

# PSS – Person Search System

Paul Ridgway  
Matthew Rowe

# Outline

- Motivation
- Approach
  - Crawling
  - Indexing
  - Information Extraction
  - Clustering
- Conclusions
- Future Work

# Motivation

- Web users increasingly use the Web to search for information about specific people, however:
  - Person names are ambiguous



Simon Tucker

Search

About 3,520,000 results (0.25 seconds)

[Advanced search](#)

Everything

Images

More

The web

[Pages from the UK](#)

All results

[Sites with images](#)

More search tools

### [Simon Tucker: Multimedia Interface Researcher](#) ☆

S. **Tucker**, A. Ramamoorthy, O. Bergman and S. Whittaker: Catchup: A Useful Application of Time-Travel in Meetings; O. Bergman, S. **Tucker**, R. Beyth-Marom, ...  
[www.dcs.shef.ac.uk/~sat/](http://www.dcs.shef.ac.uk/~sat/) - [Cached](#) - [Similar](#)

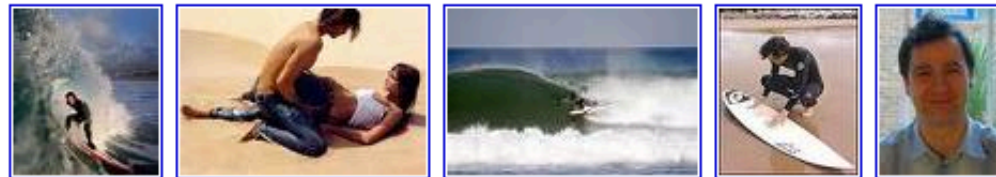
### [Simon Tucker Surf Academy](#) ☆

**Simon Tucker's** Surfing Academy, established in 2002, was opened by Tom Curren, four times World Professional Champion. Whether learning or wanting to ...  
[www.bridgend.gov.uk/web/groups/tourism/.../002130.hcsp](http://www.bridgend.gov.uk/web/groups/tourism/.../002130.hcsp) - [Cached](#) - [Similar](#)

### [Graduate School of the Environment - Simon Tucker](#) ☆

arrow About Us arrow Staff Members arrow **Simon Tucker** ... S. Tucker, (2007) Integrating energy efficiency into building design using a simplified thermal ...  
[www2.cat.org.uk/graduateschool/index.php?...id...](http://www2.cat.org.uk/graduateschool/index.php?...id...) - [Cached](#) - [Similar](#)

### [Images for Simon Tucker](#) - [Report images](#)



### [Simon Tucker resigns as number two at Telegraph Media Group ...](#) ☆

23 May 2008 ... **Simon Tucker**, chief executive Murdoch MacLennan's number two at the Telegraph Media Group, has left the company after 14 months in the job.  
[www.guardian.co.uk/.../telegraphmediagroup.pressandpublishing](http://www.guardian.co.uk/.../telegraphmediagroup.pressandpublishing) - [Cached](#) - [Similar](#)

### [Simon Tucker | Launchpad](#) ☆

**Simon Tucker**. Simon joined the Young Foundation to establish Launchpad in 2005. He oversees all Launchpad programmes and projects as well as currently ...  
[launchpad.youngfoundation.org/about/people/.../simon-tucker](http://launchpad.youngfoundation.org/about/people/.../simon-tucker) - [Cached](#) - [Similar](#)

### [simontucker.net](#) ☆

Printing a directory listing. Just thought I'd share a quick tip that will let you list all files/folders within a directory in a text file. ...  
[www.simontucker.net/](http://www.simontucker.net/) - [Cached](#)



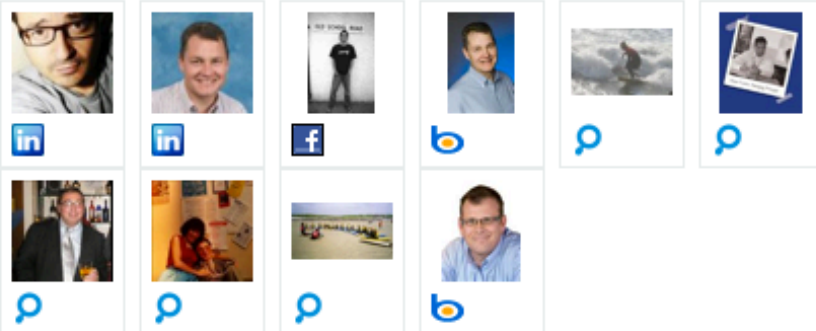
# Motivation

- Web users increasingly use the Web to search for information about specific people, however:
  - Person names are ambiguous
- Current people search engines do not disambiguate
  - Concentrate on specific content sources
  - Limits information found: domain specific patterns
    - Limited in extracting information from new web documents

► People search results for: **Simon Tucker**

## Simon Tucker's Pictures (39) ?

ALL 123people facebook linkedin bing myspace



1 2 3 4 next&gt;&gt;

## Related Domains (4) ?

ALL 1&amp;1 123people

The following domains were found for "Simon Tucker" - 4 results

1&1 [tuckersimon.co.uk](http://tuckersimon.co.uk)  
✓ This domain is still available.

Register this domain!

1&1 [tuckersimon.com](http://tuckersimon.com)  
✓ This domain is still available.

Register this domain!

[Simon Tucker](http://www.simontucker.co.uk)  
<http://www.simontucker.co.uk>

[simon-tucker.com has been registered](#)

## Premium Public Records (50) ?

1 9 2 .com

192 [Simon C Tucker](#)  
SHEFFIELD - [view details](#)

[See more people called Simon Tucker from 192.com](#)

192 [Simon P Tucker](#)  
BARNSELY - [view details](#)

[See more people called Simon Tucker from 192.com](#)

192 [Simon M Tucker](#)  
GLOSSOP - [view details](#)

[See more people called Simon Tucker from 192.com](#)

192 [Simon J C Tucker](#)  
LEICESTER - [view details](#)

[See more people called Simon Tucker from 192.com](#)

192 [Simon L Tucker](#)  
OLDBURY - [view details](#)

[See more people called Simon Tucker from 192.com](#)

1 2 3 4 5 next&gt;&gt;

## News (17) ?

ALL theguardian Telegraph STM

[Simon Tucker resigns as number two at Telegraph Media Group](#)

Simon Tucker, chief executive Murdoch MacLennan's number two at the Telegraph Media Group, has left the company after 14 months in the job. By Mark Sweney

[The Real Business - Telegraph](#)



Name [Email](#) [Username](#) [Phone](#) BETA [Business](#)

simon	tucker			
First Name	Last Name	City	State	Country

[Clear](#)

## Simon Tucker

### Quick Facts

[Simon Tucker](#) is Associate Director at the Young Foundation and director of... [www.socialinnovationexchange.org](http://www.socialinnovationexchange.org)

[Simon Tucker](#) is Manager, Corporate Strategy for Fonterra Cooperative Group Limited, based in Auckland, New Zealand... [www.nzuscouncil.com](http://www.nzuscouncil.com)

[Simon Tucker](#) is an experienced enterprise marketing professional... [www.linkedin.com](http://www.linkedin.com)

[Tucker](#) is with the Department of Information Studies, University of... [www.dcs.shef.ac.uk](http://www.dcs.shef.ac.uk)

[21 additional Quick Facts »](#)





### Personal Profiles

	<a href="#">Simon Tucker</a> . <a href="#">Simon</a> . 32 / male. London, United Kingdom... Personal Web Space - MySpace <a href="http://profile.myspace.com">profile.myspace.com</a> - Deep Web
	<a href="#">Rainman</a> . 24 / male. Reading - Caversham, South, United Kingdom... Personal Web Space - MySpace <a href="http://profile.myspace.com">profile.myspace.com</a> - Deep Web
	<a href="#">simon tucker</a> . ***super simon***. 22 / male. exeter/devon, United Kingdom... Personal Web Space - MySpace <a href="http://profile.myspace.com">profile.myspace.com</a> - Deep Web
	<a href="#">Tucker</a> . 20 / male. Leeds, East, United Kingdom... Personal Web Space - MySpace <a href="http://profile.myspace.com">profile.myspace.com</a> - Deep Web

Sponsored Tip: Find secret profiles and photos across MySpace, Facebook and 40+ networks... [www.spokeo.com](http://www.spokeo.com)





[46 additional Personal Profiles »](#)

