employed, thus reducing some English words to a common root
- Next, a vector is created where each word represents a dimension and the frequency of the word is a point on that axis

Comparing feature vectors

- Comparison is always between the original content found at the beginning of the experiment for the Web at large reference, and a subsequent interval's instance of that content
- The representation for each of these two documents is normalized (same number of slots in the matrix even if one document has count zero for a given term), so they have the same dimensions
- We then use cosine similarity to compare the two documents – which looks at the angle between the two vectors rather than the distance between n-dimensional endpoints
- This normalizes all terms, thus frequency of term occurrence is irrelevant (could argue less precise)

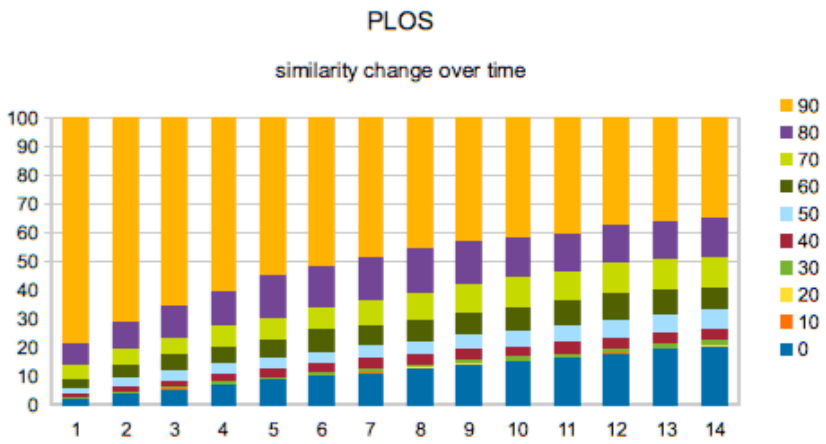
Content similarity issues

- Reformatting and rearranging content can affect the similarity hash but not necessarily the information conveyed in the document
- Rearranging the same content might effectively change the relative information content as compared to a previous copy, or to the original
- Rearranging the content or reformatting it may not change what the referenced document is about

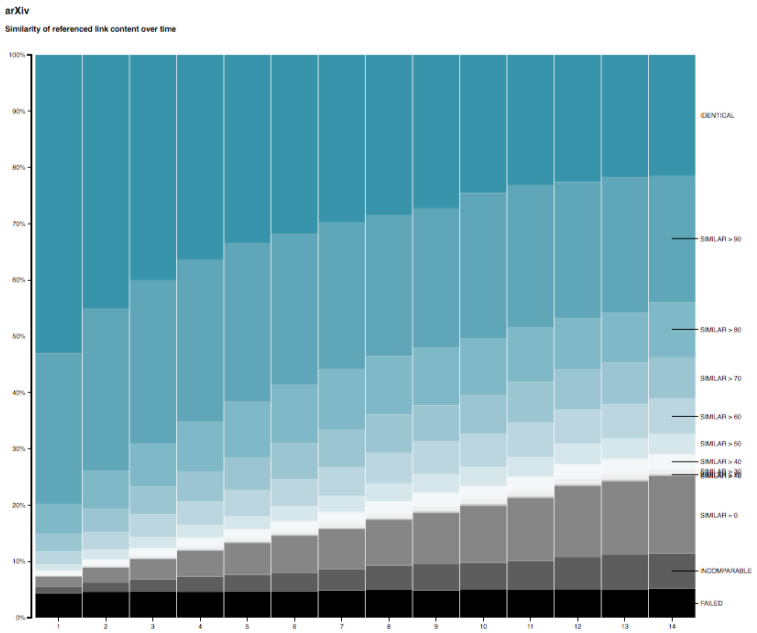
Stacked similarity plots

- Stacked bar charts that compare month over month similarity hash scores for items
- These charts show that content drift is clearly increasing as time passes

