Twittervane

Crowd sourcing for web archiving

Helen Hockx-Yu, Stephen Johnson, Maureen Pennock

British Library

Introduction and objectives

- Project funded by the IIPC
- Current selection process is largely manual by a small number of experts
  - Expensive & time consuming
  - Cannot respond to sudden events quickly
  - Subjective
  - Does not scale up
- Explore automatic selection
  - Exploit the wisdom of the crowd
  - Social setting: Twitter
  - Use popularity of websites as selection criteria
  - Compliment manual selection

Deliverables

- A prototype application for capturing and analysis of URLs contained in tweets.
  - which websites are shared most frequently around a given theme over a given time period
- Integration with existing web archiving infrastructure to harvest selected websites from the UK
- One pilot collection based on selection by Twittervane
- Vetting and assessment by curators

General Approach

- Use the Twitter Streaming API to capture tweets
  - best API for capturing tweets filtered by terms over extended period of time
- Only a single connection with Twitter is allowed at a time
  - Matches returned for all collection search terms
  - Tweets have to be re-matched to the collections post-capture
  - URLs are then extracted and expanded
- No attempt to capture backdated tweets through the search API at this stage
- Reports on URL/Domain by number of occurrences or by popularity (re-tweets
- Top URLs from TwitterVane xml output retrieved and automatically entered into our ‘Online Selection Tool’

TwitterVane

Crowd sourcing for Web Archiving

Welcome

Selection of web archive is a manual process that relies upon people to select quality sites and nominate or submit them to web archives for inclusion. Past (national and/or international) efforts to automate selection have either failed to convince staff that automation was identifying quality resources, or have focused on providing an interface for selectors to submit sites rather than carry out the selection per se. Selection has thus remained, for most institutions, an element of the workflow wholly dependent on contributions from externals. The number of selectors providing sites is typically small and their contributions are inevitably subjective. The resulting collections, whilst immensely valuable, are therefore mostly representative of the expertly selected sites and do not fully represent the sites frequently used in a more social setting.

Crowd sourcing is an opportunity to develop a new approach to this problem. It taps into the growth of social networks to outsource tasks typically performed by an employee or contractor, to an undefined, large group of people or community (a "crowd")â€™ (Wikipedia, 2011). It is a particularly attractive option in the current economic climate, where we are all being asked to 'do more with less'. A number of cultural heritage institutions and/or projects have already begun to leverage the power of the crowd for digitised collections, including the National Library of Australia (through Trove), the Transcribe Bentham project at ULCC, and the National Library of Finland (through the DigitalKoot program). No such projects have yet been launched for web archives.

This project will develop an automated approach to selection for web archiving based on the principles of crowd sourcing. It has been awarded partial funding from the International Internet Preservation Consortium and supports the forthcoming web archiving strategy by increasing the number of selections and automating the selection process.

Sidebar Menu

Configuration

Analysis Jobs

Links

Twitter API

UK Web Archive

Sponsors

IIPC

TwitterVane

Crowd sourcing for Web Archiving

Collections

Collection	Start Date	End Date	Search Terms	
Olympics	April 15, 2012	April 30, 2012	olympics paralympic london2012 lo2012	Delete
Diamond Jubilee	April 15, 2012	April 30, 2012	diamond jubilee diamondjubilee jubileeyacht	Delete
St George's day	April 15, 2012	April 30, 2012	stgeorgesday	Delete
Petrol	April 16, 2012	April 30, 2012	panic buying panicbuying fuel shortage fuelshortage nofuel no fuel	Delete
DP Reference Stack	April 24, 2012	April 30, 2013	dpref	Delete
Running Techniques	April 26, 2012	May 10, 2012	Running Technqiues	Delete

Add New Collection

Name

Start Date

End Date

Add

Sidebar Menu

Configuration

Analysis Jobs

Links

Twitter API

UK Web Archive

Sponsors

IIPC

TwitterVane

Crowd sourcing for Web Archiving

Jobs

Number of entities to process

200 Submit

Job Results

Total Processed:

Entity Summary

Total Tweets: 122094

Tweets Waiting for Analysis: 24105

URLs Analyzed: 96812

Purge all entities (Warning, This will delete all current analysis.)