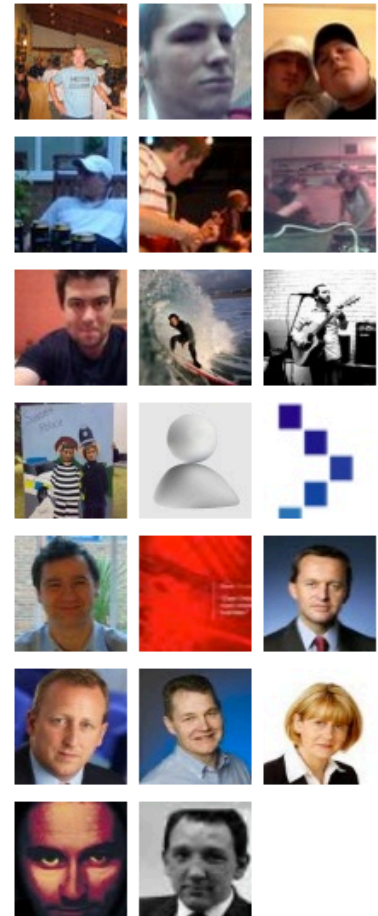
### Professional & Business

	<a href="#">Simon Tucker</a> , Attorney, Law, License: State: Dist. of Columbia , Status :... Attorney Rating - Avvo <a href="http://www.avvo.com">www.avvo.com</a> - Deep Web
	<a href="#">Simon Tucker</a> , GB, Opodo Limited... Web Extracted Biography - ZoomInfo <a href="http://www.zoominfo.com">www.zoominfo.com</a> - Deep Web
	<a href="#">Simon Tucker</a> , GB, Image Source Ltd... Web Extracted Biography - ZoomInfo <a href="http://www.zoominfo.com">www.zoominfo.com</a> - Deep Web
	<a href="#">Simon Tucker</a> , AU, Biota Holdings Limited... Web Extracted Biography - ZoomInfo <a href="http://www.zoominfo.com">www.zoominfo.com</a> - Deep Web


[46 additional Professional & Business »](#)

### Photos

	<a href="#">Jennifer MARANA</a> ... PhotoPages - Myspace <a href="http://viewmorepics.myspace.com">viewmorepics.myspace.com</a> - Deep Web
	<a href="#">Jennifer Beller</a> ... PhotoPages - Myspace <a href="http://viewmorepics.myspace.com">viewmorepics.myspace.com</a> - Deep Web
	<a href="#">Solar Captive</a> ... PhotoPages - Myspace <a href="http://viewmorepics.myspace.com">viewmorepics.myspace.com</a> - Deep Web
	<a href="#">Maxie Ray Mills</a> ...



### Sponsored Links

[Simon Tucker's Reputation](#)  
 Complete report. Manage your online reputation.  
[ReputationDefender.com](http://ReputationDefender.com)

# Motivation

- Web users increasingly use the Web to search for information about specific people, however:
  - Person names are ambiguous
- Current people search engines do not disambiguate
  - Concentrate on specific content sources
  - Limits information found: domain specific patterns
    - Limited in extracting information from new web documents
- Our solution = PSS (Person Search System)
  - Crawl the Web
  - Extract person information
  - Disambiguate between namesakes

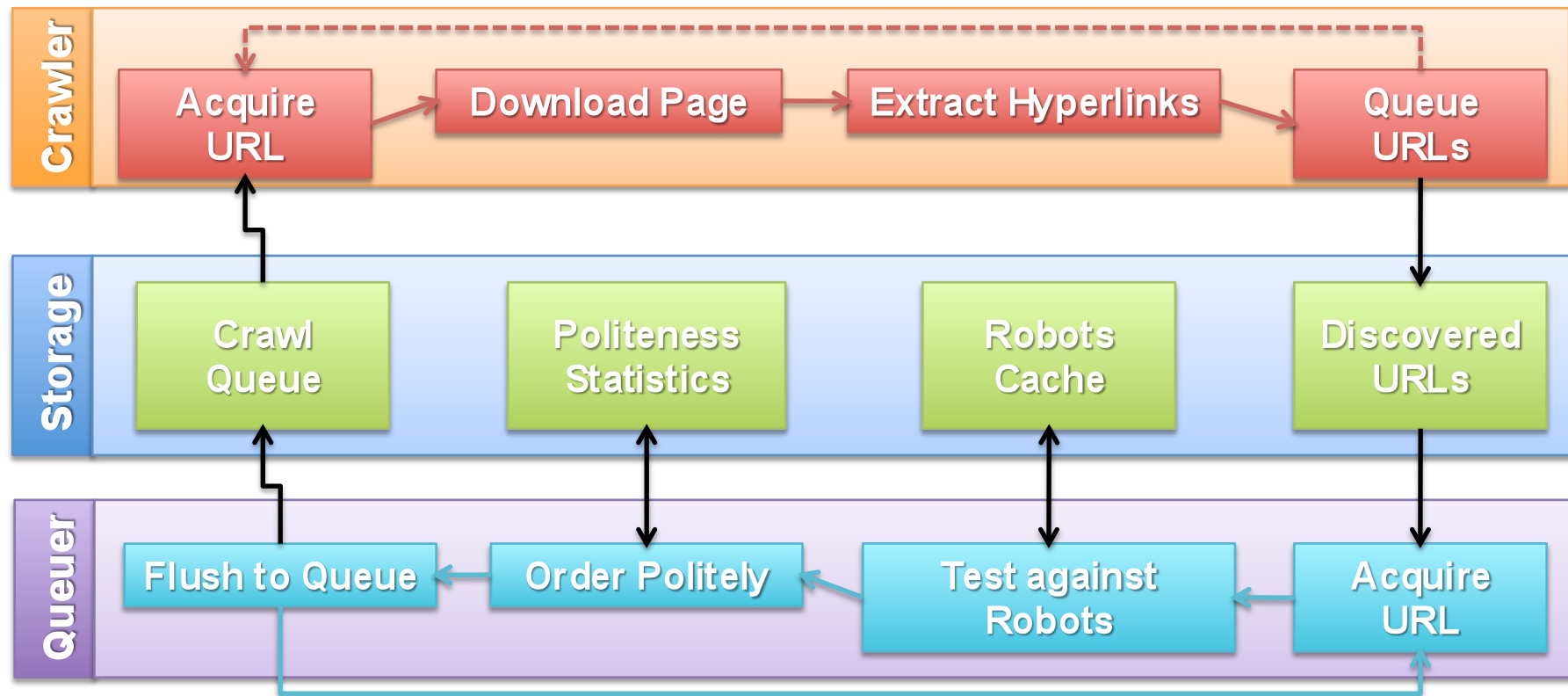
# Approach

- Build a “Reverse Index of Names”
  - Allowing pages containing specific names to be quickly located
  - Achieved by
    - Crawling The Web
    - Indexing Content
- Cluster based on Full Name
  - Create a list of URLs which contain names
    - Query the index looking of instances of adjacent First and Last names
  - Extract Content Windows from URLs
  - Identify features in Content Windows
  - Cluster URLs based on window features
- Web based interface for searching and clustering online
  - Clustering is expensive so clusters are temporarily cached
  - Offline clustering of popular names can also be enabled

# Crawling: Summary

- In the simplest form
  - Crawl a page from the Queue
    - Download the Page
    - Extract links
      - Queue links not yet crawled
    - Start over
- Problems
  - Too many to list!
  - Main problems
    - Storage (needed for indexing too)
    - Scalability
    - Large scale “politeness”
    - Required continuous monitoring
- Re-crawling
  - Not implemented, however the infrastructure supports it

# Crawling: Approach



# Crawling: Problems - Storage

- Crawling the Web requires lots of information to be stored
  - Even if not indexing or caching pages!
  - Lots of crawled pages must be stored
  - Queued URLs must be stored
  - Ideally (realistically) the data must be stored logically
- Solution:
  - Scalable Storage
    - Hadoop HDFS
  - Scalable Database System
    - Hadoop HBase



# Crawling: Problems - Storage



- The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on “commodity hardware” clusters.
  - Files are broken up into chunks (64 Mb)
  - Chunks are distributed across the cluster and replicated several (three) times.
- Hadoop HBase is a distributed database system aimed at situations where a table could have billions of rows and millions of columns requiring random read/write access
  - Scales instantly by adding new “region servers” which host row based regions of tables stored on HDFS

# HBase Table Example

The example below shows a table used to track when URLs were crawled and discovered

Row	crawled:/	crawled:/imghp	crawled:/links	seen:/about	crawled:/robots.txt
google.com	10:32 AM	10:33 AM			10:31 AM
shef.ac.uk			11:42 AM	09:12 AM	09:10 AM

- Each row can have completely different columns as they are not defined by any schema.
- The prefix on each row (crawled: or seen:) is known as the column family. These are predefined in the schema.
- Each column family can have specific parameters
  - TTL (Time to Live) – Expiry
  - Versions – Each cell also has an internal timestamp allowing the order of versions to be determined
- ALL data (rows, column families, columns and cells) are byte arrays
  - Very flexible!