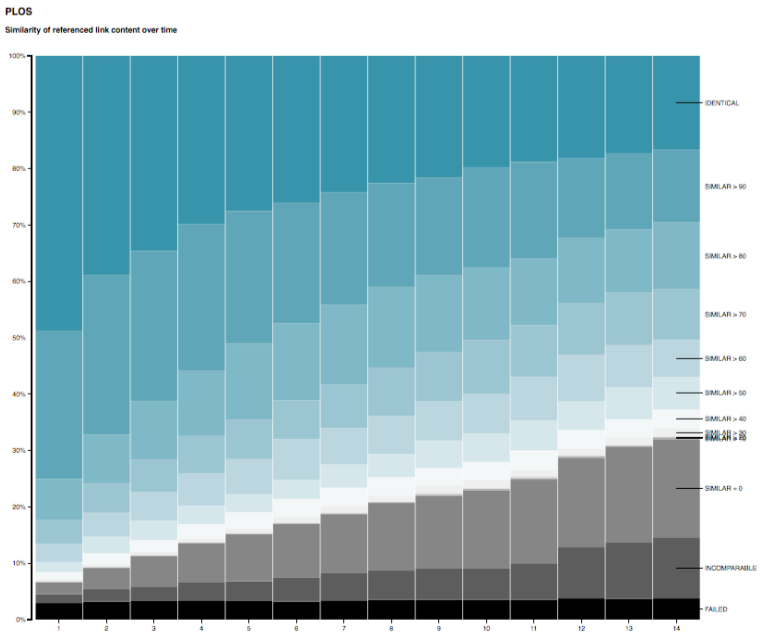
PLOS stacked bar chart



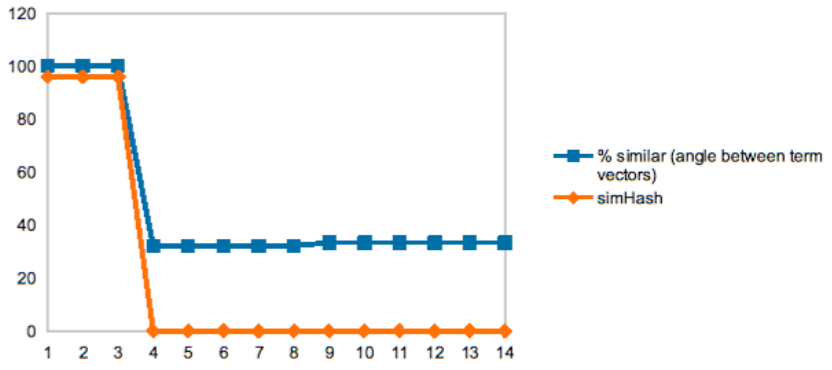
arXiv



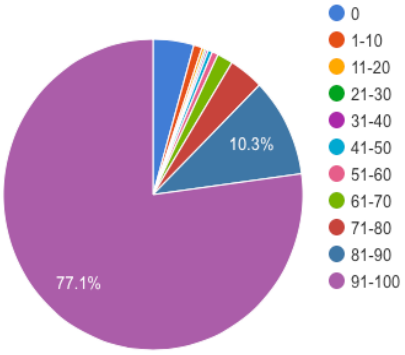
PLOS



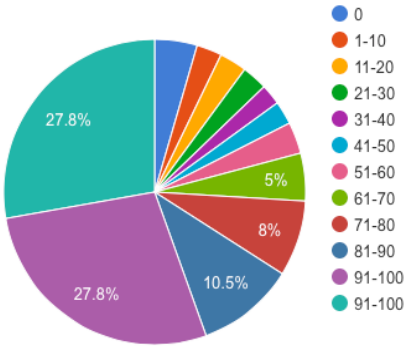
Similarity hash is far less forgiving, which can be deceiving with text



VSIM PLOS first vs last interval

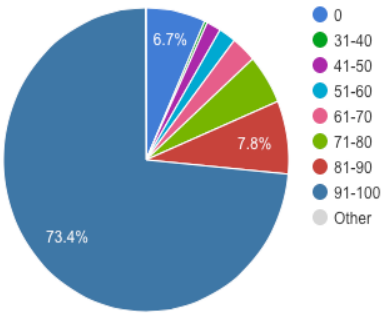


First

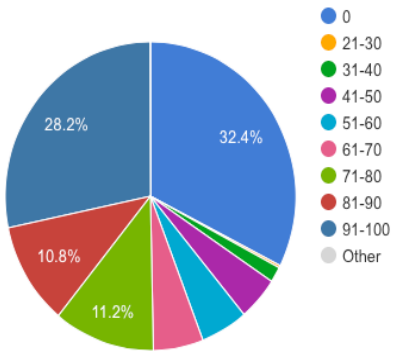


Last

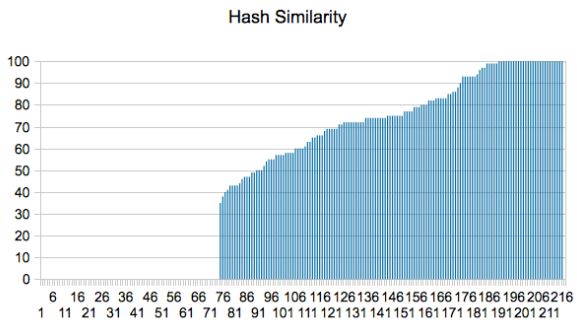
Hash similarity PLOS first vs last interval

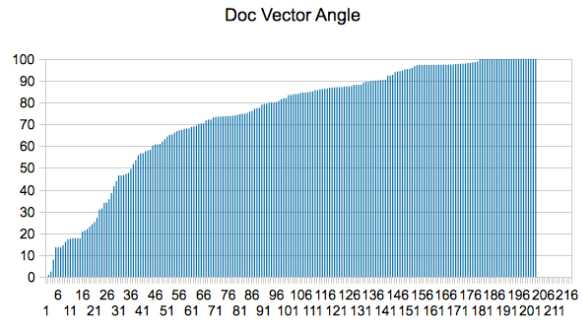


First



Last





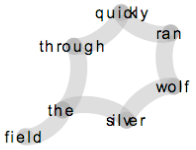
Comparing similarity measures

- Hashes are sensitive to content rearrangement
- Feature vectors are not
- Hashes are fairly unforgiving with respect to content change, large or small
- Feature vectors can illustrate more subtle content drift but won't notice content rearrangement
- Using graph document models and graph matching to compare content would reveal more subtle changes in textual documents, but this would be computationally intensive

Graph Model, Graph Edit Distance

- Really small GXL document graph model example

```
<!DOCTYPE gxl SYSTEM "http://www.gupro.de/GXL/gxl-1.0.dtd">
<gxl xmlns:xlink="http://www.w3.org/1999/xlink">
<graph id="SEN_004" edgeids="false" edgemode="directed">
<node id="_0"><attr name="word"><string>the</string></attr></node>
<node id="_1"><attr name="word"><string>silver</string></attr></node>
<node id="_2"><attr name="word"><string>wolf</string></attr></node>
<node id="_3"><attr name="word"><string>ran</string></attr></node>
<node id="_4"><attr name="word"><string>quickly</string></attr></node>
<node id="_5"><attr name="word"><string>through</string></attr></node>
<node id="_6"><attr name="word"><string>field</string></attr></node>
<edge from="_0" to="_1"/>
<edge from="_1" to="_2"/>
<edge from="_2" to="_3"/>
<edge from="_3" to="_4"/>
<edge from="_4" to="_5"/>
<edge from="_5" to="_0"/>
<edge from="_0" to="_6"/>
</graph></gxl>
```