Sidebar Menu

Configuration

Analysis Jobs

Links

Twitter API

UK Web Archive

Sponsors

IIPC

TwitterVane

Crowd sourcing for Web Archiving

Latest Tweets

Tweeter	Date	Retweet	URL Entities	Tweet
Eyad Ahmad	2012-04-26	0	View URL(s)	View Tweet
SAM-	2012-04-26	0	View URL(s)	View Tweet
fairappletraveller	2012-04-26	4	View URL(s)	View Tweet
DailyGawk	2012-04-26	0	View URL(s)	View Tweet
Gossip Detector	2012-04-26	0	View URL(s)	View Tweet
NetDoctor	2012-04-26	0	View URL(s)	View Tweet
Jutta Gue	2012-04-26	0	View URL(s)	View Tweet
Kenda Khalil	2012-04-26	0	View URL(s)	View Tweet
pete harrison	2012-04-26	0	View URL(s)	View Tweet
inicia FP	2012-04-26	0	View URL(s)	View Tweet

?@YourAnonNews: London 2012 Olympics Won?t Allow Sharing of Photos and Video via Social Networks | http://t.co/Jc6aBY47? @FarhanK501

Total Tweets: 122094

Sidebar Menu

Configuration

Twitter API

UK Web Archive

Sponsors

IIPC

TwitterVane

Crowd sourcing for Web Archiving

Reports

Collection

Olympics

Report Type

Top URLs

Filter URL

Filter Domain

.uk

Generate Report

Top URLs
Top Domains
Top URL by Retweet
Failed Analysis

XML request: http://explorer.bl.uk:8090/crowdsourcing2/rest/xmlreport/1/filterDomain/__uk

Sidebar Menu

Configuration

Analysis Jobs

Links

Twitter API

UK Web Archive

Sponsors

IIPC

TwitterVane

Crowd sourcing for Web Archiving

Reports

Total Tweets: 12955
Total URLs: 3959
Total Domains: 1390

KeyWords: olympics, paralympic, london2012, lo2012

Sidebar Menu

Configuration
Analysis Jobs

Links

Twitter API

No.	Url
586	http://Telegraph.co.uk
430	http://www.bbc.co.uk/news/uk-17741213
320	http://www.bbc.co.uk/sport/0/olympics/17742236
262	http://www.thesundaytimes.co.uk/sto/public/article1016905.ece
185	http://www.bbc.co.uk/news/uk-17747643
146	http://www.guardian.co.uk/p/3734e/tw
113	http://www.guardian.co.uk/sport/2012/apr/13/olympics-2012-branding-police-sponsorsnewsfeed=true
108	http://www.bbc.co.uk/news/world-asia-india-17737756
105	http://www.guardian.co.uk/sport/2012/apr/13/olympics-2012-branding-police-sponsors
104	http://www.bbc.co.uk/news/uk-england-london-17807502
103	http://www.bbc.co.uk/news/health-17744446
91	http://www.bbc.co.uk/news/uk-17752210
88	http://www.independent.co.uk/news/uk/politics/greenest-ever-olympics-claims-dismissed-as-corporate-spin-7647995.html
79	http://Mirror.co.uk
73	http://www.bbc.co.uk/sport/0/football/17753784
68	http://www.dailymail.co.uk/news/article-2133408/London-2012-Olympics-Worlds-biggest-McDonalds-1-500-seats-built-games.html
67	http://yourdailynews.co.uk/new/thisiswhatiwouldcallatest.php
65	http://www.guardian.co.uk/p/36pzb/tw
61	http://www.lrb.co.uk/blog/2012/04/20/evgeny-morozov/shop-your-neighbours/
58	http://www.guardian.co.uk/p/37vmc/tw
55	http://UrbanKronix.co.uk