# Crawling: Problems – Scalability

- As the crawl space grows operations often get slower
  - Removing duplicate URLs can be especially slow
    - Using a binary search method each lookup will take *roughly*  $\log_2[urls\ crawled]$  operations.
      - Checking against 256 URLs will take 8 operations
      - Checking against 1,000,000 URLs will take 20 operations
      - This might not seem much, but when crawling *up to* 400 page/sec when each page could have 10s or 100s of links on this adds up rapidly
    - Centralized ‘history’ of URLs crawled is ideally required
      - May put the load on one machine
      - Solution
        - » HBase – indexed regions across many nodes with much RAM reduces lookup time
- Each type of operation should ideally have the same complexity, regardless of the crawl space so that crawling can occur at a steady speed
  - Otherwise exponential slow down

# Crawling: Problems – Scalability

- As URLs are discovered they are added to a HBase table 'discovered'
  - The row they are added to corresponds to time of discovery, to the nearest second. This is known as a discovery window.
    - Each column is a URL found
    - The cell value is essentially pointless
  - The windows are processed sequentially by the Queuer
    - The window is emptied
    - Each URL is checked against the 'history' table and duplicates are removed
    - The URL is then checked against the appropriate robots file
      - Robots data is cached and prefetched for performance
    - The resulting URLs are queued AND added to the 'history' table
      - Domain based rows for quick lookups
      - Once in the history table a URL will not make it through the queuer again



# Crawling: Problems - Politeness

- When crawling lots of pages as quickly as possible it is easy to over crawl a site
  - especially if it is a link-rich seed site
  - or contains a sitemap, allowing rapid discovery of all internal links
- Solution
  - Time frame based crawl queue
    - As URLs are discovered they are put into the queue
    - The queue is broken up into frames
    - Each frame, at most can only contain one URL from each domain
    - If a frame is emptied in under a second a wait (sleep) is imposed until a second has elapsed
    - Therefore:
      - NO domain is crawled more than once per second
      - Re-crawl delay in robots can be adhered to
        - » Add domain every  $n$  frames, where  $n$  is the re-crawl delay in seconds

# Crawling: Problems - Politeness

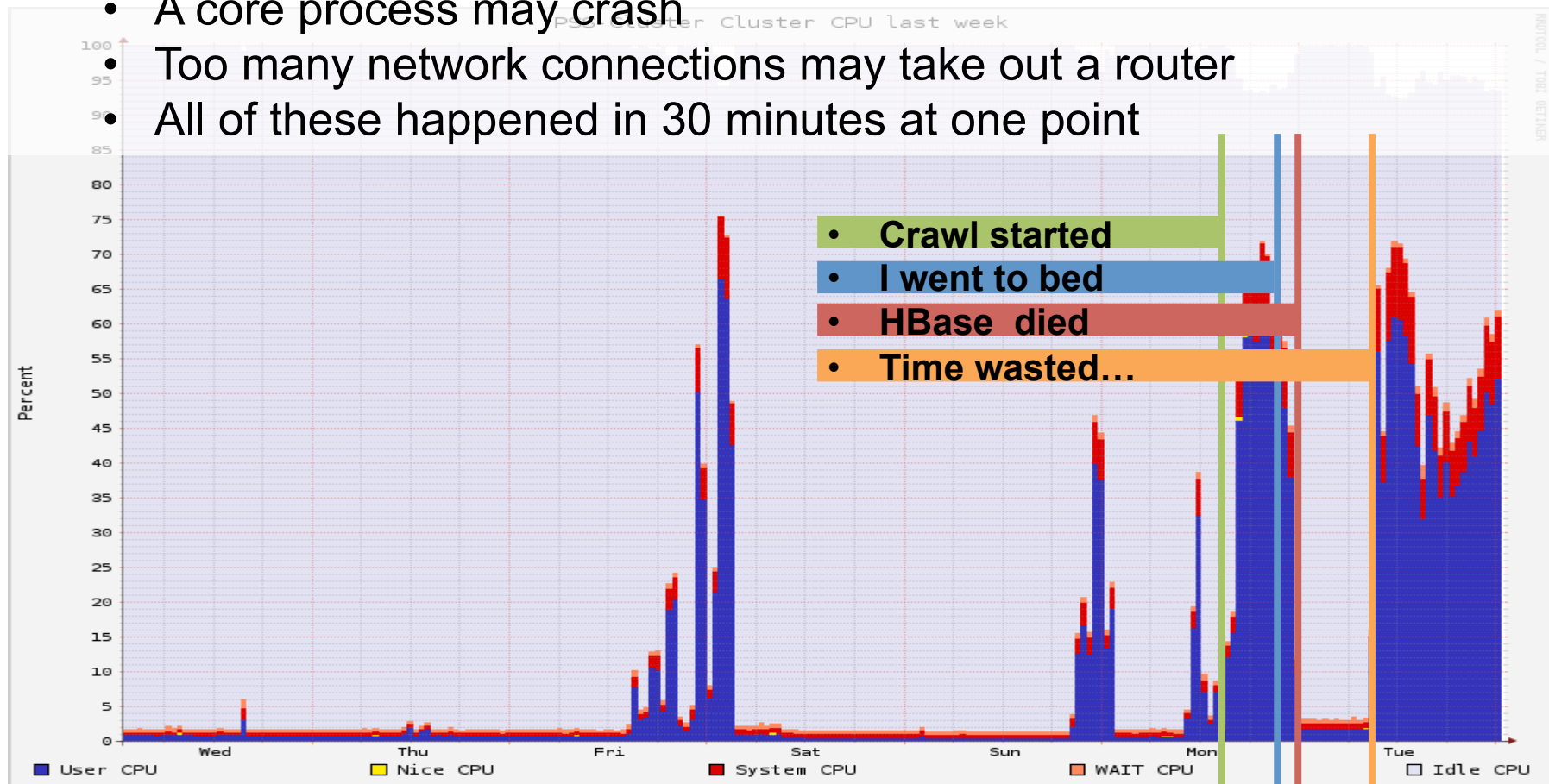
- Frames are allowed a limited number of URLs to ensure emptier frames can be balanced out so that sleeps are not required
- Framed Queue example:

'second' 1	'second' 2	'second' 3
<a href="http://www.shef.ac.uk/">http://www.shef.ac.uk/</a> <a href="http://www.dcs.shef.ac.uk/">http://www.dcs.shef.ac.uk/</a> <a href="http://www.google.com/">http://www.google.com/</a> <a href="http://www.w3c.org/">http://www.w3c.org/</a> <a href="http://www.internic.net/">http://www.internic.net/</a> <a href="http://www.dell.com/">http://www.dell.com/</a> <a href="http://www.192.com/">http://www.192.com/</a> <a href="http://www.123people.com/">http://www.123people.com/</a> <a href="http://oak.dcs.shef.ac.uk/">http://oak.dcs.shef.ac.uk/</a>	<a href="http://www.shef.ac.uk/links/">http://www.shef.ac.uk/links/</a> <a href="http://www.dcs.shef.ac.uk/intranet/">http://www.dcs.shef.ac.uk/intranet/</a> <a href="http://www.google.com/imghp/">http://www.google.com/imghp/</a> <a href="http://www.dell.com/sales/">http://www.dell.com/sales/</a> <a href="http://www.192.com/directory/">http://www.192.com/directory/</a> <a href="http://oak.dcs.shef.ac.uk/tools/">http://oak.dcs.shef.ac.uk/tools/</a>	<a href="http://www.shef.ac.uk/about/">http://www.shef.ac.uk/about/</a> <a href="http://www.google.com/news/">http://www.google.com/news/</a> <a href="http://www.192.com/search/">http://www.192.com/search/</a> <a href="http://oak.dcs.shef.ac.uk/blog">http://oak.dcs.shef.ac.uk/blog.</a>

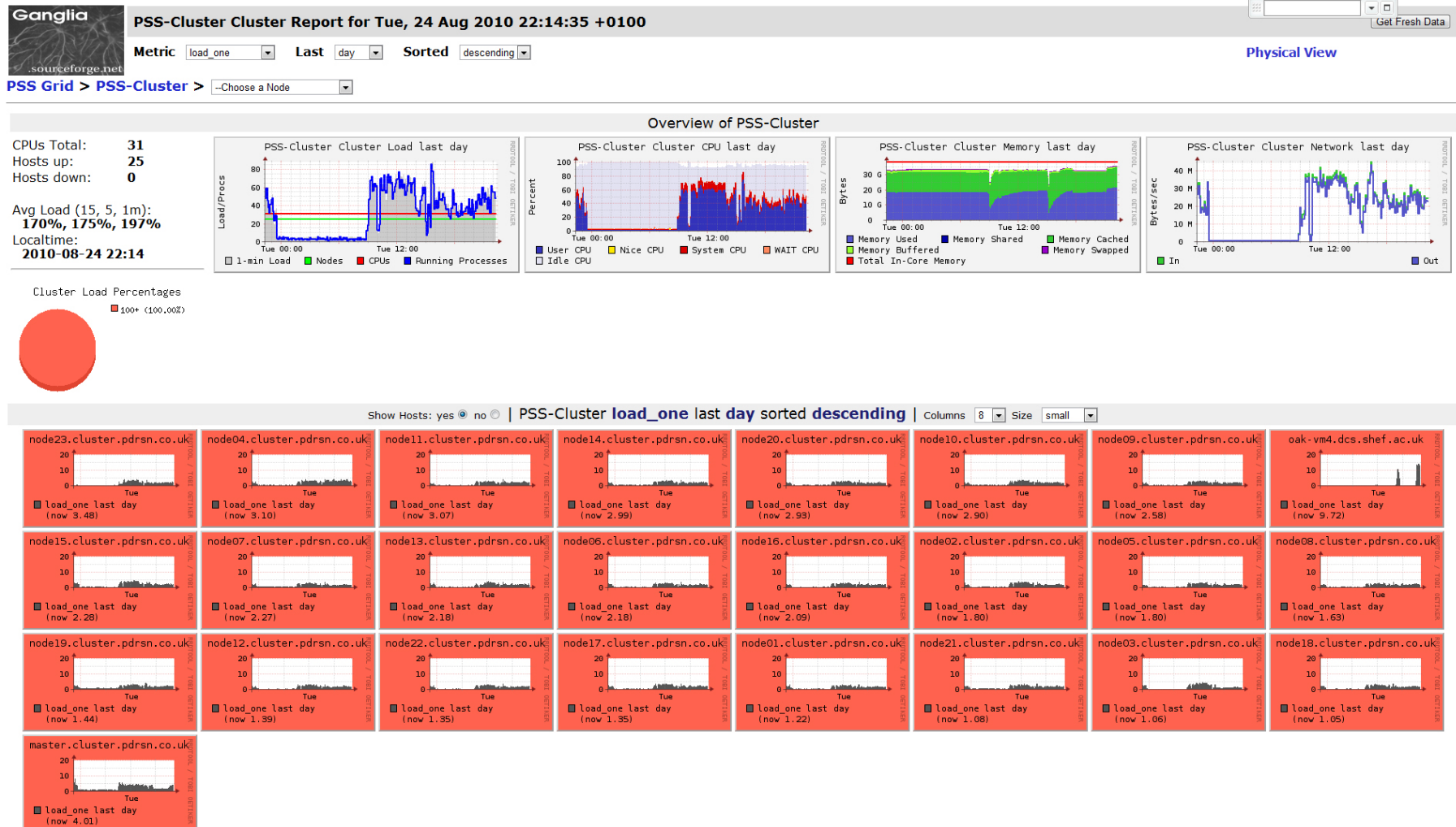
- Quick note: Robots rules are applied to discovered URLs before adding them to queue frames
- The last frame used by a domain and robots data is stored in a `politeness` table