- Output of a Graph Edit Distance run

```
Graph edit distance procedure:  Beam

Edge mode:              directed
Cost for node deletion/insertion:      3.0
Cost for edge deletion/insertion:      3.0

Alpha weighting factor between node and edge costs:     0.8

Node attribute 0:       word;   Cost function:  discrete;       mu = 0
nu = 2;  Soft factor:    1.0

No attributes for edges defined

Individual node costs are added
Individual edge costs are added
(Combined node cost)^(1/1.0)
(Combined edge cost)^(1/1.0)

*** The distance matrix ***

1.20000,3.40000,2.80000,3.40000
3.40000,1.20000,5.00000,1.20000
2.80000,5.00000,1.20000,5.00000
3.40000,1.20000,5.00000,1.20000
```

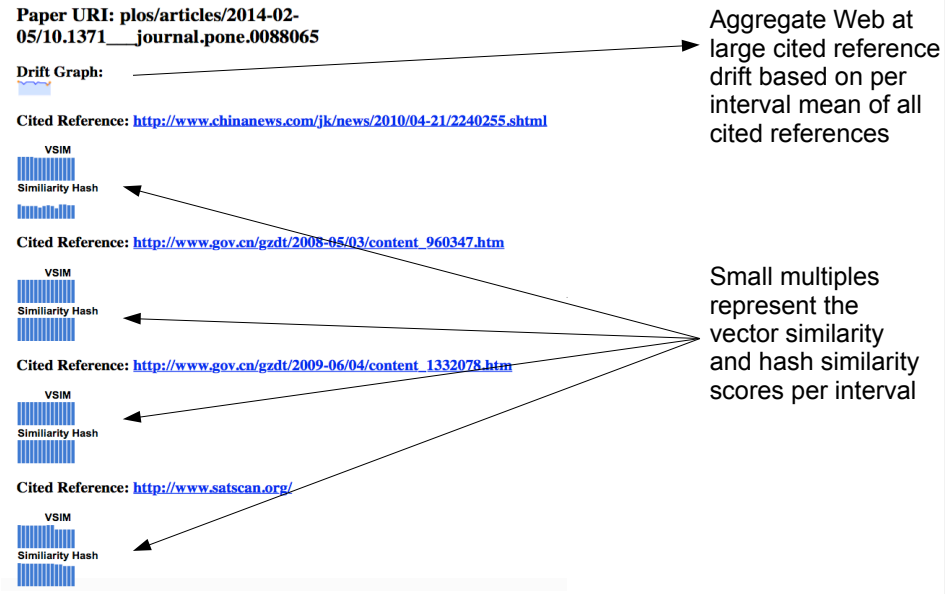
Use entropy to compare original to harvested copies

- Entropy as defined in information theory: the amount of surprise, the probability of the occurrence of a given bit, byte, character, token, etc.
- Use the original document as a model
- Use English string tokenizer, as with feature vectors, to create a granular pair of data streams to compare
- Low entropy, little change, high entropy, significant content changes

Another way to look at content drift

- The web-at-large links associated with a given paper represent a portion of its context
- So, what is happening to the aggregate context for a given paper?
- Can we visualize the aggregate content drift?
- Can we calculate some kind of meaningful aggregate drift score for a paper?

Citing paper exploring aggregate drift



A paper that saw aggregate drift

Paper URI: plos/articles/2014-02-04/10.1371___journal.pone.0087534

Drift Graph:

Cited Reference: http://crocodilian.com/cnhc/cnhc.html

VSIM

Similiarity Hash

Cited Reference: http://www.lirmm.fr/~caraux/Bioinformatics/NegativeMultinomial/

VSIM

Similiarity Hash

Cited Reference: http://www.ncbi.nlm.nih.gov/

VSIM

Similiarity Hash

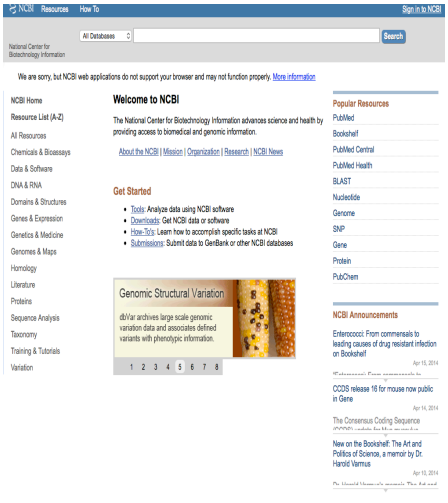
Cited Reference: http://www.r-project.org

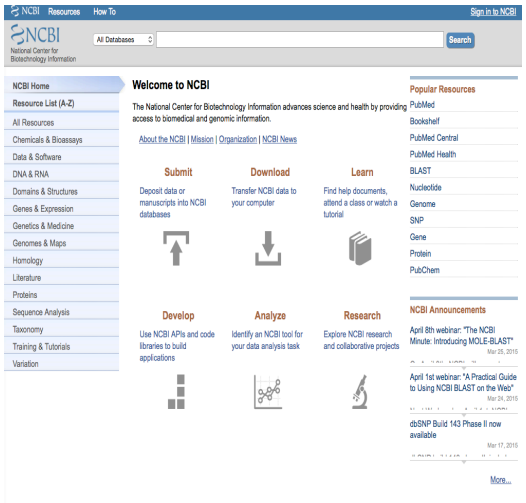
VSIM

Similiarity Hash

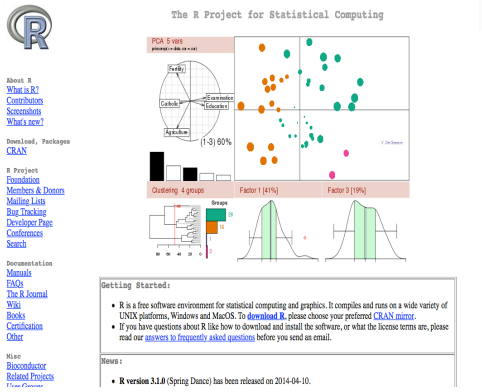
Drift Graph:

NCBI Interval 1 and Interval 15

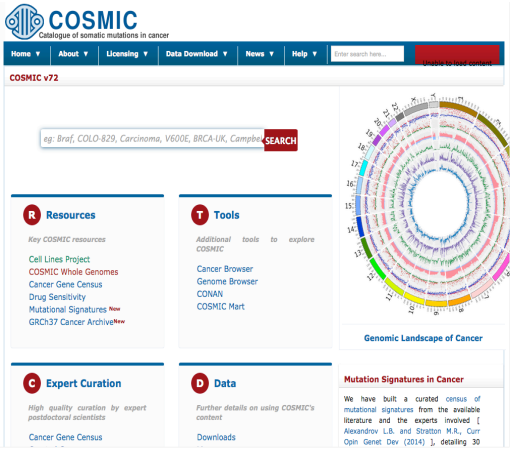




R-Project Interval 1 and 15



COSMIC Interval 1 and 14



Another example

Paper URI: plos/articles/2014-02-05/10.1371___journal.pone.0088388

Drift Graph:

Cited Reference: http://www.ars.usda.gov/ba/fsrg

VSIM

Similiarity Hash

Cited Reference: http://www.bhf.org.uk

VSIM

Similiarity Hash

Cited Reference: http://www.cdc.gov/nchs/nhanes.htm

VSIM

Similiarity Hash

Cited Reference: http://www.cdc.gov/nchs/nhanes/nhanes2003-2004/analytical_guidelines.htm

VSIM

Similiarity Hash

Drifting reference

URI http://www.cdc.gov/nchs/nhanes.htm

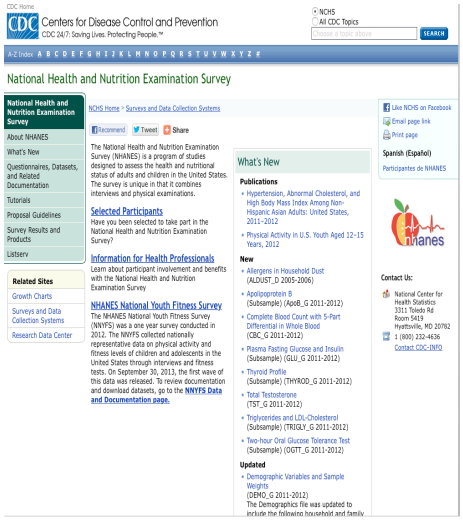
Interval	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
VSIM		80.5495328035	78.688475321	48.1812683986	48.4617977811	46.6842573389	47.7946530164	47.1557947414	45.721880693	47.996968254	46.3997696048	46.7327838301	47.1335448909	46.6950464289	47.1437659677
Similarity Hash		82	79	54	54	54	50	44	44	54	50	50	50	60	57

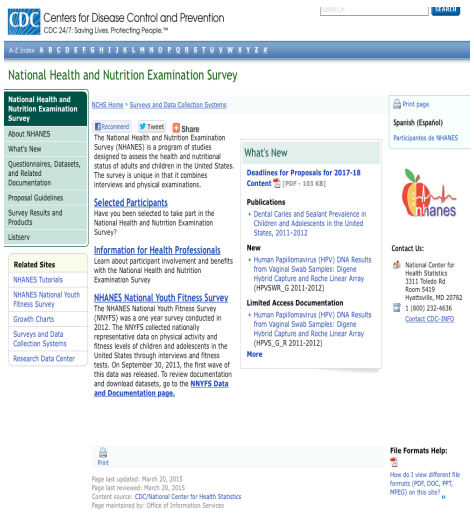
Visualizations of...

VSIM

Similiarity Hash

But not necessarily





An arXiv example

Paper URI: arxiv/articles/2014-02-10/1402.1693.pdf

Drift Graph:

Cited Reference: http://en.wikipedia.org/wiki/Radio-frequency_

VSIM

Similiarity Hash

Cited Reference: http://www.ijcttjournal.org

VSIM

Similiarity Hash

Cited Reference: http://www.rfidreader.info

VSIM

Similiarity Hash

Cited Reference: http://www.webopedia.com/TERM/R/RFID.html

VSIM

Similiarity Hash

Drift over intervals

URI http://www.webopedia.com/TERM/R/RFID.html

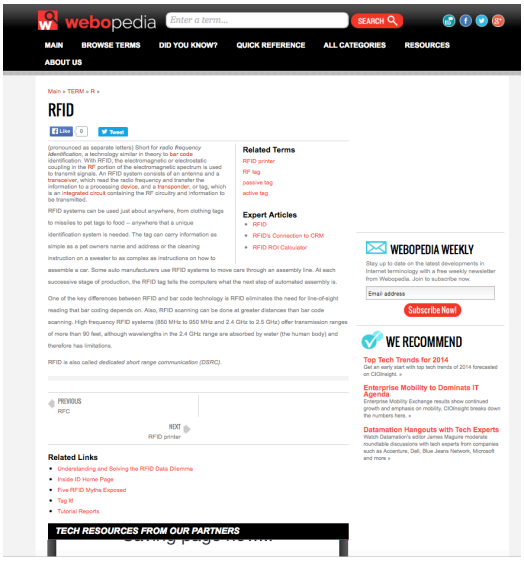
Interval	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
VSIM		94.6808587881	81.4394182743	80.9194831356	80.9752024027	74.3773192546	72.2005775786	72.7674750684	72.6705682691	66.3972676361	67.9518357817	67.5069360714	66.6890930887	66.3690611449	48.5046277953	48.4777986223
Similarity Hash		80	0	0	0	0	0	0	0	0	0	0	0	0	0	0