BRITISH LIBRARY

Online Web Selection

Welcome Stephen Johnson Logout

Bookmark quick save

Create New Selection

Home | Bookmarks | Selections | Templates | Tags | Reports | My Profile

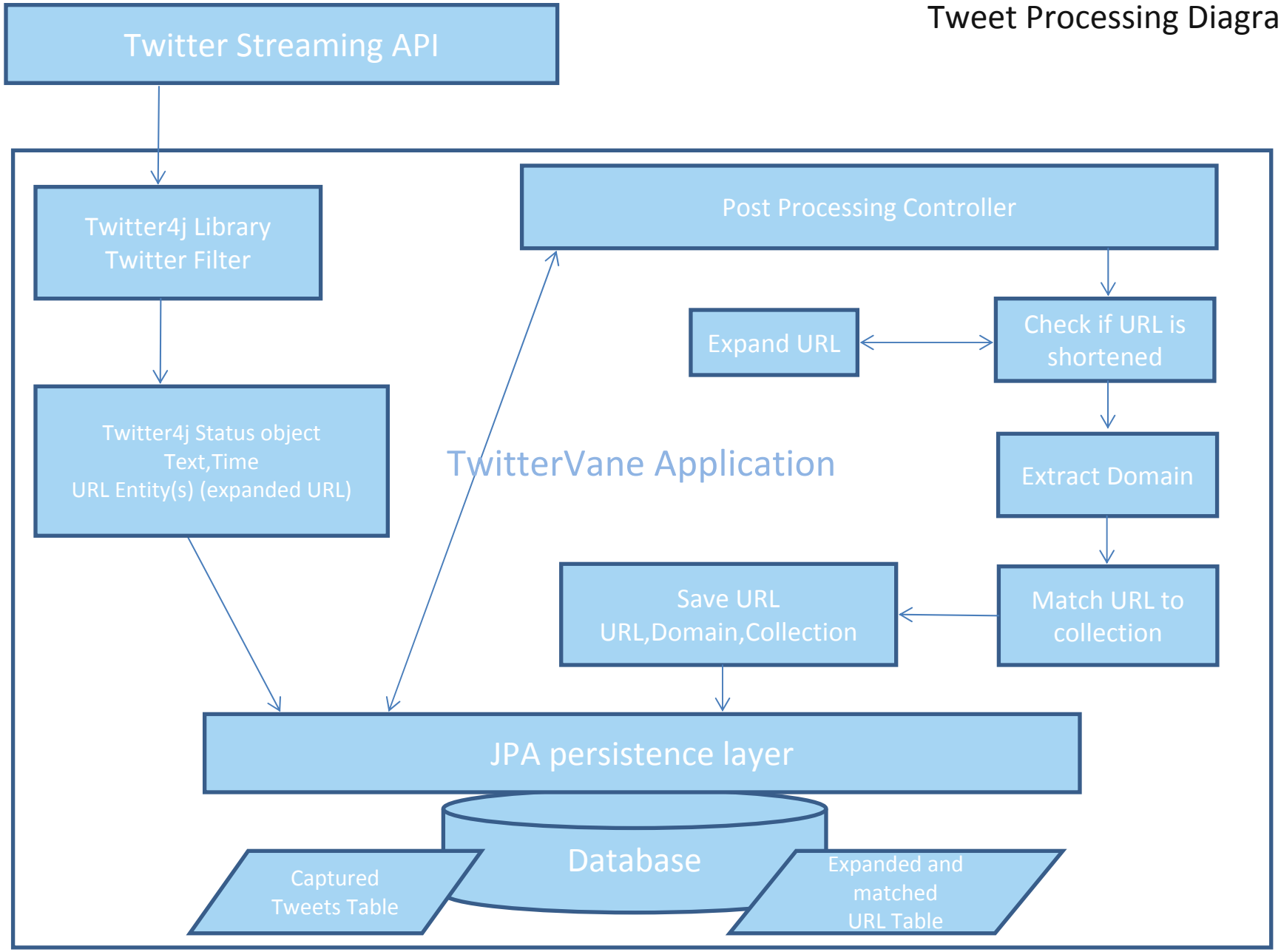
New Submissions | Permission Queued | Pending Permission | Permission Granted | Email Templates | Change Requests | News

Selection List

599 results Page 1 of 30 — 1 2 3 4 5 6 7 8 9 next> last>>

Title	Selector	Contact Email	Email Template	Permission	Opt Out
⚠ BBC News: Diamond Jubilee: Queen celebrating 60-year reign URL: http://www.bbc.co.uk/news/uk-16896731	Nicola Johnson		General	MPR	MOO
⚠ Samsung: London 2012 Olympics Games URL: http://www.samsung.com/uk/london2012/	Nicola Johnson		General	MPR	MOO
⚠ Team South West URL: http://www.teamsouthwest.co.uk/	Nicola Johnson	info@teamsouthwest.co.uk	General	APR MPR	AOO MOO
⚠ Championing the East Midlands URL: http://2012.emda.org.uk/	Nicola Johnson	regionalcoordinator@emd.org.uk	General	APR MPR	AOO MOO
2012/04/17) On the matters of the Olympics in London, Nazis, and Listener Mail | Citizen Radio, Tele URL: http://barginhunting.net/rand2/rand1.php	Twitter Vane		General	MPR	MOO
2012/04/17) On the matters of the Olympics in London, Nazis, and Listener Mail | Citizen Radio, Tele URL: http://callofdutyeliteclub.com/rand2/rand1.php	Twitter Vane		General	MPR	MOO
URL: http://funnygirlsvideos.in/i_always-knew-women-are-crazy-but-not-this-much.php	Twitter Vane		General	MPR	MOO
2012/04/17) On the matters of the Olympics in London, Nazis, and Listener Mail | Citizen Radio, Tele URL: http://www.nocnsf.nl/cms/showpage.aspxid=9236	Twitter Vane		General	MPR	MOO
2012/04/17) On the matters of the Olympics in London, Nazis, and Listener Mail | Citizen Radio, Tele URL: http://mediadecoder.blogs.nytimes.com/2012/04/17/at-london-olympics-nbc-says-if-cameras-are-on-it-well-stream-it/	Twitter Vane		General	MPR	MOO
2012/04/17) On the matters of the Olympics in London, Nazis, and Listener Mail | Citizen Radio, Tele URL: http://www.forbes.com/sites/marketshare/2012/04/13/the-london-olympics-are-coming-so-is-new-global-brand-building/	Twitter Vane		General	MPR	MOO

Tweet Processing Diagram



Lessons learnt & thoughts

- When possible it is better to search on a hash tag related to an event and to avoid using common terms
- Top URLs shared on Twitter seem to point to many of online news sites
- Time consuming to “select the selection”?
- Use as a selection tool for focused crawls?