# Crawling: Monitoring

- Events occur very rapidly when crawling on a large scale
  - A bug may cause a server to get 'abused' by the crawler
  - A core process may crash
  - Too many network connections may take out a router
  - All of these happened in 30 minutes at one point



# Crawling: Monitoring



Organisations,  
Information and  
Knowledge



# Crawling: Monitoring

<pre>ps@node06: ~\$ top CPU: 1.3% (idle) Tasks: 33 total, 1 running Mem: 1.24/1.31 G (93%) Load average: 1.24 1.31 1.63 Swap: 0/707MB Uptime: 8 days, 05:21:29  PID USER      NI  VIRT  RES  SHR S  CPU% MEM%   TIME+  Command 4863 root        19   703M 442M 7780 S 22.0 35.3 10:21.25 /usr/share/jdk/bin/java - 4776 root        18   676M 43212 8440 S 9.0 3.4 2:38.16 /usr/share/jdk/bin/java - 24403 pss         20   2456 1092 864 R 1.0 0.1 3:52.72 http 10145 nobody      20   77528 576 852 S 0.0 0.2 2:04.67 /usr/sbin/inetd 7034 root         20   8772 3086 2412 S 0.0 0.2 0:00.04 sshd: pss [priv] 723 root         20   21252 556 1512 S 0.0 0.2 0:04.50 /usr/sbin/console-kit-daem 597 messageb    20   2660 860 580 S 0.0 0.1 0:00.96 dbus-daemon --system --for 1 root         20   2760 1268 872 S 0.0 0.1 0:00.45 /sbin/init 260 root         20   2312 660 476 S 0.0 0.1 0:00.13 upstart-udev-bridge --daem 263 root        16   2304 620 252 S 0.0 0.0 0:00.03 udevd --daemon 348 root        18   2308 572 248 S 0.0 0.0 0:00.01 udevd --daemon F1 help F2Setup F3Search F4Invert F5Tree F6SortBy F7Nice F8Pause F9Kill F10Quit</pre>	<pre>ps@node07: ~\$ top CPU: 37.9% (idle) Tasks: 35 total, 1 running Mem: 1.39/1.46 G (95%) Load average: 1.39 1.46 1.42 Swap: 2/303MB Uptime: 8 days, 10:20:53  PID USER      NI  VIRT  RES  SHR S  CPU% MEM%   TIME+  Command 9229 pss         20   207M 67472 7356 S 0.0 45.4 1:43:25 bash /local/hbase/bin/./h 2130 pss         20   2456 1164 928 R 0.0 0.1 1:25:03 http 3438 root        18   672M 40268 8380 S 0.0 3.1 5:19.11 /usr/share/jdk/bin/java - 25602 root       20   4200 1220 1088 S 0.0 0.1 8:57:52 bash -c while true; do cle 11263 nobody     20   7692 3680 944 S 0.0 0.2 2:48.63 /usr/sbin/inetd 9225 pss         20   8772 1484 852 S 0.0 0.1 0:02.81 sshd: pss@notty 9207 root        20   8772 3036 2412 S 0.0 0.2 0:00.04 sshd: pss [priv] 711 root         20   21240 2860 1584 S 0.0 0.2 0:04.61 /usr/sbin/console-kit-daem 11855 ganglia     20   4228 1424 904 S 0.0 0.1 0:00.47 /usr/sbin/gmond 598 messageb    20   2660 744 588 S 0.0 0.1 0:00.91 dbus-daemon --system --for 555 syslog      20   3572 1208 916 S 0.0 0.1 0:01.56 rsyslogd -c4 F1 help F2Setup F3Search F4Invert F5Tree F6SortBy F7Nice F8Pause F9Kill F10Quit</pre>	<pre>ps@node08: ~\$ top CPU: 3.3% (idle) Tasks: 34 total, 1 running Mem: 1.04/1.05 G (98%) Load average: 1.24 1.63 1.70 Swap: 0/1447MB Uptime: 8 days, 09:57:03  PID USER      NI  VIRT  RES  SHR S  CPU% MEM%   TIME+  Command 17104 root       18   930M 38112 4316 S 0.0 43.5 22:01:69 bash /local/hbase/bin/./h 7394 pss         20   2460 1072 840 R 0.0 0.1 6:37:63 http 18949 pss        20   205M 66192 3408 S 0.0 3.2 1:52:26 java -cp out/production/ps 17191 root       19   935M 59900 4440 S 0.0 2.9 0:39.00 /usr/share/jdk/bin/java - 11089 nobody     20   7436 3968 860 S 0.0 0.1 2:47.34 /usr/sbin/inetd 18945 pss        20   8772 1132 496 S 0.0 0.1 0:02.89 sshd: pss@notty 18927 root        20   8772 1556 928 S 0.0 0.1 0:00.04 sshd: pss [priv] 722 root         20   21412 2388 680 S 0.0 0.1 0:05.30 /usr/sbin/console-kit-daem 1 root         20   2764 1024 628 S 0.0 0.0 0:00.45 /sbin/init 264 root         20   2312 500 320 S 0.0 0.0 0:00.13 upstart-udev-bridge --daem 267 root         16   2312 560 268 S 0.0 0.0 0:00.04 udevd --daemon F1 help F2Setup F3Search F4Invert F5Tree F6SortBy F7Nice F8Pause F9Kill F10Quit</pre>
<pre>ps@node09: ~\$ top CPU: 2.7% (idle) Tasks: 34 total, 1 running Mem: 1.73/1.80 G (96%) Load average: 1.47 1.61 1.60 Swap: 12/1447MB Uptime: 8 days, 11:22:27  PID USER      NI  VIRT  RES  SHR S  CPU% MEM%   TIME+  Command 10800 pss         20   8772 1312 456 S 0.0 0.1 0:02.92 sshd: pss@notty 722 root         20   21332 744 200 S 0.0 0.1 0:05.27 /usr/sbin/console-kit-daem 604 messageb    20   2660 356 212 S 0.0 0.0 0:01.14 dbus-daemon --system --for 593 syslog      20   3536 924 644 S 0.0 0.1 0:01.86 rsyslogd -c4 1 root         20   2760 636 372 S 0.0 0.1 0:00.42 /sbin/init 265 root        16   2312 328 244 S 0.0 0.0 0:00.04 udevd --daemon 341 root        18   2308 364 228 S 0.0 0.0 0:00.00 udevd --daemon 342 root        18   2308 176 120 S 0.0 0.0 0:00.00 udevd --daemon 612 root         20   2548 476 316 S 0.0 0.0 0:00.29 /usr/sbin/sshd 618 avahi        20   3024 588 408 S 0.0 0.1 0:00.34 avahi-daemon: running [node 620 avahi        20   2924 224 140 S 0.0 0.0 0:00.00 avahi-daemon: chroot helpe F1 help F2Setup F3Search F4Invert F5Tree F6SortBy F7Nice F8Pause F9Kill F10Quit</pre>	<pre>ps@node10: ~\$ top CPU: 69.2% (idle) Tasks: 34 total, 1 running Mem: 1.04/1.05 G (98%) Load average: 2.37 1.67 1.47 Swap: 0/1447MB Uptime: 8 days, 10:24:51  PID USER      NI  VIRT  RES  SHR S  CPU% MEM%   TIME+  Command 14243 root       20   206M 65088 7324 S 1.0 3.2 1:47:02 java -cp out/production/ps 13131 root       19   936M 64652 7720 S 0.0 3.1 0:39.11 /usr/share/jdk/bin/java - 13044 root       18   937M 45112 8420 S 25.0 2.2 20:41.14 nice -n -2 /local/hadoop/b 715 root         20   21332 1280 1424 S 0.0 0.2 0:05.94 /usr/sbin/console-kit-daem 889 root         20   8772 3052 2424 S 0.0 0.1 0:00.04 sshd: pss [priv] 14272 root       20   8772 3040 2412 S 0.0 0.1 0:00.04 sshd: pss [priv] 909 pss         20   5652 2776 1316 S 0.0 0.1 0:00.38 bash 11437 nobody     20   7612 2716 944 S 0.0 0.1 2:46.69 /usr/sbin/inetd 572 syslog      20   3572 1440 932 S 0.0 0.1 0:01.59 rsyslogd -c4 908 pss         20   8772 1512 864 S 0.0 0.1 0:00.62 sshd: pss@notty 14290 pss        20   8772 1488 852 S 0.0 0.1 0:02.95 sshd: pss@notty F1 help F2Setup F3Search F4Invert F5Tree F6SortBy F7Nice F8Pause F9Kill F10Quit</pre>	<pre>ps@node23: ~\$ top CPU: 3.0% (idle) Tasks: 40 total, 1 running Mem: 1.32/1.33 G (99%) Load average: 3.24 2.80 2.22 Swap: 0/768MB Uptime: 8 days, 11:14:54  PID USER      NI  VIRT  RES  SHR S  CPU% MEM%   TIME+  Command 19706 root       20   416M 55212 308 S 80.0 5.4 1:31:29 java -Xmx256M -cp out/prod 24240 root       20   8684 3092 2440 S 0.0 0.3 0:05.10 sshd: [accepted] 733 root         20   20988 3088 1704 S 0.0 0.3 0:04.29 /usr/sbin/console-kit-daem 32386 root       20   8528 3072 2440 S 0.0 0.3 0:01.36 sshd: root@pts/1 20083 root       20   8528 3072 2440 S 0.0 0.3 0:00.07 sshd: pss [priv] 6630 nobody     20   76364 2788 1200 S 0.0 0.3 3:10.03 /usr/sbin/inetd 667 haldaemon   20   6004 2136 1616 S 0.0 0.2 0:02.41 hald -daemon=yes 32450 root       20   4400 1520 1476 S 0.0 0.2 0:00.02 bash 24504 root       20   4400 1516 1472 S 0.0 0.2 0:00.03 bash 20147 pss         20   8528 1600 972 S 0.0 0.2 0:02.29 sshd: pss@notty 7836 ganglia     20   4144 1572 1060 S 0.0 0.2 0:21.03 /usr/sbin/gmond F1 help F2Setup F3Search F4Invert F5Tree F6SortBy F7Nice F8Pause F9Kill F10Quit</pre>