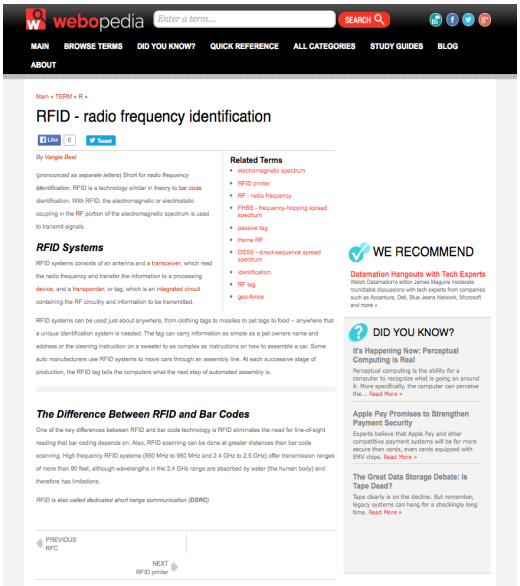
Visualizations of...

VSIM

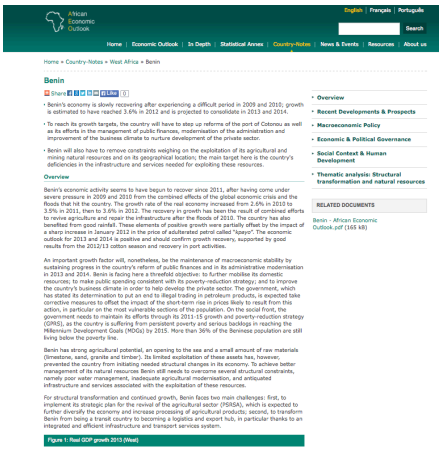
Similiarity Hash

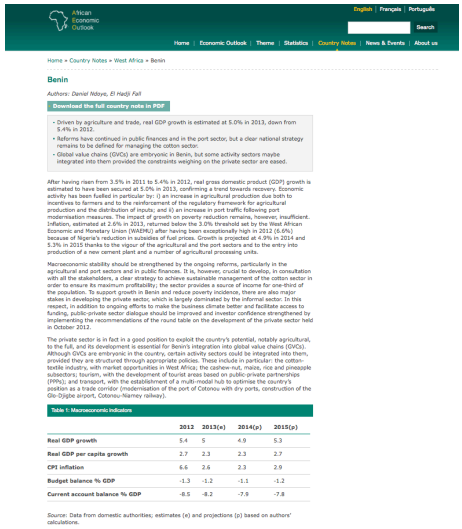

A definition of RFID rewritten, expanded





On the other hand...





What a difference a year makes – completely rewritten economic summary of Benin with additional year's data

Drift for collaborative knowledge archives

- DBPedia

Average vsim score for arXiv Dbpedia references	83.1609944762
Median	83.3099367787
Standard Deviation	14.5843111047
Minimum	0
Count	114

Average vsim score for PLOS Dbpedia references	90.742775143
Median	91.5110663275
Standard Deviation	14.4120110005
Minimum	15.2128779096
Count	45

- Wikipedia

Average vsim score for arXiv Wikipedia references	72.454779475
Median	81.1939942706
Standard Deviation	15.9157933471
Minimum	10.2760503055
Count	4103

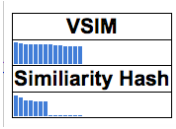
Average vsim score for PLOS Wikipedia references	82.3733833883
Median	84.7083559866
Standard Deviation	12.005865573
Minimum	48.1399118472
Count	203

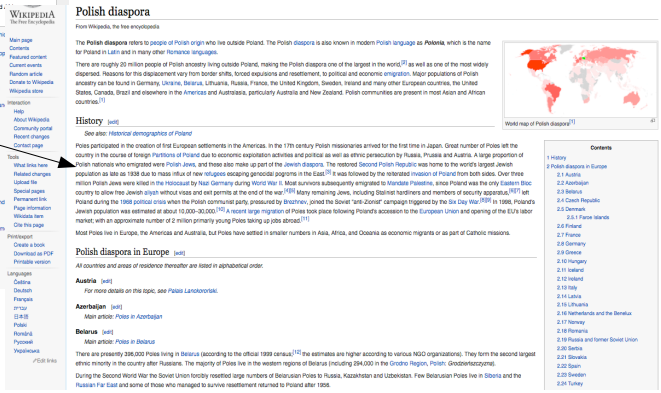
Paper URI: plos/articles/2014-02-26/10.1371___journal.pone.0089094

URI http://www.wikipedia.org/wiki/Polish_diaspora

Interval	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
VSIM		95.6340673805	95.2581050605	90.8034715828	89.0253273915	89.4879046515	89.1197618124	88.5628927583	88.0126419665	81.8988033806	81.9102912652	81.7504468501	80.5717285329	81.27371033	81.3690217007
Similarity Hash		91	83	72	69	65	65	65	0	0	0	0	0	0	0

Interval 1





Interval 14

Observations

- Drift is sometimes the result of very general, or lazy, references, such as references to the root level of a project or organization website
- Although these references make sense in some ways, they are exceptionally vulnerable to content drift
- Some references drift less than the similarity measures indicate – the core of the page in the CDC example changed little over 13 months, it was the “What's New” section that changed.

Summary

- No matter content format or similarity measure used, measuring and characterizing content drift remains difficult and challenging
- In one sense, drift is a subjective event
- Some authors may consider significant changes in peripheral aspects of a cited Web at large reference as immaterial
- Perhaps some authors actually intend to cite the "*timeless*" version of a reference
- Yet activities such as analysing data and reproducing results might require precise alignment of software versions and other details which a snapshot would increase the likelihood of a reader finding

Weka

- Converting forward experiment metadata and content metrics to ARFF attributes
  - Nominals, strings, numeric
- Unsupervised clustering
- Classifier for identifying drift based on attribute values
- Association algorithm and resulting rules
- Potential uses for...

ML algorithms work with “concepts”

- Just converting your existing data and importing it may not yield good results
- What are the characteristics of the thing or phenomena that you are describing?
- Are they adequately represented within the data?
- Categorical data and numeric data is useable by more algorithms, strings are generally not useful unless you are specifically analysing text
- Concepts can be used to classify, cluster, find associations or to predict values

Prepping data for Weka

- Metadata about cited references

@attribute interval {0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16}
@attribute url string
@attribute citedby string
@attribute last_val_laststatus {200, 203, 301, 302, 400, 401, 403, 404, 405, 406, 408, 409, 410, 416, 429, 490, 500, 501, 502, 503, 504}
@attribute current_val_laststatus {200, 203, 301, 302, 400, 401, 403, 404, 405, 406, 408, 409, 410, 416, 429, 490, 500, 501, 502, 503, 504}
@attribute status_chain string
@attribute vsim {0,1-9,10-19,20-29,30-39,40-49,50-59,60-69,70-79,80-89,90-99,100}
@attribute sim {0,1-9,10-19,20-29,30-39,40-49,50-59,60-69,70-79,80-89,90-99,100}
@attribute htype { *, application/json, application/msword, application/octet-stream, ... }
@attribute size numeric
@attribute tld { *nominal value list* }

- Derived attributes about the content

@attribute vsim numeric
@attribute sim numeric

Building training data for classifier my rules for content drift

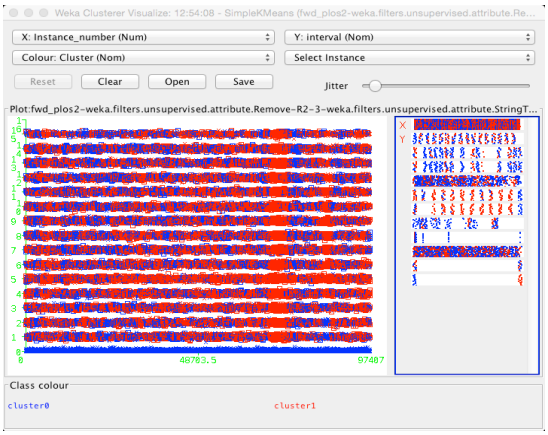
- If it's the zero'th interval and the vector similarity is greater than 80%, then flag as no drift
- If vector similarity is less than 80%, content drift has occurred, drift set to yes
- If size has changed more than 10% then drift has occurred, drift set to yes
- Otherwise flag as no drift

Preprocessing the data

- Many algorithms don't work with strings, used unsupervised text filter to convert status chains (StringToNominal in Weka)
- All Url strings were omitted
- Tests consisting of applying various classes of algorithms against the data were conducted on instance data for the PLOS collection.
- Each instance represents a concept that has various retrieve-time metadata, a previous retrieval HTTP status and some derived values based on the actual retrieved content

Cluster: Simple Kmeans

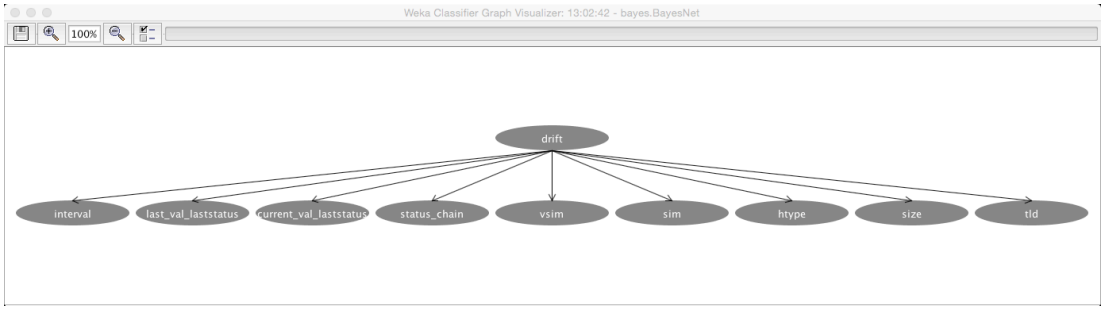
```
Clustered Instances

0       41599 ( 43%)
1       55809 ( 57%)


Class attribute: drift
Classes to Clusters:

     0     1  <-- assigned to cluster
 12817 29006 | yes
 28782 26803 | no


Cluster 0 <-- no
Cluster 1 <-- yes

Incorrectly clustered instances : 39620.0  40.6743 %
```



Classifiers: BayesNet

```
=== Summary ===

Correctly Classified Instances       95087               97.6172 %
Incorrectly Classified Instances      2321                2.3828 %
Kappa statistic                          0.9511
Mean absolute error                      0.0337
Root mean squared error                  0.1336
Relative absolute error                  6.8687 %
Root relative squared error             26.9827 %
Total Number of Instances            97408

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC
Area  Class
                 0.947     0.002      0.998     0.947     0.972      0.997
   yes
                 0.998     0.053      0.961     0.998     0.98       0.997
   no
Weighted Avg.    0.976     0.031      0.977     0.976     0.976      0.997

=== Confusion Matrix ===

     a     b   <-- classified as
 39601  2222 |     a = yes
    99 55486 |     b = no
```

BayesNet visualized



Probability Distribution Table For current_val_laststatus

drift	200	203	301	302	400	401	403	404	405	406	408	409	410	416	429	490	500	501	502	503	504
yes	0.928	0	0.003	0.003	0.001	0.002	0.009	0.047	0	0.001	0	0	0.002	0	0	0	0.002	0	0.001	0.001	0.001
no	0.941	0	0	0	0.002	0.003	0.005	0.043	0	0.003	0	0	0	0	0	0	0.002	0	0	0.001	0