<pre>ps@master: ~\$ top bwm-ng v0.6 (probing every 1.000s), press 'h' for help input: /proc/net/dev type: rate -----             Rx              Tx              Total ----- lo:          105.56 KB/s      105.56 KB/s      211.11 KB/s eth0:        291.92 KB/s      195.02 KB/s      486.94 KB/s eth1:        117.90 KB/s      34.72 KB/s       152.62 KB/s ----- total:       515.38 KB/s      335.29 KB/s      850.67 KB/s</pre>	<pre>ps@master: ~\$ top SUCCESS URL: http://www.ddj.com/windows/226900048 SUCCESS URL: http://truetheadvertiser.com/blog/691 SUCCESS URL: http://religion.blogs.cnn.com/2010/08/22/my-interview-with-obama-on-his-chr istianity-and-the-muslim-issue/?replytocom=84727 SUCCESS URL: http://video.ign.com/dor/articles/965543/category/daily-fix-videos/thefix_spc_0 70810.html?download=true SUCCESS URL: http://vampiretravel.wordpress.com/category/heralders/ SUCCESS URL: http://aws.amazonaws.com/solutions/solution-providers/ SUCCESS URL: http://www.epicponyz.com/2009_09_28_archive.html SUCCESS URL: http://atlascosplay.com/globe/north-america/united-states/california/bodye- mountain SUCCESS URL: http://www.nickcannon.com/tag/nickcannon.com/ SUCCESS URL: http://mediosmexico.blogspot.com/2010/01/caracas-autoriza-reponer-la-tran- smision.html SUCCESS URL: http://anneandday.com/?p=219 SUCCESS URL: http://www.monergismbooks.com/skin/reformedbooks.html SUCCESS URL: http://ancienthebrewpoetry.typepad.com/ancient_hebrew_poetry/2005/05/the_gre- at_arrai.html</pre>	<pre>ps@master: ~\$ top Reducing: Kept: 9320 from 23248, 2591/4178 - com.wordpress.lindsaymeyer Reducing: Kept: 9320 from 23248, 2592/4178 - com.best-geiger-counter.www Reducing: Kept: 9333 from 23263, 2593/4178 - com.wordpress.ccasstuff Reducing: Kept: 9333 from 23265, 2594/4178 - org.market-ticker Reducing: Kept: 9337 from 23269, 2595/4178 - com.blogspot.vampibots Reducing: Kept: 9337 from 23270, 2596/4178 - com.educationfutures.www Reducing: Kept: 9338 from 23271, 2597/4178 - com.silicon.www Reducing: Kept: 9338 from 23272, 2598/4178 - com.wordpress.minilaptop410 Reducing: Kept: 9338 from 23273, 2599/4178 - com.wordpress.snoopyk2 Reducing: Kept: 9338 from 23274, 2600/4178 - com.wordpress.blognotherun Reducing: Kept: 9338 from 23275, 2601/4178 - com.kiroty.www Reducing: Kept: 9338 from 23276, 2602/4178 - com.coreofcorruption Reducing: Kept: 9342 from 23281, 2603/4178 - com.goldmoney Reducing: Kept: 9343 from 23282, 2604/4178 - com.radio.929dave Reducing: Kept: 9353 from 23308, 2605/4178 - com.wordpress.humairamarby Reducing: Kept: 9353 from 23309, 2606/4178 - de.jenny-engelchen.www Reducing: Kept: 9354 from 23310, 2607/4178 - com.stripegameer.www Reducing: Kept: 9354 from 23311, 2608/4178 - com.amazon.www</pre>
<pre>ps@node23: ~\$ top cc - Remain: 100 + 21. Succ: 60106, Err: 1954 Prefetch complete: net.soccerblogs (344) cc - Remain: 100 + 22. Succ: 60107, Err: 1954 Prefetch complete: com.wordpress.wompwomp vomp (273) cc - Remain: 100 + 23. Succ: 60108, Err: 1954 Prefetch complete: cn.js-edu.www (68) cc - Remain: 100 + 23. Succ: 60109, Err: 1954 Prefetch complete: com.wordpress.puzzling nyc (272) cc - Remain: 100 + 23. Succ: 60110, Err: 1954 Prefetch complete: com.u2ipo.www (166) cc - Remain: 100 + 23. Succ: 60111, Err: 1954 Prefetch complete: gov.bellevue.www (18 968) cc - Remain: 100 + 24. Succ: 60112, Err: 1954 Prefetch complete: cn.edu.sbc.www (0) cc - Remain: 100 + 24. Succ: 60113, Err: 1954 Prefetch complete: com.ssbj (0) cc - Remain: 100 + 24. Succ: 60114, Err: 1954 Prefetch complete: com.nwsource.seattletl se.forums (0) cc - Remain: 100 + 24. Succ: 60115, Err: 1954 Prefetch complete: fr.actu-entreprise-cit oyenne.www (24) cc - Remain: 100 + 25. Succ: 60116, Err: 1954 Prefetch complete: com.kus.tech (0) cc - Remain: 100 + 25. Succ: 60117, Err: 1954 Prefetch complete: net.ad-cn.www (0)</pre>	<pre>ps@master: ~\$ top node10: regionserver running as process 13240. Stop it first. node11: regionserver running as process 11446. Stop it first. node14: regionserver running as process 9919. Stop it first. node20: regionserver running as process 16199. Stop it first. node22: regionserver running as process 3237. Stop it first. node15: regionserver running as process 497. Stop it first. node19: starting regionserver, logging to /local/hbase/bin/./logs/hbase-root-regionser- er-node19.out node09: regionserver running as process 9607. Stop it first. node01: regionserver running as process 17049. Stop it first. node17: regionserver running as process 7718. Stop it first. node21: regionserver running as process 11137. Stop it first. node07: regionserver running as process 3577. Stop it first. node13: regionserver running as process 22019. Stop it first. node18: regionserver running as process 27848. Stop it first. node08: regionserver running as process 17305. Stop it first. node16: regionserver running as process 12775. Stop it first. node04: regionserver running as process 7239. Stop it first. root@master: /local/hbase/bin#</pre>	<pre>ps@master: ~\$ top Queue: 0, Gets: 2509210 (0/sec), Over: 112331 sec, TotAvg: 22. MinAvg: 26 Queue under 200, loading more... Active bucket: 2011 Last Full bucket: 2011 Last Empty bucket: 2011 No usable bucket - stall Queue: 0, Gets: 2509210 (0/sec), Over: 112333 sec, TotAvg: 22. MinAvg: 26 Queue under 200, loading more... Active bucket: 2011 Last Full bucket: 2011 Last Empty bucket: 2011 No usable bucket - stall Queue: 0, Gets: 2509210 (0/sec), Over: 112335 sec, TotAvg: 22. MinAvg: 26 Queue under 200, loading more... Active bucket: 2011 Last Full bucket: 2011 Last Empty bucket: 2011 No usable bucket - stall</pre>

12:58 3.13 3.10 1 Tue Aug 24 22:13:53 BST 2010



Organisations,  
Information and  
Knowledge

# NameNode 'master.cluster.pdrsn.co.uk:54310'

**Started:** Tue Aug 24 15:07:58 BST 2010  
**Version:** 0.20.2, r911707  
**Compiled:** Fri Feb 19 08:07:34 UTC 2010 by chrisdo  
**Upgrades:** There are no upgrades in progress.