Probability Distribution Table For vsim

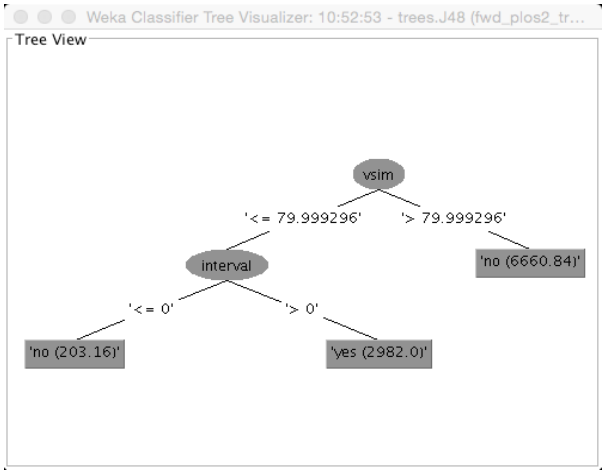
drift	0	1-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	100
yes	0.011	0.049	0.046	0.036	0.034	0.048	0.06	0.107	0.206	0.3	0	0.103
no	0	0	0	0	0	0	0	0	0	0	0.429	0.571

Probability Distribution Table For sim

drift	0	1-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	100
yes	0.216	0	0	0.002	0.02	0.058	0.081	0.107	0.106	0.102	0.08	0.226
no	0.018	0	0	0.001	0.001	0.005	0.007	0.022	0.047	0.117	0.264	0.519

Decision trees

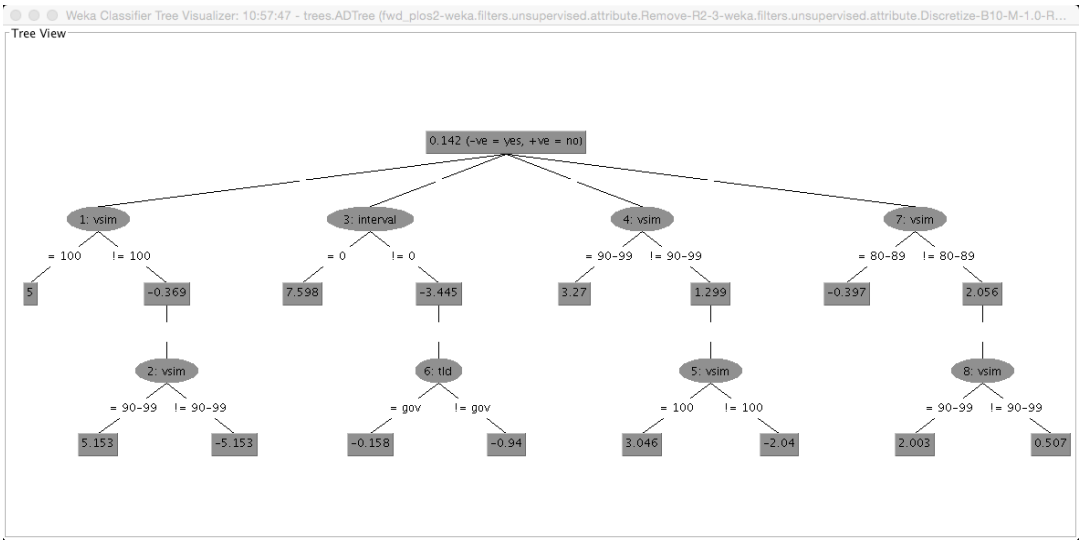
- J48* Decision Tree



* http://www.d.umn.edu/~padhy005/Chapter5.html

Decision trees

- ADTree* (alternating decision)



* https://en.wikipedia.org/wiki/Alternating_decision_tree

Alternate concept model

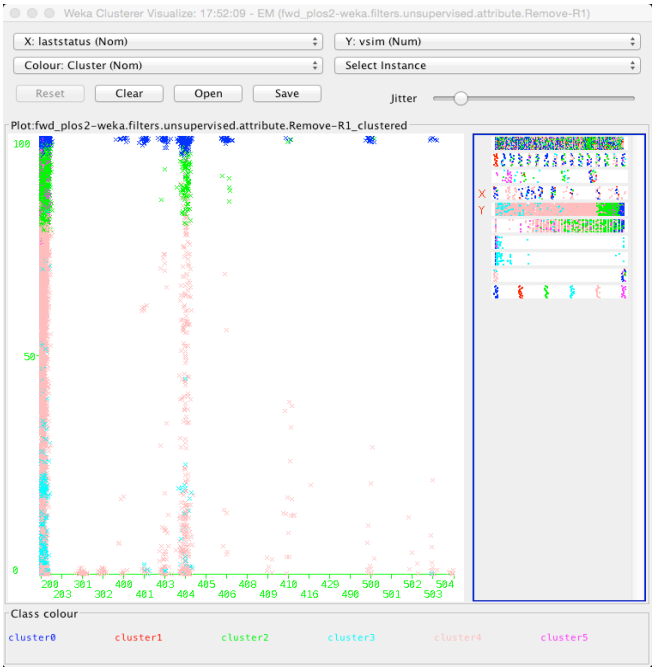
- Metadata about cited references

@attribute url string
@attribute interval {0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15}
@attribute htype { *, application/json, application/msword, application/octet-stream, application/pdf, application/postscript, application/rar, application/rdf+xml, application/rss+xml, application/unknown, application/vnd.google-earth.kmz, application/vnd.ms-excel, application/vnd.ms-powerpoint, application/x-dvi, application/x-gzip, application/x-rar, application/x-rar-compressed, application/x-tar, application/x-troff-man, application/xhtml+xml, application/xml, application/zip, Content-Type:text/html, image/jpeg, image/png, image/svg+xml, text/csv, text/html, text/plain, text/turtle, text/vnd.wap.wml, text/x-csrc, text/x-server-parsed-html, text/xml, video/mp4, video/quicktime, video/x-msvideo }
@attribute laststatus {200, 203, 301, 302, 400, 401, 403, 404, 405, 406, 408, 409, 410, 416, 429, 490, 500, 501, 502, 503, 504}
@attribute size numeric
@attribute drifted {yes, no}