[Browse the filesystem](#)  
[Namenode Logs](#)

## Cluster Summary

8010 files and directories, 5786 blocks = 13796 total. Heap Size is 12.38 MB / 739.56 MB (1%)

**Configured Capacity** : 1.35 TB  
**DFS Used** : 293.87 GB  
**Non DFS Used** : 101.11 GB  
**DFS Remaining** : 988.32 GB  
**DFS Used%** : 21.24 %  
**DFS Remaining%** : 71.45 %  
**Live Nodes** : 21  
**Dead Nodes** : 0

## NameNode Storage:

Storage Directory	Type	State
/hadoop-name	IMAGE_AND_EDITS	Active

[Hadoop](#), 2010.

Node	Last Contact	Admin State	Configured Capacity (GB)	Used (GB)	Non DFS Used (GB)	Remaining (GB)	Used (%)	Used (%)	Remaining (%)	Blocks
node01	0	In Service	72.5	16.3	5.16	51.04	22.49		70.4	785
node02	0	In Service	73.21	14.86	5.09	53.26	20.3		72.74	900
node04	0	In Service	71.44	14.86	5.11	51.47	20.8		72.05	780
node05	2	In Service	72.5	14.85	5.18	52.46	20.49		72.36	836
node06	1	In Service	73.21	15.56	5.08	52.57	21.25		71.8	801
node07	2	In Service	4.39	2.42	1.62	0.35	55.07		8	148
node08	0	In Service	72.5	15.41	5.14	51.96	21.25		71.66	791
node09	0	In Service	72.5	15.1	5.09	52.31	20.82		72.15	879
node10	1	In Service	72.5	14.37	5.08	53.05	19.82		73.17	759
node11	0	In Service	70.73	15.05	4.96	50.71	21.28		71.7	813
node12	1	In Service	70.73	15.21	5.01	50.5	21.51		71.4	857
node13	0	In Service	73.21	14.65	5.07	53.49	20.02		73.07	769
node14	1	In Service	72.5	14.9	5.12	52.48	20.56		72.38	876
node15	1	In Service	8.17	6.02	1.82	0.33	73.71		3.99	342
node16	2	In Service	70.73	14.88	5.07	50.77	21.04		71.79	741
node17	0	In Service	72.85	15.56	5.17	52.12	21.36		71.55	812
node18	1	In Service	72.5	16.13	5.14	51.24	22.24		70.67	888
node19	2	In Service	71.79	14.83	5.17	51.78	20.66		72.13	907
node20	2	In Service	71.08	16.12	5.11	49.85	22.68		70.14	886
node21	1	In Service	71.08	14.96	5.15	50.97	21.04		71.72	893
node22	0	In Service	73.21	15.02	5.07	53.13	20.51		72.57	933



Organisations,  
Information and  
Knowledge

# Crawling: Monitoring

Browser tabs: Hadoop NameNode mas..., HBase Master: master.clu...  
Address bar: http://rdfa.dcs.shef.ac.uk:60010/master.jsp  
Bookmarks: Mail, Calendar, Tasks, Archived, Dissertation/Summe..., Server Stuff, Programming, master:54310, master:60000, Ganglia: PSS Cluster..., dev-node01:54310, dev-node01:60000, Other bookmarks

**HBase** Master: master.cluster.pdrsn.co.uk:60000  
[Local logs](#), [Thread Dump](#), [Log Level](#)

### Master Attributes

Attribute Name	Value	Description
HBase Version	0.20.6, r965666	HBase version and svn revision
HBase Compiled	Mon Jul 19 16:54:48 PDT 2010, stack	When HBase version was compiled and by whom
Hadoop Version	0.20.2, r911707	Hadoop version and svn revision
Hadoop Compiled	Fri Feb 19 08:07:34 UTC 2010, chrisdo	When Hadoop version was compiled and by whom
HBase Root Directory	hdfs://master:54310/hbase	Location of HBase home directory
Load average	76.04761904761905	Average number of regions per regionserver. Naïve computation.
Regions On FS	1640	Number of regions on FileSystem. Rough count.
Zookeeper Quorum	node21:2222,node20:2222,node19:2222,node16:2222,node18:2222,node14:2222,node17:2222,node12:2222,node11:2222,node10:2222,node09:2222,node08:2222,node05:2222,node01:2222,master:2222	Addresses of all registered ZK servers. For more, see <a href="#">zk.dump</a> .

### Catalog Tables

Table	Description
<a href="#">.ROOT.</a>	The .ROOT. table holds references to all .META. regions.
<a href="#">.META.</a>	The .META. table holds references to all User Table regions

### User Tables

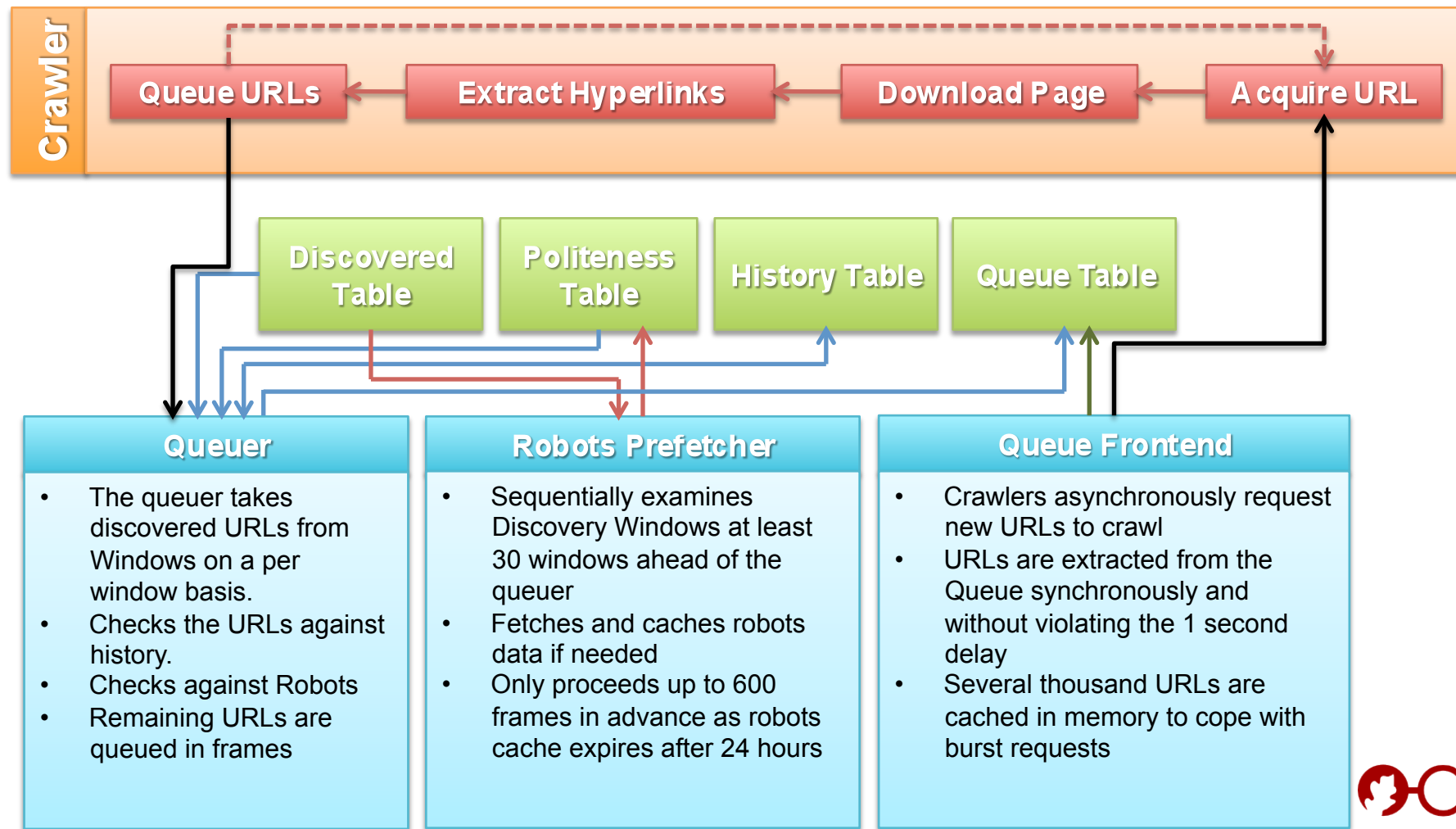
6 table(s) in set.