- Derived attributes about the content

@attribute sizechange numeric % current content length as compared to original
@attribute vsim numeric
@attribute sim numeric

Simple EM (expectation maximisation)



Classifiers: Logistic Regression

```
=== Summary ===

Correctly Classified Instances       19259               99.4218 %
Incorrectly Classified Instances       112                0.5782 %
Kappa statistic                          0.9862
Mean absolute error                      0.0116
Root mean squared error                  0.0653
Relative absolute error                  2.78   %
Root relative squared error             14.2883 %
Total Number of Instances            19371

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC
Area  Class
                 0.991     0.004      0.989     0.991     0.99        1
   yes
                 0.996     0.009      0.996     0.996     0.996       1
   no
Weighted Avg.    0.994     0.008      0.994     0.994     0.994       1

=== Confusion Matrix ===

     a     b   <-- classified as
  5699    51 |     a = yes
    61 13560 |     b = no
```

Some classification rules, by algorithm

- RIpple DOwn Rule Learner(Ridor) rules

```
drifted = yes
Except (vsim > 80.00467) => drifted = no
Except (interval = 0) => drifted = no
```

- Jrip

```
(vsim <= 79.997759) => drifted=yes
=> drifted=no
```

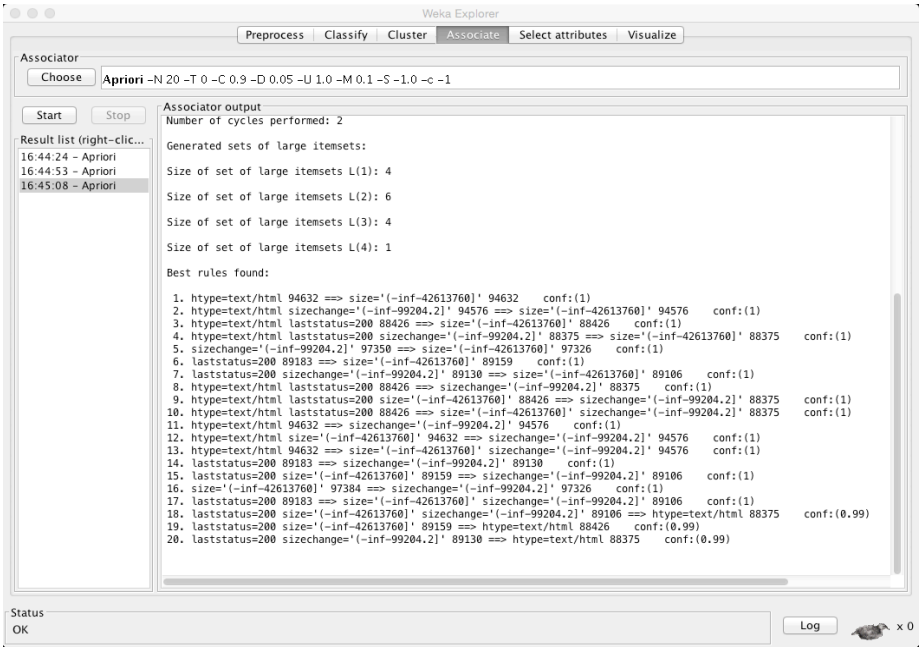
- OneR

```
vsim:
< 80.00218866579999    -> yes
>= 80.00218866579999   -> no
?   -> no
```

Applying association learner

- Algorithm will not work with numeric data
- Discretize step required for these values
  - Vsim
  - Sim
  - Size
  - Sizechange
- Tuning the Discretize algorithm
  - Somewhat arbitrarily chose 100 bins for nominals derived from numeric values – poor results. 1000 bins yields more reasonable nominal values.
  - “Use equal frequency” setting also yields better nominal values

Association in the data: PLOS



Apriori association learner

- High confidence level rules tend to be obvious and/or not useful associations within the data
- Setting “number of rules to find” higher is one way to see more variety and more interesting associations
- Of course, confidence level is lower for some of these rules so you have to keep this in mind when reviewing them

Examples

- Spurious associations
  - 1. sim='(99.5-inf)' 23768 ==> htype=text/html 23768 conf:(1)
  - 4. laststatus=200 sim='(99.5-inf)' 21178 ==> htype=text/html 21178 conf:(1)
  - 30. drifted=no 68142 ==> htype=text/html 67380 conf:(0.99)
  - 38. vsim='(100-inf)' sizechange='(-0.5-0.5]' drifted=no 17066 ==> htype=text/html 16746 conf:(0.98)
- Valid associations
  - 49. sim='(99.5-inf)' sizechange='(-0.5-0.5]' 21514 ==> drifted=no 20402 conf:(0.95)
  - 53. drifted=no 68142 ==> laststatus=200 64395 conf:(0.95)
  - 79. vsim='(100-inf)' 18520 ==> sizechange='(-0.5-0.5]' drifted=no 17066 conf:(0.92)
  - 99. sim='(99.5-inf)' 23768 ==> sizechange='(-0.5-0.5]' 21514 conf:(0.91)

What's a use cases for ML?

- Let's say our goal is to warn a user about the potential for content drift having occurred...
  - Were a classifier to arrive at a combination of size, status and content type/content length as another path to accurate drift classification, then yes but preliminary tests didn't result in anything quite so simple.
  - A couple of rules-based classifiers did arrive at some very simple rules which all emphasized vector similarity
  - Generally, size and similarity measures are the best indicators of content drift. Size is available via HEAD HTTP, but similarity requires calculation prior to the classification step.