Table	Description
<a href="#">discovered</a>	{NAME => 'discovered', MAX_FILESIZE => '67108864', MEMSTORE_FLUSHSIZE => '16777216', FAMILIES => [{NAME => 'seen', COMPRESSION => 'NONE', VERSIONS => '1', TTL => '2147483647', BLOCKSIZE => '65536', IN_MEMORY => 'false', BLOCKCACHE => 'true'}]}
<a href="#">history</a>	{NAME => 'history', MAX_FILESIZE => '67108864', MEMSTORE_FLUSHSIZE => '16777216', FAMILIES => [{NAME => 'queued', COMPRESSION => 'NONE', VERSIONS => '1', TTL => '2147483647', BLOCKSIZE => '65536', IN_MEMORY => 'false', BLOCKCACHE => 'true'}]}
<a href="#">index</a>	{NAME => 'index', MEMSTORE_FLUSHSIZE => '16777216', MAX_FILESIZE => '67108864', FAMILIES => [{NAME => 'count', VERSIONS => '1', COMPRESSION => 'NONE', TTL => '2147483647', BLOCKSIZE => '65536', IN_MEMORY => 'false', BLOCKCACHE => 'true'}, {NAME => 'locations', COMPRESSION => 'NONE', VERSIONS => '1', TTL => '2147483647', BLOCKSIZE => '65536', IN_MEMORY => 'false', BLOCKCACHE => 'true'}]}
<a href="#">politeness</a>	{NAME => 'politeness', MAX_FILESIZE => '67108864', MEMSTORE_FLUSHSIZE => '16777216', FAMILIES => [{NAME => 'buckets', VERSIONS => '1', COMPRESSION => 'NONE', TTL => '2147483647', BLOCKSIZE => '65536', IN_MEMORY => 'false', BLOCKCACHE => 'true'}, {NAME => 'robots', COMPRESSION => 'NONE', VERSIONS => '5', TTL => '2147483647', BLOCKSIZE => '65536', IN_MEMORY => 'false', BLOCKCACHE => 'true'}]}
<a href="#">queue</a>	{NAME => 'queue', MEMSTORE_FLUSHSIZE => '16777216', MAX_FILESIZE => '67108864', FAMILIES => [{NAME => 'entry', VERSIONS => '1', COMPRESSION => 'NONE', TTL => '2147483647', BLOCKSIZE => '65536', IN_MEMORY => 'false', BLOCKCACHE => 'true'}, {NAME => 'stats', COMPRESSION => 'NONE', VERSIONS => '1', TTL => '2147483647', BLOCKSIZE => '65536', IN_MEMORY => 'false', BLOCKCACHE => 'true'}]}



Organisations,  
Information and  
Knowledge

# Crawling: Implementation

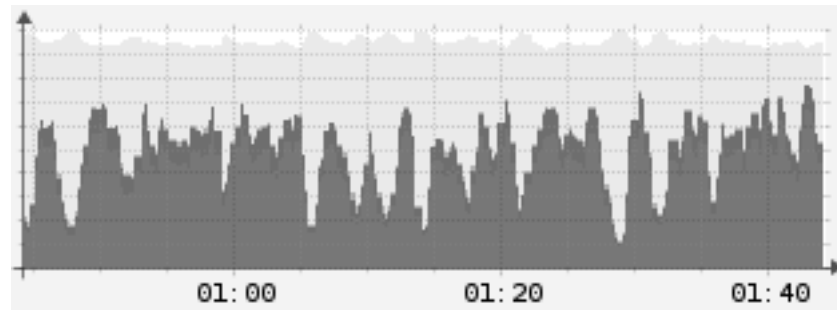


# Crawling: Conclusions

- Lots (too much) to potentially talk about
- Most importantly:
  - Scalability is always a major problem
    - Constant (or at least not exponential) complexity of operations
      - Especially duplication checking
  - Politeness
    - It is easy to annoy a lot of people (we got 2 complaints, so far...)
    - Especially as small sites pay for bandwidth
  - Bottlenecks always shift, you **cannot** eliminate them!
  - Strange patterns occur even when running well
    - Ripple effect

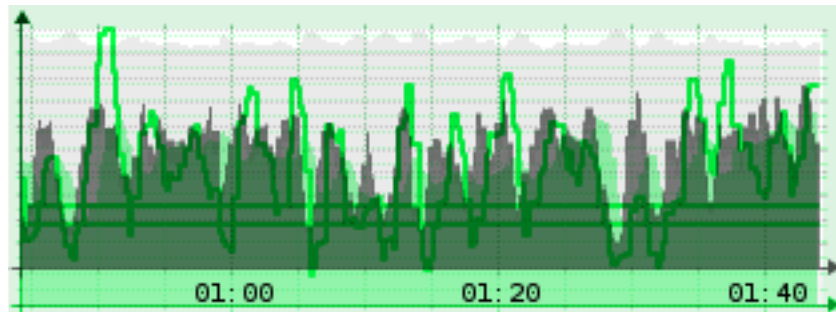
# Crawling: Conclusions

- Lots (too much) to potentially talk about
- Most importantly:
  - Scalability is always a major problem
    - Constant (or at least not exponential) complexity of operations
      - Especially duplication checking
  - Politeness
    - It is easy to annoy a lot of people (we got 2 complaints, so far...)
    - Especially as small sites pay for bandwidth
  - Bottlenecks always shift, you **cannot** eliminate them!
  - Strange patterns occur even when running well
    - Ripple effect



# Crawling: Conclusions

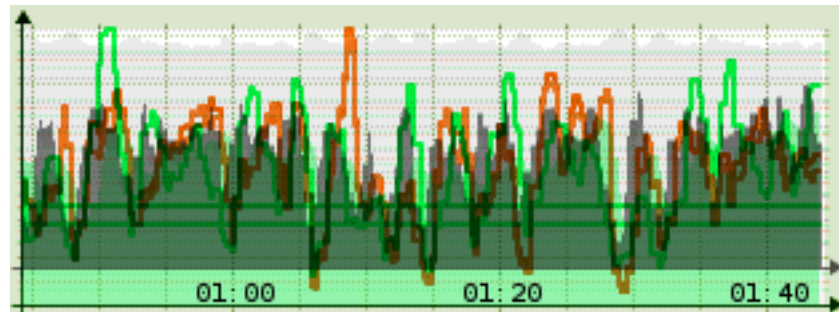
- Lots (too much) to potentially talk about
- Most importantly:
  - Scalability is always a major problem
    - Constant (or at least not exponential) complexity of operations
      - Especially duplication checking
  - Politeness
    - It is easy to annoy a lot of people (we got 2 complaints, so far...)
    - Especially as small sites pay for bandwidth
  - Bottlenecks always shift, you **cannot** eliminate them!
  - Strange patterns occur even when running well
    - Ripple effect





# Crawling: Conclusions

- Lots (too much) to potentially talk about
- Most importantly:
  - Scalability is always a major problem
    - Constant (or at least not exponential) complexity of operations
      - Especially duplication checking
  - Politeness
    - It is easy to annoy a lot of people (we got 2 complaints, so far...)
    - Especially as small sites pay for bandwidth
  - Bottlenecks always shift, you **cannot** eliminate them!
  - Strange patterns occur even when running well
    - Ripple effect



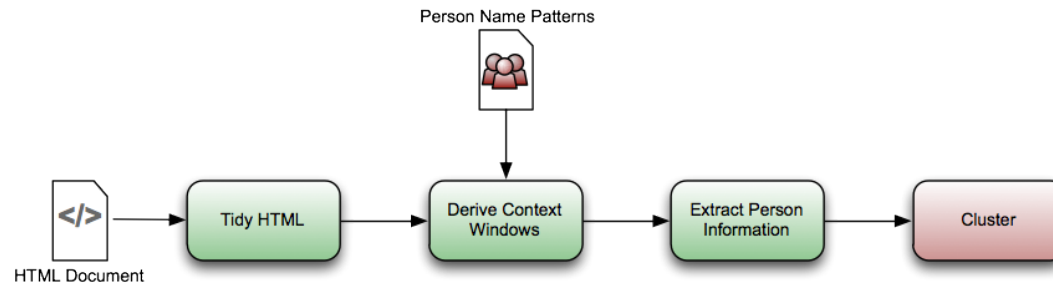


# Indexing

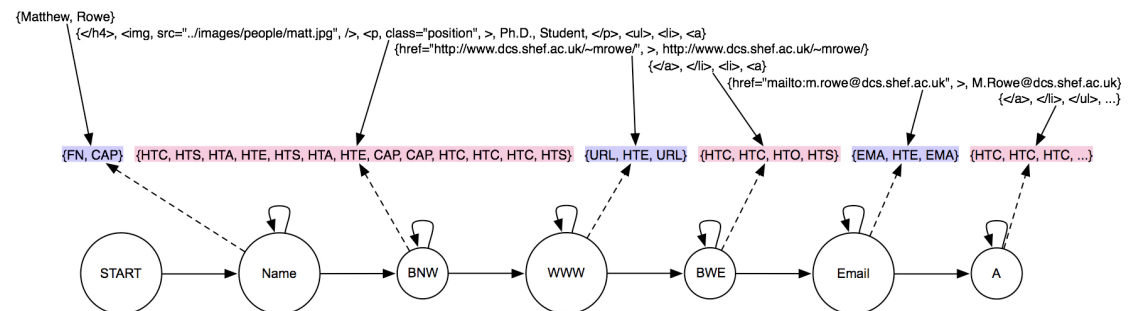
- Indexing is carried out as part of the crawling process
  - Once a page is downloaded it is scanned
  - Each word is identified (**not** using RegExs)
    - Locations of Pronouns (words starting with a capital letter) are recorded
    - Each page ends up with a list of Pronouns which Resolve to their original position in the Page
      - `HashMap<String, HashSet<Integer>>`
  - The Index table is then updated
    - Each Pronoun has it's own row
    - Each URL has a column
    - The value at the coordinates (row,column) represents a list of locations stored as sequential 32-bit integers
- For clustering later it would be useful to cache pages, unfortunately this requires lots of disk space...



# Information Extraction



- Extracts person information (i.e., relation extraction)
  - Pages to extract from identified by index
    - Spot occurrences <first\_name> <last\_name>
  - Content Windows derived using name patterns
  - Use HMMs to extract information regarding a single person
    - Name
    - Email
    - Website
    - Location



# Clustering

- Goal =
  - Query index using: <first\_name> <last\_name>
  - Returns set of documents: contains different namesakes
  - Group documents into separate clusters, where
    - **One cluster contains documents referring to one namesake**
- Tested 3 methods for clustering
  - DBScan (measures density outliers)
  - Xmeans (extension of k-means: estimates k)
  - Agglomerative Hierarchical Clustering (singleton cluster merging)
- Feature vector composition
  - Bag-of-words model for each document (from extracted features)
  - Features weighted based on TF/IDF scores
- Tuning: Web People Search Evaluation (WEPS) 07 training split
  - Choosing maximum params based on F0.5(purity,inverse\_purity)
- Testing: WEPS07 and WEPS09 test splits
  - Measured: purity, inverse purity, Bcubed Precision, Bcubed Recall, F0.2(prec, rec), F0.5(prec, rec), F0.5(pur,inv\_pur)



# Clustering

		Purity	Inverse Purity	B3 Precision	B3 Recall	F02	F05	F05(P,IP)
weps07	DBScan (m=1, e=0.9)	0.476	<b>0.759</b>	0.494	<b>0.748</b>	<b>0.620</b>	<b>0.550</b>	0.500
	Xmeans	<b>0.929</b>	0.263	<b>0.897</b>	0.155	0.171	0.214	0.528
	Agglomerative (t=0.1)	0.692	0.428	0.657	0.345	0.321	0.367	<b>0.556</b>
weps09	DBScan (m=1, e=0.9)	0.263	<b>0.693</b>	0.408	<b>0.679</b>	<b>0.544</b>	<b>0.466</b>	0.284
	Xmeans	<b>0.927</b>	0.436	<b>0.919</b>	0.276	0.304	0.366	<b>0.677</b>
	Agglomerative (t=0.1)	0.476	0.553	0.530	0.444	0.396	0.391	0.438

- With respect to SoA systems using same datasets
  - WEPS07
    - Agglo ranks 12<sup>th</sup> out of 16 Entries for F0.5(P,IP)
    - Xmeans ranks 1<sup>st</sup> out of 16 Entries for Purity
  - WEPS09
    - Xmeans ranks 1<sup>st</sup> out of 16 Entries for Purity
    - Xmeans ranks 12<sup>th</sup> out of 22 Entries for F0.5(P,IP)

per.sn

## Query:

Enter a name in the format *first last*:

☐ Ignore cache

## Results

**Note:** This result was retrived from the cache

### Cluster 1

- <http://www.iuiconf.org/pastui/09program.html>
- <http://www.informatik.uni-trier.de/%7Eley/db/conf/naacl/naacl2007.html>
- [http://www.informatik.uni-trier.de/%7Eley/db/indices/a-tree/m/Moore:Johanna\\_D=.html](http://www.informatik.uni-trier.de/%7Eley/db/indices/a-tree/m/Moore:Johanna_D=.html)
- [http://www.informatik.uni-trier.de/%7Eley/db/indices/a-tree/m/Moore:Johanna\\_D=.html](http://www.informatik.uni-trier.de/%7Eley/db/indices/a-tree/m/Moore:Johanna_D=.html)
- <http://www.informatik.uni-trier.de/%7Eley/db/indices/a-tree/c/Carletta:Jean.html>
- <http://www.informatik.uni-trier.de/%7Eley/db/indices/a-tree/c/Carletta:Jean.html>
- <http://www.informatik.uni-trier.de/%7Eley/db/indices/a-tree/h/Hsueh:Pei=yun.html>
- <http://www.informatik.uni-trier.de/%7Eley/db/indices/a-tree/h/Hsueh:Pei=yun.html>
- <http://www.informatik.uni-trier.de/%7Eley/db/indices/a-tree/w/Wellner:Pierre.html>
- <http://www.informatik.uni-trier.de/%7Eley/db/indices/a-tree/w/Wellner:Pierre.html>
- <http://www.informatik.uni-trier.de/%7Eley/db/indices/a-tree/r/Renals:Steve.html>
- <http://www.informatik.uni-trier.de/%7Eley/db/indices/a-tree/r/Renals:Steve.html>
- <http://www.informatik.uni-trier.de/%7Eley/db/indices/a-tree/t/Tucker:Simon.html>
- <http://www.informatik.uni-trier.de/%7Eley/db/indices/a-tree/t/Tucker:Simon.html>
- <http://www.informatik.uni-trier.de/%7Eley/db/conf/mlmi/mlmi2008.html>
- <http://oak.dcs.shef.ac.uk/people/>
- <http://www.informatik.uni-trier.de/%7Eley/db/indices/a-tree/m/Murray:Gabriel.html>
- <http://www.informatik.uni-trier.de/%7Eley/db/indices/a-tree/m/Murray:Gabriel.html>
- <http://www.informatik.uni-trier.de/%7Eley/db/conf/mlmi/mlmi2004.html>
- <http://www.informatik.uni-trier.de/%7Eley/db/conf/mlmi/mlmi2005.html>
- [http://videolectures.net/mlmi04ch\\_martigny/](http://videolectures.net/mlmi04ch_martigny/)
- <http://www.informatik.uni-trier.de/%7Eley/db/indices/a-tree/k/Kilgour:Jonathan.html>
- <http://www.informatik.uni-trier.de/%7Eley/db/indices/a-tree/k/Kilgour:Jonathan.html>
- <http://www.informatik.uni-trier.de/%7Eley/db/indices/a-tree/f/Flynn:Mike.html>
- <http://www.informatik.uni-trier.de/%7Eley/db/indices/a-tree/f/Flynn:Mike.html>
- <http://www.idiap.ch/events/workshop-mlmi04>

### Associated People

, Andreas Pleuss, Robert Farrell, Shlomo Argamon, Anja Belz, John S., Kofi Boakye, Aria Haghighi, Alfred Dielmann, Mari Ostendorf, Stephen Isard, Edward Loper, Ye-Yi Wang, Xavier Anguera, E., Brian Lathrop, Nicolas Moënné-Loccoz, .O. Box, Mike Hochberg, Joe Mattis, Kenneth Ward, Sameer Pradhan, Jose San, .J. Watson, Michael Voit, D. A., Stefan Sauer, Tracy Hammond, Stanley Peters, Ilana Bromberg, Ben J., Lucian Galescu, Liang Zhou, Chuck Wooters,, Gabriel Murray,, Robert Moore, Daniel Jurafsky, Laura Smith, Katherine Forbes-Riley, Mary Shaw, Angel Puerta, Irene Kimbara, Lukas Burget, Jean, Barry Schiffman, Stephen Perona, Emilia Stoica, Iain Mccowan, Karl Gyllstrom, Frank Shipman, Martin Karafiát, Vivek Kumar, Jon Oberlander, Giuseppe Attardi, Jamie Callan, Simone Stumpf, Steve Renals,, Melissa Kronenthal, Gregoire Burel, Antonio Sanfilippo, James Allan, Ulrich

# Conclusions

- Crawling
  - Scalability: hard to scale linearly
  - Politeness: increased speed inhibits politeness
- Information Extraction
  - Limited by supervised training
  - Semi-supervision could increase features collected
- Clustering
  - Performance is too low
    - Xmeans: inverse purity is too low
      - Produces insufficient cluster numbers (low k)
    - Agglomerative: purity is too low
      - Produces too many clusters (high k)

# Future Work

- Information Extraction
  - Retrain HMMs using induced observation patterns
    - Boost coverage over URLs
- Clustering
  - Increase features used:
    - Additional person attributes
      - From WEPS Attribute Extraction Challenge
    - Pronouns
  - 2 phase clustering:
    - Apply Xmeans followed by Agglomerative
      - Split initial high purity clusters to boost recall

# Questions?